Random Aspects of Beam Physics and Laser-Plasma Interactions

by

Andrew Emile Charman

A.B. (Harvard University) 1991
M.A. (University of California, Berkeley) 1996

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Physics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Jonathan S. Wurtele, Chair
Professor Robert G. Littlejohn
Professor Philip B. Stark

Spring 2007

The dissertation of Andrew Emile Charman is approved.

Chair	Date
	Date
	Date

University of California, Berkeley

February 2007

Random Aspects of Beam Physics and Laser-Plasma Interactions

by

Andrew Emile Charman

Abstract

Random Aspects of Beam Physics and Laser-Plasma Interactions

by

Andrew Emile Charman

Doctor of Philosophy in Physics

University of California, Berkeley

Professor Jonathan S. Wurtele, Chair

Aspects of the dynamics of charged particle and radiation beams, and of the interaction of plasmas with radiation are investigated, informed by concerns of classical and quantum mechanical uncertainty and noise, and related by notions of particle and radiation phase space manipulation, overlap, and control.

We begin by studying questions of optimal longitudinal pulse-shaping in laser wakefield accelerators, based on a one-dimensional model with prescribed laser drive and either a linearized or fully nonlinear quasi-static plasma response. After discussing various figures-of-merit, we advocate maximizing the peak wake amplitude instead of the transformer ratio. A number of new results are demonstrated, certain conjectures are rigorously proved for the first time, and some erroneous claims corrected. Generally speaking, shorter is always better for fixed pulse energy, at least until dispersive effects become important. In the nonlinear regime two (or more) pulses are always better than one, if they can be appropriately timed. For fixed peak intensity, "bang-bang" solutions consisting of sequences of square waves are found to be optimal. With effective bandwidth constraints imposed on the laser amplifier by modeling a realistic transfer function with possible phase or amplitude masks, we find no amplitude masking should be performed in the linear regime. The phases should be chosen to vary linearly with frequency to again ensure the shortest possible pulse in the time domain. A preliminary analysis is made of the Colliding Beam Accelerator, which aims to increase the wake excitation produced by a short seed pulse through the addition of a longer counter-propagating pump laser. If the intensity of the seed laser is limited, then this scheme can enhance the wake well beyond that achievable by the seed laser alone,

but this approach is almost always less effective than simply adding the extra energy to the short seed, if possible.

Instead of using short laser pulses to excite plasma waves, one can employ the beat wave between two co-propagating lasers to excite a Langmuir wave with high phase velocity suitable for acceleration of relativistic electrons. A modified version of this plasma beat-wave accelerator scheme is introduced and analyzed, which is based on autoresonant phase-locking of the nonlinear Langmuir wave to the slowly chirped beat frequency of the driving lasers via adiabatic passage through resonance. This new scheme is designed to overcome some of the well-known limitations of previous approaches, such as relativistic detuning and nonlinear modulation of the driven Langmuir wave amplitude, as well as sensitivity to frequency mismatch due to measurement uncertainties and density fluctuations or inhomogeneities. As in previous schemes, modulational instabilities of the ionic background ultimately limit the useful interaction time, but nevertheless, peak electric fields at or approaching the wave-breaking limit seem readily attainable. Compared to traditional approaches, the autoresonant scheme achieves larger accelerating electric fields for given laser intensity, or comparable fields for less laser power. The plasma wave excitation is much more robust to variations or uncertainties in plasma density. Further, autoresonance is largely insensitive to the precise choice of chirp rate, provided only that chirping is sufficiently slow. The quality and uniformity of the resulting plasma wave and its suitability for accelerator applications may be superior. In underdense plasmas, the total frequency shift required is only of the order of a few percent of the laser carrier frequency. For possible experimental proofs-of-principle, the scheme might be implemented with relatively little additional modification to existing laser systems.

From radiation exciting plasmas, we turn to issues of plasmas or beams emitting radiation. We develop a Hilbert-space and operator-based approach to electromagnetic radiation, and use this formalism to derive a maximum-power variational principle (MPVP) for spontaneous radiation from prescribed classical harmonic sources. Results are first derived in the paraxial limit, based on well-known analogies between paraxial optics and the Schrödinger equation for a single non-relativistic particle, and then generalized to non-paraxial situations. In essence, the variational principle says that prescribed classical charges radiate "as much as possible," consistent with energy conservation. The techniques are developed to model undulator radiation from relativistic electron beams, for which an example involving high harmonic generation is reviewed. However, they are more broadly applicable to other synchrotron radiation problems. Although similar to well known variational princi-

ples widely used in electromagnetic theory, the MPVP appears to be an independent result, and thus adds to the large family of variational techniques available for electromagnetic problems in general, and undulator radiation in particular. The MPVP is a straightforward consequence of simple physical and mathematical constraints: the outgoing power in any one source-free mode of the electromagnetic field may not exceed the total outgoing power in all the modes (i.e., Bessel inequality); while the power radiated must be attributable to power delivered by the sources, even in the regime where we ignore back-action on the sources (i.e., conservation of energy). However simple, these observations have practical implications for undulator and other radiation problems.

We next study a situation where wiggler radiation is both emitted from particles and re-applied to them. In stochastic cooling, information in the radiation induced from a particle bunch, if suitably amplified and fed back on the beam, can decrease entropy and increase phase space density. Specifically, we analyze and assess possible quantum mechanical effects in optical stochastic cooling. Fast stochastic cooling (i.e., on microsecond time-scales) would be desirable in certain applications, for example, to boost final luminosity in the proposed muon collider, where the short particle lifetimes severely limit the total time available to reduce beam phase space. But fast cooling requires very high-bandwidth amplifiers to limit the incoherent heating effects from neighboring particles. Transit-time optical stochastic cooling employs high-gain, high-bandwidth, solid-state lasers to amplify the spontaneous radiation from the charged particle bunch in a strong-field magnetic wiggler. This amplified light is then fed back onto the same bunch inside a second wiggler, with appropriate phase delay to effect cooling. Prior to amplification, the usable coherent signal from any one particle is quite small, on average much less than one photon for each pass through the wiggler. This fact suggests that the radiation must be treated quantum mechanically, and raises doubts as to whether this weak signal even contains sufficient phase information for cooling and whether it can be reliably amplified to provide cooling on each pass. A careful treatment of the dynamics, where the radiation and amplification processes are treated quantum mechanically, indicates that the feared quantum cooling catastrophe is in the end a red herring, and that fast cooling is in principle possible. Cooling rates essentially agree with a simple classical calculation, provided that the added effects of the unavoidable amplifier noise arising from quantum mechanical uncertainty are included. Thus, quantum mechanical uncertainties do not present insurmountable obstacles to optical cooling, but do establish an upper limit on cooling rates and a lower limit on achievable emittances.

Further examining the possibility of quantum mechanical effects of charges and their

radiation, we turn to quantum treatments of Electromagnetically-Induced-Transparency (EIT) in magnetized plasmas, in which the medium – normally opaque to a resonantly-polarized EM probe field at the cyclotron frequency – can be made transparent by the application of an intense EM pump at a frequency detuned below the cyclotron frequency by the plasma frequency. The phenomenon has a completely satisfactory classical explanation, yet is evocative of EIT in cold atomic vapors, down to the required "counter-intuitive" turn-on of the pump field prior to the arrival of the probe field, and the consequent "slow light", or small effective group velocity for the probe, which can be continuously tuned in proportion to the applied pump strength if the latter is slowly varied after the probe has entered the medium. However, the case of atomic EIT is often upheld as an example of an intrinsically quantum mechanical effect, where transparency is induced as a result of destructive interference between quantum amplitudes for different excitation pathways into the excited atomic state. This raises fundamental questions as to how and to what extent a seemingly classical phenomena in plasma can mimic a quantum mechanical effect in atoms. We address these questions by describing both systems in a common quantum mechanical language, where in the cold, unsaturated limit, the relevant excitations are associated with collective Bosonic modes, or quasi-particles. EIT can be understood in terms of the dressing of these modes via the pump-mediated interaction, leading to a dark-state polariton coherently combining both field and particle excitations that is largely immune to the cyclotron resonance. The analogy between the atomic and plasma systems is essentially exact, apart from differences of degree due to different orderings for the relevant frequency-scales in accessible parameter regimes, which can make certain scattering process that compete with the EIT interaction more prominent in the plasma case. Plasma electrons gyrating transversely about the guiding centers, and oscillating longitudinally about their equilibrium positions in a plasma wave, behave much like electrons bound in the Coulomb field of the nucleus.

Professor Jonathan S. Wurtele
Dissertation Committee Chair

At the end of a long journey and the start of another, this thesis is dedicated to my family....

To the generation that is now gone:

My grandmother, Carol Grace Faure, who knew little physics, but taught me so much about life, and love, and laughter; showed me that one should attempt to cultivate good sense, good humor, and good friends along with a good intellect and good education; and taught me that knowledge must begin and end in wonder....

My grandfather, Emile Faure, whom I never knew, but whose humor and *joie de vivre* touched me all the same, and whose hard work and devotion to family helped finance my own education....

My grandfather, Howard Ross "Bill" Charman, who tried to teach me the importance of character, the value of hard work, and the utility of careful writing....

My grandmother, Martha Mossholder Charman Reib, who died before we could truly share our mutual curiosity in all manners of things, or our stories of Berkeley....

My great-uncle, Max Miller, and great-aunt, Jean "Mimi" Miller, who shared with me their exquisitely bemused outlook on the world, a keen appreciation of both the absurd and the beautiful in life, a love of travel, and a sense for adventures both big and small....

My great-uncle, Raymond Sherwin, who near the end of his life shared with me his dedication to fair play, his sense of justice, and his love of nature....

My great-aunts, Dorothy "Dee" Sherwin Vellom and Gwen Sherwin Thompson, who kept alive our family's past....

To those who showed me the way, and caught me when I stumbled:

My late mother, Linda Charman, the most generous woman I have known, friend to multitudes and best friend to so many, whose love poured out for family and friends, for children and animals, for the bounties and beauties of the world, but who had so much more love to give, and wisdom to share....

My father, Charles Charman, the most decent man I know, dedicated to protecting and providing for his family, and whose quiet strength has been tested too often by loss, but has not wavered....

My uncle, William Ross Charman, whose own journey was cut all too short, with whom I shared a love of good books, good food, good company, and a good toast, as well as a suitably outraged political outlook appropriate for our times; and who showed me the power of an informal education that ventures freely wherever curiosity leads; and set an example of a life lived seeking balance of mind and hand and heart....

My uncle, Leon Faure, who challenged with many mathematical curiosities, entertained with strange but true tales of fishing, soccer, and classroom antics, and inspired by passionately articulating the importance of family and the value of teaching....

My aunt, Linda Charman, who remained aunt-like beyond any statute of limitations....

My uncle, Howard Charman, who shared with me many books, many family stories, and many ideas....

My aunt, Lorraine Simon, who offered, above all, quiet acceptance, and constant friendship....

To those who have walked with me:

My sister, C. Lorraine Hebert, who has remained ever patient and tolerant and supportive through all our years together; who on many early mornings struggled mightily to wake me up in time to learn; who manages to adopt an easy-going attitude amidst an exhausting schedule; and who during the long, long process of earning my Ph.D. knew to ask "how is it going?" occasionally but not too often....

My brother-in-law, Curtis James Hebert II, who often knows what to take seriously and what to laugh about, and who with good humor manages to get along with a partially-Yankee-educated physicist with what he undoubtedly regards as antithetical political inclinations, unusual interests, and occasionally dark or at least quiet moods....

My cousins, Lance, Mat, and Brian Faure, Thane, Travis, and Tyler Charman, and Bill and Chris Charman, who shared many meals and many laughs at family gatherings....

My cousin, Chris Simon, who shared his art and a few waves with a scientist....

My second cousins, Dan and Paul Vellom, for always making a point stop by....

And to those who have newly arrived, and are only beginning a journey of their own:

My niece, Chloe Lorraine Hebert, and my nephew, Curtis James Hebert III, who are the sources of endless laughter and joy, and whose delight and enthusiasm and curiosity have helped renew my own....

I see trees of green, red roses too;
I see them bloom for me and you,
And I think to myself, what a wonderful world....

I see skies of blue and clouds of white:
The bright blessed day, the dark sacred night;
And I think to myself, what a wonderful world....

The colours of the rainbow, so pretty in the sky
Are also on the faces of people going by...
I see friends shakin' hands, sayin' "How do you do?"
They're really saying, "I love you...."

I hear babies cryin'; I watch them grow;
They'll learn much more than I'll ever know.
And I think to myself, what a wonderful world.
Yes, I think to myself, what a wonderful world....

George David Weiss and Bob Thiele, 1967
What a Wonderful World

Contents

List of Figures

List of Tables

Acknowledgements

For this relief much thanks....

William Shakespeare
Hamlet, Act I, Scene 1

The University of Berkeley, where I was to teach, is the loveliest place one can imagine....

I believe in these people, even after seeing them at work in a setting where they are not at their best: integrating and differentiating at a theoretical physics seminar....

Ludwig Boltzmann (1905)
A German Professor's Journey to El Dorado

While writing a thesis may often seem like a quintessentially solitary, even lonely, activity, any graduate-level education in physics culminating in a doctorate is anything but an individual achievement. Late in his life, Albert Einstein wrote

> Many times a day I realize how my own endeavors rely on the labors of others, both living and dead, and how earnestly I must exert myself to give in return as much as I have received and am still receiving...

and that someone of such such extraordinary accomplishment, producing the most profoundly original, revolutionary, and singular physics of the 20th century would express this overwhelming sense of debt to his fellow-travelers only punctuates how much more grateful and humble we mere mortals ought to be, as we struggle even to accomplish a minor, derivative, incremental, or incidental bit of physics that may earn us our degree.

I mention Einstein deliberately, because, while working throughout 2005 towards what I then hoped was the completion of my dissertation, physicists everywhere enjoyed the World Year of Physics, endorsed by the International Union of Pure and Applied Physics, both the American and European Physical Societies, and many other organizations, to celebrate Einstein, his ideas, and their influence on life and science in the new century, as a way to commemorate the centennial of his *annus mirabilis* of 1905, in which he single-handedly published five papers revolutionizing most of physics: he more or less irrefutably demonstrated the then-still-controversial existence of atoms by analysis of Brownian motion; developed a method to determine the size of molecules and the value of Avogadro's constant through hydrodynamic behavior; explained the then-mysterious properties of the photoelectric effect

by essentially inventing the concept of what is now called the photon and taking the first real steps towards a quantum theory of light; transformed our very understanding of space and time and forever intertwined their meanings; and famously elucidated the connections between mass and energy — all while working a day job as a clerk at the Bern patent office! I joked with friends that it would be a minor miraculous year for me if I could just complete my mediocre and most definitely un-revolutionary thesis and at long last graduate.

Alas, perhaps characteristically, that moment for me has now slipped through all of 2006, to the beginning of 2007. Somewhat miscast in my previous role as a consultant, I came back to physics and to physics graduate school out of a desire to learn more deeply about what know of the world and how we know it. Undoubtedly, the part of this quest spent in graduate school has been a rather protracted one, but in its original purpose not entirely unsuccessful. Indeed, I have learned much, of all manner of things, some of which even found its way into this thesis. I can only hope that the great Eugene Wigner was at least slightly mistaken (not a very frequent occurrence!) when he said,

> Physics is becoming so unbelievably complex that it is taking longer and longer to train a physicist. It is taking so long, in fact, to train a physicist to the place where he understands the nature of physical problems that he is already too old to solve them....

Although not without many intellectual rewards and cherished moments, my somewhat plodding graduate school trajectory has tended toward the slow and tortuous, involving a plethora of interesting coursework, an immersion in teaching and learning, and a somewhat slow beginning, middle, and denouement to a research program, which alternated between periods where progress stalled altogether, veered off along interesting but ultimately often unproductive tangents, with curiosity sometimes overcoming capability or even common sense, and even included what might be called retrograde motion, where I seemed to learn only that what I previously thought I knew was wrong, or excitedly pursued promising approaches only to realize I had already once tried and rejected similar routes months before, or after much work discovered that my ostensible "discoveries" were actually discovered by Russians decades earlier. In between the set-backs, dead-ends, blind alleys, and wrong turns, the work of my thesis was gradually accomplished, taking form in fits and starts and periods of gradual accumulation, in episodes in which, in the inimitable words of Ogden Nash, "progress was fine for awhile, but it went on too long."

Time has slipped by (too much befitting one doctorate, but that cannot be changed now), and life has unfolded — loves and losses, joys and disappointments, feuds and friend-

ships, triumphs and travail — marked out by many milestones and memories both good and bad.

I have mourned the passing of far too many close to me — my grandmother died not long before my arrival here, followed by maternal and paternal great-aunts, my grandfather, and an uncle, taken far too soon.

Unaware of the bitter irony, my operating system blandly informs me that I wrote that sentence on February 21, 2006, exactly one month to the day before my Mother's death due to complications from a heart attack. I will miss her in more ways and more deeply than I can possibly describe – her laughter and her wisdom, her positive outlook and gregarious disposition that counterbalanced by sometimes introverted and cynical ways, her generosity of spirit and deed, her oceanic capacity for love. I happen to be writing this on a birthday of both of Mozart and Lewis Carol. The former once said "neither a lofty degree of intelligence nor imagination nor both together can create genius: love, love, love, that is the soul of genius." My mother truly possessed a genius for the interpersonal. Most of us are lucky to have a handful of good friends, but she was not just a friend, but a best friend, to many. She understood, as Lewis Carol did, that "one of the secrets of life is that all that is really worth doing is what we do for others."

My mother was always trying to teach me, with only partial success, to avoid wasting time with regrets for what we cannot change. As I reflect back upon a particularly painful year, and a long graduate program with more than a few wasted opportunities, I try to embrace this perspective that came so naturally to my mother, and, heed the advice of Longfellow: "Look not mournfully into the past. It comes back not again. Wisely improve the present, it is thine. Go forth to meet the shadowy future, without fear, and with a courageous heart."

Yet the cycle of life turns inexorably, and I have also celebrated many additions amongst friends and family to the next generation, including the births of my own niece and nephew. I have attended more weddings than I can count, across the bay and across oceans, from Washington to Wisconsin to Wales, from Hawaii to Houston, from Shasta to Chicago, from California to Cleveland to the Czech Republic; and learned of altogether too many divorces.

My sister got married, moved from Texas to Alaska, found new jobs, re-modeled houses, and started her family. My mother become a surrogate grandmother of sorts to many, and after years of waiting, official grandmother to two of her own, with all rights and privileges granted thereto, for a few beautiful years. She weathered two worrisome medical scares,

only to be taken without warning or reason by a third. My father left the company where he had worked for almost four decades and started his own engineering consulting business. My parents sold the family home that my father and uncle built, and in an unusual demographic reversal retired from sunny San Diego to snowy Anchorage.

I have had the opportunity to visit many countries on four continents: Canada, U.S.A., and Mexico in North America; Argentina and Brazil in South America; England, Wales, Scotland, France, Italy, Switzerland, Czech Republic, Slovakia, Poland, and Hungary in Europe; and the Republic of South Africa, Botswana, and Zimbabwe in Africa. I have attended meetings and conferences in San Diego, San Francisco, Santa Cruz, Santa Helena, and Santa Fe; in Aspen and Atlanta, in Baltimore and Boston, and in Oxnard and Orlando; in Cambridge (Massachusetts) and Oxford (England); in the "other" Lake Geneva (Wisconsin); in spots along the California coast from Long Beach to Monterey, and on the Mediterranean jewel of Capri.

I have suffered through presidential elections of somewhat questionable legitimacy, and witnessed in our country's so-called governance an absolutely frightening flirtation with fascism that I would not have thought possible. Like Tennyson's Ulysses, "all times I have enjoy'd /?Greatly, have suffer'd greatly, both with those/That loved me, and alone..." yet realize that "Tho' much is taken, much abides...."

Given the length of my graduate school journey, naturally I have been the object of some much-deserved but good-natured ribbing and ridicule, some of it from myself — a good friend suggested that if it dragged on much longer I might be eligible to become a graduate student *emeritus*, while I countered by suggesting that perhaps my time in graduate school should be divided not into years but into geologic epochs. I have earned special commendation for decrepitude at the departmental holiday party, and have provided some solace to other graduate students in their own delays, consoled by the thought that at least they did not take as long as that Charman fellow.... While I am almost certainly the longest-running current student, I am told that I am still years away from the Departmental Record.

But behind any teasing, friends, family, and fellow graduate students have been extremely supportive during the sometimes Sisyphean struggle. "A joy hard earned is doubly-savored" wrote the Jesuit philosopher Baltazar Gracián, while Goethe opined "a joy shared is a joy doubled," so my friends, family, and I will no doubt indulge in redoubled relief as this task is at last completed.

"Time is the worst place, so to speak, to get lost in," Douglas Adams deadpanned[1], but If one is to be stuck rather too long for one's own good (but admittedly, mostly through one's own failings) in graduate school, Berkeley is indeed one of the best spots imaginable to be so delayed — not just the renowned university, with all the talent and visitors that attracts, and a physics department with storied history, but also its spectacular skies at sunset, its parks and hills and gardens, the views of the bridge and the bay and the land beyond; its love affair with good food; its ability, as you walk in some seemingly forgotten corner of campus, or along a quiet street in the town beyond, to surprise you with some exquisite exemplar from a golden age of California architecture: a building by Coxhead, Hall, or even Greene and Greene, by John Galen Howard or John Hudson Thomas, by Maybeck or Morgan or Mullgardt, by Polk, Schweinfurth, Yelland, or many others whose names I do not know....

Besides being the world year of physics, 2005 A.D. also marked the centennial of a more minor footnote in the history of science: the visit of Ludwig Boltzmann to teach at Berkeley in the summer of 1905 (a little less than a year prior to his tragic suicide), an opportunity which elicited his travelogue quoted above. In his memoir, he noted in the Californians he encountered a unique blend of pragmatism and idealism, not readily found in the Old World, and expressed a belief that they were destined "to achieve great things." In the intervening century, much has transpired. Some things have remained constant — he admired the "trees that have seen the centuries go by – or is it millennia?" and wrote approvingly of the steam heat "that was often welcome, even in July...." Many things have changed, some for the better: long before Alice Waters and others transformed Berkeley into a culinary beacon, it was evidently something of a backwater, where Boltzmann complained not a little both about the mediocrity of the food (referring at a dinner hosted by Mrs. Phoebe Hearst to a gruel "one might use to fatten geese in Vienna..... but I doubt that Viennese Geese would touch it...") and about the lack of suitable beverages (it was then a dry town, so he had to smuggle wine in from Oakland) — while some things serving to remind us that change is not always progress — reporting on a flurry of new campus construction, he said, "there is, after all, plenty of room and money...." But undoubtedly he would have been delighted by the accomplishments of this department here at the edge of the Western world, perhaps especially in its golden age before and after World War II, but also today, as the department emerges with renewed vigor and creativity from what some called a period of "genteel decline."

[1] *Apropos* of my graduate school career, he also remarked "it takes an awful long time to not write a book..." and "I love deadlines — I like the whooshing sound they make as they fly by...."

I cannot objectively comment on how someone like myself managed to slip quietly into and ever-so-slowly through the hallowed halls of LeConte and Birge, but I have been constantly surprised, challenged, and stimulated by the curiosity, cleverness, creativity, and courtesy of my many colleagues — the faculty from whom I have learned so much, the undergraduates whom I have helped to teach and who constantly forced me to understand and explain physics better, perhaps especially the fellow graduate students with whom I enjoyed many musings, meals, and memories, and of course the department's staff who like the Wizard of Oz make really make all the magic possible from behind the curtain.

My advisor, Professor Jonathan Wurtele, generously provided many years of academic, intellectual, financial, material, and logistical support; secured precious office space in which to labor; supplied many fancy computers on which to work; engaged in interesting and free-ranging discussion of science and many other topics; and nurtured a talented group of students, post-doctoral fellows, visitors, collaborators, and miscellaneous others, with whom to ponder, discuss, calculate, read, write, and publish. He gave me the freedom (some might instead venture rope...) to pursue many intellectual passions and distractions and to follow eclectic interests wherever they might lead (often down somewhat unorthodox back-alleys or into unconventional corners, and not infrequently into complete dead-ends), even when many of these investigations were perhaps in the end not particularly productive or pertinent to the main thrust of our group's research. I can only hope that in my many wanderings through obscure intellectual thickets and brambles I managed to stumble upon just enough green meadows and ripe fruit to make his considerable investment of time and money worthwhile. He has been kind enough to presume that my glacial pace was less from a leaning toward laziness than a thirst for thoroughness, a certain excess of inquisitiveness, or a propensity to ponder rather than produce. Along the way, he has patiently tried to teach me that sometimes "better is the enemy of good enough"[2] and the important distinction between learning interesting physics and actually being a successful physicist in academia or elsewhere. Any persistent incapacity to internalize such lessons remains my own failing, not his.

All extremely busy with their own careers, research, families, and lives, Professor Jonathan S. Wurtele and Professor Robert G. Littlejohn from the Physics Department, and Professor Philip B. Stark from the Department of Statistics, generously agreed to serve

[2]Attributed to Voltaire (1794), although some evidence suggests he may have been paraphrasing a still older Italian proverb. This or similarly-worded sentiments have also been widely attributed to various subsequent military figures, such as Clausewitz, General Patton, and Soviet Admiral Sergey Georgievich Gorshkov.

as my dissertation committee, reading this turgid thesis and offering their expertise, and who along with Professor Raymond Y. Chiao (now of U.C. Merced) also wasted what I am sure was a somewhat painful afternoon at my oral qualification exam. Over the years I have enjoyed many conversations with Professor Chiao listening to his unique perspective on quantum mechanics, and with Professor Littlejohn, gaining much insight from his expertise in mathematical physics and issues at the interface of classical and quantum dynamics. Professor Littlejohn agreed to read this thesis at short notice at the busy beginning of a new semester, just after returning from sabbatical. Professor Stark read with a careful and incisive eye under particularly trying conditions, while convalescing from complications from a major back injury. All of my committee members patiently tolerated my idiosyncrasies of style, occasional obscurity of substance, and persistent procrastination in delivering either.

Along the way, I have taken or audited a number of classes, and benefited from the skills and efforts of many dedicated teachers, especially: Professor Robert Littlejohn in quantum mechanics, semi-classical mechanics, and topological methods; Professor J. D. Jackson in classical mechanics and electromagnetic theory; Professor Allan Kaufman in Hamiltonian dynamics and in plasma physics; Professor Eugene Commins in statistical mechanics; Professor Eyvind Wichmann in group theory; Professor Hirosi Ooguri in general relativity; and Professor Dan Rokhsar and Professor Harold Lecar (the latter of the Department of Molecular and Cell Biology) in biophysics.

Jim Morehead began as my teaching assistant and has remained a friend, and over the years has illuminated much, especially about short-wavelength asymptotics, optics, and the intricacies and delights of phase space. Bruce Birkett has helped me to become a better teacher, a better gourmand, and better friend, but I have still have much to learn.

In our combined group meetings, Professor Joel Fajans attempted to explain the intricacies of real experimental physics to a bunch of ignorant and naive theorists, patiently answered many of our stupid questions but asked many smart ones; shared his insight and his humor; and often offered an amusing political counterweight to my advisor, whose own opinions are what might be described as somewhat atypical for a town like Berkeley....

I have also attended more talks and seminars than I can count or name, where I have encountered many stimulating ideas, but would like to praise as particularly useful the dynamics seminar series, sadly now extinct, established by Professor Allan Kaufman along with colleagues from the the Department of Electrical Engineering and continued for a while by Professor Edgar Knobloch; the biophysics and neurobiology seminar run by Professor Harold Lecar; and the the broad-ranging atomic physics seminar series run by Professor

Dmitry Budker followed by Professor Dan Stamper-Kurn, who were generous enough to allow a classical plasma physicist the chance to drag down their average by twice contributing presentations.

At the Center for Beam Physics at the Lawrence Berkeley National Laboratory, many have been generous with their time and expertise. In particular, I thank Max Zolotorev for sharing his sometimes idiosyncratic but always interesting insights and ideas, and for being patient with an American physicist who would undoubtedly fall far short of Russian standards; Eric Esarey for helpful discussions and an invaluable review article of laser-plasma acceleration; and Bill Fawley for good-natured advice and assistance, kind words of concern and encouragement, and some healthy venting of political frustrations.

Among full-time, one-time, or sometime members of the group, I should single out my office-mates for special thanks for long tolerating my curmudgeonly ways and questionable office organizational techniques – stacks of books that tempt onset of gravitational instabilities, and piles of papers that give aid and comfort to the Second Law of Thermodynamics. I shared space with Katya Backhaus during the "paleolithic" epoch of my graduate school career, and Ryan Lindberg during the "neolithic" phase. Katya is one of those people blessed by a true excess of talent who remind the rest of us really how much less we should be sleeping and harder we should be working in order to keep up. Watching her work, she seemed to approach physics with an efficiency and ease that escapes most of its practitioners, with little outward manifestation of the all too common sense of exasperation that threatens to overcome us as we edge tentatively along the precipice of that gaping chasm between question and answer, or between equation and solution. Besides doing theoretical plasma physics better than most, in her spare time she also cooks, sings, plays the piano, paints in oils and watercolors, reads extensively in at least two languages, volunteers to help the elderly, adopts stray dogs, works as a professional aerobics instructor, and now also operates an organic farm, and is mother to two toddlers. Seriously, I am not exaggerating....

Ryan, in turn, volunteers to teach mathematics at San Quentin, presides over the department's semi-official Bud Lite Fan Club, and plays a mean game of soccer and ping pong (not, as far as I know, simultaneously), and has been a good friend and office-mate, never too embarrassed to include an over-the-hill colleague in various activities. He has collaborated on several projects, helped me dot not a few 'i's' and cross many a 't' in my own research, successfully picked up a few of my suggestions and pursued them much further and more completely than I, acted as effective sounding-board and listened patiently to many strange ideas as well as various political tirades, and challenged me to think more

carefully and explain more clearly. I hope that I have been able to teach him something in return, if only by counter-example or object lesson as to how not to organize his own graduate studies and time-table.

Fellow graduate student Vladimir Gorgadze tolerated my very un-Russian willingness to numerically integrate when in only a matter of days or weeks an analytic solutions might be found, my possibly overly-harsh criticisms of the Soviet Union and post-Soviet dis-Union and various other odd ideas and unfounded opinions, and what, by his own Slavic standards, must have been seen as an abominably poor ability to hold my liquor. For my part, I enjoyed many conversations on various scientific, geopolitical, and other topics, and I can only hope that these same conversations were not too aggravating from his own perspective.

During our time of overlap in Berkeley, Palma Catravas gave generously of her time and of her reserves of good sense and good cheer, and provided a wonderful piano sound-track to the activities at the Center for Beam Physics.

Over the years, Carl Schroeder has provided helpful advice, discussion, perspective, and encouragement, as well as much hospitality together with his wife Sarah. Despite the demands of career and new family, he has always taken time to answer questions or offer suggestions. In recent years we have shared several meals and conversations, as well as an expensive hotel room and much of the local *mozzarella di bufala* cheese and *lacrima de Christi* wine at a conference on Capri, which he still insists is only the second best site ever for a beam physics conference....

Among the many post-doctoral fellows that have at one time belonged to the Wurtele research group, I would like to acknowledge in particular Min-sup Hur for his positive outlook and productive collaboration; Gregg Penn for giving freely of his time and ideas while quietly and serenely going about the sometimes overlooked business of producing good physics; and especially Brad Shadwick, who in many conversations and email exchanges over the years has generously shared his insightful and original outlook on mathematical and computational physics, his formidable expertise with TEX, C/C++, and all things numerical, and a Canadian's perspective on the American experience, and who within the plasma and beam physics communities has played both sherpa helping to maintain Apple Computer's foothold in scientific computing, and devil's advocate to the more over-zealous Particle-in-Cell proselytizers.

Beyond my group, in my entering class or elsewhere in the department, I have been fortunate to meet and interact with many extremely talented students and fellow-travelers,

too many to mention by name, and even more fortunate to be befriended by a few while we struggled with puzzles, pondered questions, and exchanged ideas.

In his comparably short time here at Berkeley, Ari Mizel was a true friend, offering collegial companionship and conversation, intriguing speculation, bad jokes and a good appetite, an original and creative outlook on all manner of things. He went out of his way to include me in many adventures and activities. I am grateful that despite tough times he did not completely give up on me. His mind is as incisive as his soul is gentle, and I only hope that he finds what he seeks.

Eric Chang and Kevin Mitchell provided insights into physics and mathematics, as well as several great meals. Kevin and his wife Diana have continued to offer good advice, good food, and good cheer. Beth Chen and Pavel Volfbeyn provided friendly outlooks, encouraging words, and a large amount of grilled meat. With Elliot Greenfield, I shared many interesting classes, reading groups, and conversations about nonlinear dynamics, evolutionary theory, neurobiology, information theory, Italian wine, Prague restaurants, Scottish single Malts, and many mutual political exasperations, sometimes joined by Elliot's advisor, Harold Lecar. An informal reading group with Elliot, Dan Butts, and Anne Hsu led to many interesting ideas.

Through many years here, especially as more and more of our colleagues graduated, Roy Therrien remained a loyal friend. First a soldier and helicopter pilot, then by turns low-temperature experimentalist, NATO peacekeeper, defense analyst, and now diplomat, Roy is the one physicist you would want at your back in the proverbial or literal dark allies of life. He also makes an excellent minestrone. Over a pint of beer or glass of wine, we have shared many conversations on everything from evolutionary psychology to European history. I will cherish memories of our bourbon-tasting, comet-watching, and other adventures. More than most, he speaks his mind, lives without regrets, happily ignores any constraints of political correctness, questions your assumptions, and keeps you honest. Agreeing and disagreeing with him are almost equally enjoyable. If anything, he is too prone to admit and exaggerate ignorance. Trained in the controlled use of violence and the practical implementation of peace, he put up with my pathetic physical conditioning and tendency to over-intellectualize just about everything. He has almost literally picked me up when I was down. It is one thing for your friends to fail to see in you the faults that might immediately bother others; it is altogether another for them to be fully aware but tolerate them anyway.

I would venture to say Jason Zimba is a kindred spirit, except that I am afraid that might be heard more as an insult than as a complement coming from the likes of me. It

now seems vaguely out of fashion to heed Socrates and live the Examined Life, but Jason manages to do so without apology, affectation, or awkwardness, yet with a demeanor that instantly diffuses any possible resentment over his probably being smarter than almost everyone else in almost any room. He is an immensely talented person, an extraordinary intellect, a gifted teacher, and a generous friend. Few can accomplish so much on so little sleep. Interested and conversant in just about everything, he is the kind of person who not only decides to master the New York Times crossword puzzle, but out of curiosity also plots his completion times to quantify his progress and correlate difficulty with day of the week, and then goes on to compose his own complete puzzle, themed around twentieth-century African literature. I used to think I knew something about poetry, at least for a physicist, but no longer can entertain delusions of being well-read or well-educated. Two and a half millennia ago, Heraclitus is said to have written in *On the Universe*, "It is better to hide ignorance, but it is hard to do this when we relax over wine." Friends like Jason do not hold this against you. Together we discovered that graduate students can sometimes live the Good Life. The Italians have a saying to the effect that "time spent at the dinner table does not count against one's hours allotted on this earth," and for our sake I hope it is true, although if not, I can still scarcely imagine hours better spent, lingering over good food and good wine, maybe even good whisky, discussing in turn the intricacies of quantum measurement or a Milanese risotto.

Celeste Winant shared her Medieval, Renaissance, Baroque, and Classical choral music, while Lorraine Sadler shared restaurant gossip and cooking advice, and Colin McCormick shared something approaching stand-up physics comedy. I have been fortunate to meet and interact with all these students and many more besides, who have immeasurably enriched my inner and outer life as a graduate student.

The physics department staff, singly and collectively, has been enormously helpful, invariably friendly, and infinitely patient when inevitably I failed to turn in the right form with the right signatures to the right place at the right time. In the main office, Professors Roger Falcone and Marjorie Shapiro have as departmental chairs never complained as I single-handedly skewed their time-to-graduation statistics, and Carol Dudley kindly tolerated my continuing presence in the department for a duration longer than she should have been forced to endure, and recently acted as my official champion in battling bureaucratic dragons.

In Student Services, Claudia Trujillo helped me reserve classrooms and secure textbooks at the last possible minute, organized the famous Physics of Music and Motion "seminar"

series, and always offered a kind greeting and friendly smile, however harassed and overworked she might be.

Anne Takizawa amazed all of us with her appearance of effortless competence, but while the competence of the highest order was very real, we knew of the prodigious amount of effort required behind the scenes to make things look effortless. She saved me from many potential missteps and missed deadlines, arranged preliminary and qualifying exams, and always offered a calm perspective and sound advice, as well as many good cinematic recommendations.

Often acting as semi-official departmental morale officer, and self-proclaimed Physics Phairy, Donna Sakima reminded me of important deadlines before they passed, or somehow managed to circumvent them after I forgot them anyway, solved my problems, patiently answered questions she had answered a thousand times before (or even more, if you count other students...), offered good humor, exotic foodstuffs, and corny emails, and arranged outings for *dim sum* and Neapolitan pizza.

Together they planned holiday parties, Spring picnics, poster sessions, welcome receptions, and graduation ceremonies, forgave our failings and diligently refused to pass judgement on our mistakes and delays, and generally worked tirelessly with little reward to make the requisite hoop-jumping more pleasant, and our experiences more enjoyable. They played Beatrice to our Dante, guiding us safely through the various circles of Berkeley's bureaucratic hell, past physics purgatory, and toward the paradisiacal Ph.D.— or maybe Ariadne to our Theseus, cleverly leading us through the labyrinth of graduate school at a Large Public University. Metaphorical meandering aside, I cannot thank them enough for their support and hard work. When once asked whether he would like to have a statue erected in his memory, Cato the Elder (Marcus Porcius Cato) reportedly said, "after I am gone, I would rather have people ask why I have no monument than why I have one." I suspect our office staff is of similar outlook, but I do hope they fully realize their importance to this department and this university, and know what a difference they make in the lives of the students here, from the moment we visit and decide what school to attend, and compare the friendly smiles here to the dour and dismissive faces in certain other premier physics departments, through all the exams and paperwork and procedures along the way, to the fresh strawberries greeting us at graduation.

Before her well-deserved retirement, Laura Fan at the Physics Library patiently helped me track down books and forgave more than a few fines. She was generous with her time,

expertise, and numerous volumes withdrawn from the physics library that managed to find their way into my personal collection.

Although my family was featured in the dedication, I would be remiss if I did not acknowledge them explicitly and thank them profusely again here, particularly my parents Charles Charman and the late Linda Charman, a fathomless wellspring of emotional, material and moral support, of solace, forbearance, encouragement, advice, and assistance throughout my life, in good and bad times. How do you thank those who gave you this mysterious and inexplicable gift of life, only to dedicate a good portion of their own to your protection and nurturance? I would also like to acknowledge my sister, Lorraine Hebert, for her even-tempered disposition, calm outlook, easy-going acceptance, and endless patience. She always looked after me while we were growing up, and still does.

Many friends from college and even high school that also settled in the Bay Area offered some social connection outside the physics department, and waited patiently in between my sometimes infrequent phone calls or visits: from college, George Lumpkin, Ann Ryu, and my old roommate Josh Peterson, and from high school, Chris Piccioni, Jason Harris, and Heather Chung for a time, and Francis Kelly, who kindly invited me to many get-togethers around San Francisco and holiday meals at his family's table. In my early years of graduate school, Michael Mitzenmacher and Steve Lumetta, both friends from the Research Science Institute, a secondary school science summer program founded by Admirial Rickover, offered collegial advice and hospitality to a newcomer to Berkeley. While Lauren Ancel and Jay Koh lived in the area and when Rich Simon and Olgica Bakajin, all alumni from the same program, moved to the area, they too offered an occasional respite from LeConte and Birge Halls.

I also enjoyed memorable visits from other college friends, including my other college roommates Scott Meltzer, Brian Landzberg, Ian Steaman, and Jeremy Druker and his Prague family, as well as Laurie Belin and Dave Friedman.

In the department or around campus, various random encounters with Andrea Jani, Michelle Hoffman, Jennifer Nickel, and Live Rekvig led to rewarding friendships, now mostly lapsed for no particular reason, the way friendships sometimes do. Jeff Grossman encouraged me to explore ballroom dancing, taught the infamous "Physics of Music and Motion" seminar series, otherwise known as salsa lessons, in LeConte Hall, and with his wife Kate Moschandreas has offered much hospitality and many opportunities to deconstruct the miasma that it our current political landscape.

Over the last twenty years, the hard-working staff of the Center for Excellence in Education, in particular the president, Joann P, DiGennaro, her assistant, Mariana Pestana, and vice president for programs, Maite Ballestero, have offered friendship, professional and personal support, and everything from letters of recommendation to free trips to Africa.

The Berkeley ballroom dance club provided many enjoyable distractions from physics, many memories, and many friends, including Patty Linden, Eric Luke, Kristy Wirthlin, Ann Nguyen, Rebecca Melton, and Richie Hom. Over the years, Elaine Ashby – doctor, dancer, medical researcher, engineer, seamstress, chef, and auto-mechanic, among other roles – has provided outings, meals, and medical advice with her inimitable combination of Southern hospitality and Northern efficiency.

Among the many people whom I was fortunate to meet at or through the ballroom dance club, I would especially like to thank Tasha Fairfield for all her love, kindness, and support over the last few years, sticking by me through some dark days, and believing in me when I did not. In the final stretches of this intellectual and emotional marathon, she has shown immense patience, and tried to bring to my life some sanity and sunshine – I only hope that I can now offer more in return. During the tense and sleep-deprived end-game pitting me against both my own considerable powers of procrastination and arbitrary decrees from the administration, she has provided everything from proof-reading to catering, with an amazing forbearance – even I did not want to be around me and my moods.

It has been a long journey, and an especially painful year. Living under a dark cloud, it has been particularly difficult to concentrate sufficiently on physics to complete this thesis, but I am grateful for all the expressions of sympathy and support – the phone calls, letters, cards, emails, and personal visits – from many acquaintances, friends, and relatives. Many in my mother's expansive circle of family and friends still feel as if there is a terrible rent in the tapestry of life, but together we are trying to reconnect threads and restore some of the beauty, although the pattern will never quite be the same.

Turning from the existential to the mundane, this thesis was composed, compiled, and typeset on a Macintosh G5 running MacOS X 10.4, using the PDFLaTeX engine within Gerben Wierda's distribution of TeX for MacOS X, based on the TeXLive distribution of the TeX User Group (TUG) and Thomas Esser's teTeX code, together with the application TeXShop (version 2), developed by Richard Koch, *et al.*, as text editor and front-end, as well as the extremely useful September 2004 version of the U.C. Berkeley thesis LaTeX template updated by Matthew A. d'Alessio, derived from the modifications of John T. Whelan and the changes of Chris Martin to an update by Blaise B. Frederick of the LaTeX2e

ports by Dan Gildea and Karl Berry of the original thesis style file written by Ethan V. Munson, based on the original report class with modifications by Frank Mittelbach and Rainer Schopf, all of which constitutes a contemporary implementation of Leslie Lamport's LaTeX language, which is an extension of Donald E. Knuth's original TeX typesetting and formatting program. Their innumerable hours of hard work immeasurably reduced my own (possibly less the time spent in the construction of the previous sentence).

In my first year here, I was supported by a University/Department of Education Fellowship, as well as through appointments as a graduate student instructor that year and for a few years directly and intermittently thereafter. The research reported in this thesis and elsewhere was completed with the financial support of my advisor, Professor Wurtele, through grants awarded to him as Principle-Investigator or Co-Principle-Investigator from the Advanced Accelerator Concepts initiative of the Office of High Energy Physics at the U.S. Department of Energy, and from DARPA, U.S Department of Defense.

One of the interesting things about the provisional nature of science is that in the end there is no shame in being wrong, as long as you are wrong for the right reasons. I have exerted myself to ensure that what is reported here is mostly true, to the best of my limited abilities, and original to the best of my knowledge, if neither especially interesting nor important. Still, I have learned much, and lived much, along the way. "Our science seems primitive and childish in comparison with reality," Einstein said, "but it is the most precious thing we have."

If, as is often said, the true value of science is to generate not answers, but more questions, questions we did not know to ask, or how to ask, until our ignorance was refined, then my protracted stay in graduate school shall count as something of a success, in least insofar as my own personal ignorance has progressed. It is somewhat sobering to realize how often even (or perhaps especially) the greatest scientists have tended to describe human ignorance, either individual or collective, in terms of the oceanic. In 1887 T.H. Huxley wrote:

> The known is finite, the unknown infinite; intellectually, we stand on an islet in the midst of an illimitable ocean of inexplicability. Our business in every generation is to reclaim a little more land.

echoing (I would imagine deliberately) the famous recollection of Newton[3]:

> I seem to have been only like a boy playing on the seashore, and diverting myself

[3]Newton is painted by his biographers as being essentially humorless, but surely he puns deliberately, here, since pebble in Latin is *calculus*?

> in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me.

If Titans such as Newton find mere pebbles here and there on the shore, what hope do we mere mortals have, except to keep our eyes open and perhaps catch a glimpse of a speck or fleck or mote in a sunbeam? I leave graduate school with more questions than answers, but perhaps that is exactly how it should be.

Physics progresses only because it is both cumulative, building on the hard-won knowledge of previous generations, and provisional, destined to be proven false in some sense, but in the interim offering insight into the way the world seems to work, and hopefully proving useful to those who come after and help to uncover something truer still.

I never cease to be awestruck when contemplating how human knowledge in general, and science in particular, can can keep present the past, or be whispered into the future, how it can speak across continents and down through the ages, or how we may learn something of the mind of another far away or long gone by a few marks on a page. As Lord Dunsaney wrote

> ...he knew of the things that ink may do, how it can mark a dead man's thoughts for the wonder of later years, and tell of happenings that are gone clean away, and be a voice for us out of the dark of time, and save many a fragile thing from the pounding of heavy ages; or carry to us, over the rolling centuries, even a song from lips long dead on forgotten hills.

In 1676, Newton[4] famously wrote in a letter that "If I have seen a little further than others, it is by standing on the shoulders of giants." We of course have something Newton did not, which is to say the legacy of Newton himself, as well as Laplace, Lagrange, Hamilton, Fourier, Faraday, Maxwell, Hertz, Boltzmann, Gibbs, Kelvin, Thomson, Rutherford, Planck, Einstein, Bohr, Schrodinger, Heisenberg, Dirac, Bethe, Feynman, Schwinger, and hundreds and hundreds more between and beyond.

Countless individuals, past and present, near and remote, have helped me in countless ways with this thesis, directly and indirectly. Now that I am graduating, I can only hope to find some way to begin to repay my intellectual debt to those who have taught me.

[4]Newton may have been paraphrasing a comment by George Herbert in 1651, who probably borrowed it from a passage of Robert Burton written in 1621, which quoted a 16th century Spanish theologian named Didacus Stella, who probably borrowed the idea from a saying attributed to Bernard of Chartres by John of Salisbury in 1159, while Bernard may or may not have read of attribution of a similar metaphor to the early sixth-century Roman grammarian Priscian, which all just goes to prove Newton's point.

It is part of this sometimes paradoxical nature of science – cumulative but provisional, incremental but original, collaborative but solitary – that any successful results, however marginal or minor, are surely communal in origin, to be shared by my advisor and mentors and teachers, by co-workers at the next desk and colleagues or contemporaries across the country or across oceans, and by all those, living and dead, with names we utter in awed reverence or with names we never knew, whose hard-won gains in knowledge and insight have been gifted to us across time and space, who pondered and toiled and meticulously observed and calculated so that we might live in this remarkable age where we have wrested, however tentatively, a bit of understanding of nature's beauty and mystery; while it is equally true that all of the inevitable errors mathematical, physical, historical, rhetorical, grammatical, or typographical, all lapses or leaps of logic, all sins of omission or commission, all blunders and bungles, defects and deficiencies, faults and failings, all flaws, falsehoods, fallacies, and *faux pas* in theory or fact, principle or practice, any inconsistency or incompleteness, any slights or slips, any mistakes, missteps, misstatements, or misattributions, any miscalculations, misunderstandings, or misprints, and any and all misjudgement, misreading, misapprehension, or misinterpretation, are mine alone.

Chapter 1

Introduction

...let's prepare to grapple with the ineffable itself, and see if we may not eff it after all.

Douglas Adams

The night is large, and full of wonders....

Lord Dunsany

Et ignem regunt numeri.
(Even fire is ruled by number)

Jean Baptiste Joseph Fourier

Those who venture into this thesis will quickly discover that the title embodies an intentional *double entendre*. By "random aspects" I refer to the fact that the various chapters constitute a sundry collection of topics in the physics of under-dense plasmas and of particle and radiation beams, but more importantly to the observation these specific investigations, as well as my entire graduate research program more broadly, are unified by a focus on issues of randomness and noise – classical and quantum mechanical, statistical and numerical.

This introduction therefore serves as both preface to the following chapters and a *précis* in a larger sense of the major themes of my research. A diverse but representative sample of topics are presented in detail in the following chapters. These investigations and associated mathematical techniques form part of a larger body of ongoing research in various stages of development. I have confronted limitations of time and space, but I mention some other projects of similar theme or approach in passing to position the following chapters in a broader context, and to reinforce the common threads that run through my research.

Chapter 3 addresses issues in the area of classical noise and fluctuations in beams and plasmas. I report on possibilities for using the nonlinear phase-locking phenomenon known as autoresonance in the plasma beat-wave accelerator (PBWA), as a mechanism for robust excitation of Langmuir waves. This approach can overcome relativistic detuning and other limitations, but more importantly, it is much more robust than conventional schemes in the presence of inevitable experimental *uncertainties* as to the exact plasma density, or actual physical *variations or fluctuations* in the density. Ongoing research on similar themes includes: a) the use of fluctuational spectroscopy of spontaneous undulator radiation to reconstruct particle bunch length, transverse emittance, and other aspects of the particle beam phase space, leading to a non-destructive, single-shot beam diagnostic applicable to very short bunches; and b) the effects of plasma density fluctuations on gain and coherence in the plasma-based Raman amplifier.

While noise and uncertainty are ubiquitous in classical plasma and beam physics, they are truly fundamental to quantum mechanics. Much of my research has focused on possible quantum mechanical or semi-classical effects in the interactions between beams or plasmas and radiation fields. In Chapter 5, I assess possible quantum mechanical limitations on a proposed method of ultra-fast, transit-time optical stochastic cooling, applicable to final cooling of muon beams, where opportunities for increasing phase space density are intrinsically limited by finite particle lifetimes, or in other applications where cooling rates faster than those achievable with conventional radio-frequency (RF) techniques might be desirable. Because stochastic cooling makes use of a very small cooling signal emitted by each particle in a large noise background, it is important to address whether quantum effects might obliterate or overwhelm this information. In Chapter 6, I study the classical/quantum correspondence in Electromagnetically-Induced-Transparency (EIT). A careful quantum mechanical formulation of both systems within a collective-mode formalism reveals close analogies in the unsaturated limit between EIT in magnetized plasmas and the more familiar version in cold atomic vapors. A related topic that I have pursued involves derivation and analysis of a one-dimensional, but fully quantum-mechanical, model of the Free Electron Laser (FEL) in the self-amplified spontaneous emission (SASE) regime.

Thinking about quantum mechanical and statistical issues led me to contemplate the application of established mathematical techniques from these disciplines to new problems, either by physical or purely formal analogies. In Chapter 2, I adopt a control-theory approach to optimal pulse-shaping in the Laser Wake-Field Accelerator (LWFA). Results in the nonlinear regime are derived from a Hamiltonian formalism due to Pontryagin, while

optimizations in the linear regime rely on manipulations of Fourier integrals and certain function-space inequalities familiar to physicists from quantum mechanics. In Chapter 4, inspired by the well-known parallels between classical paraxial optics and non-relativistic single-particle quantum mechanics, I elucidate a Hilbert-space framework for radiation and an associated maximum-power variational principle (MPVP), applicable to problems involving wiggler, bending-magnet, or other synchrotron radiation from prescribed classical harmonic sources. The MPVP is applicable in more general geometries outside the paraxial regime as well. Despite the formal connections between paraxial optics and quantum mechanics, the MPVP is not as originally suspected an analog of the well-known Raleigh-Ritz variational principle in quantum mechanics, nor is it simply equivalent to the Principle of Least Action that governs classical electromagnetism. In fact, the MPVP is a distinct addition to the many variational principles for optical or electromagnetic problems in general, and wiggler radiation in particular. We continue to pursue several topics of similar theme or temperament, including: a) the use of the paraxial/quantum correspondence to develop a novel derivation of the envelope equations describing the transverse profile of a laser propagating in a weakly inhomogeneous plasma; b) the use of the Wigner function, familiar from semi-classical physics, and ideas from scattering theory to study stochastic EM wave propagation in a noisy plasma subject to local density fluctuations at various length-scales; and c) work toward a universal definition of beam emittance, valid for quantum or classical beams of material particles or radiation, based on notions of statistical entropy and a model of continuous quantum measurement.

Although due to limitations of space no examples are included here, our last major area of research focuses on issues of numerical noise in simulation of a variety of physical systems, based on the growing realization that not all errors are created equal. In the numerical integration of dynamical systems, the standard concerns of classical numerical analysis (truncation and roundoff error, order of accuracy, stability, consistency, and convergence) obviously remain important. However, they are insufficient to address essential notions of fidelity of the numerical approximations to the underlying ordinary or partial differential equations. In many physical problems, these equations possess various types of kinematic, dynamic, geometric, or other structure, which constrains the actual solutions in important ways that generic discretizations cannot capture. I have investigated some ideas for what are known variously as structure-preserving, exactly-conservative, geometric, or "high-fidelity" numerical integration algorithms, which maintain certain underlying structure or invariants of the original dynamical system, such as dynamical or phase space invariants. I have

applied these ideas to the Vlasov equation, the Korteweg-de Vries (KdV) equation, finite-dimensional quantum systems, and three-wave interactions used to model Raman Back Scatter (RBS).

Other ongoing investigations that fit into one or more of these categories, or are informed by similar concerns, include: a) the investigation of stochastic effects in the RBS spectrum; b) investigations into the connections between the accuracy of WKB theory and supersymmetry (SUSY) in single-particle quantum mechanics; c) a study of stochastic particle dynamics in the LWFA, to assess whether RBS plays a role in self-injection and the resulting exponential tail in the electron energy distribution; d) use of Fokker-Planck methods and analogs of SUSY to analyze the movement of so-called hybrid-zones in biological speciation models in evolutionary dynamics; e) an exploration, analysis, and cataloguing of scattering, absorption, and loss mechanisms in low-Z plasmas, in support of plasma-based Raman amplifier experiments; f) investigation into reduced magnetized kinetic transport equations for non-neutral plasmas, with the goal of working toward the simulation of current anti-hydrogen experiments; and g) the use of Wannier function bases for the numerical simulation of possible laser-driven photonic bandgap (PBG) accelerating structures for low-emittance electron beams.

If concerns of uncertainty, noise, and fluctuations in particle beams and laser-plasma systems and issues growing out of these considerations constitute the warp for the fabric of this research, then the woof consists of an emphasis on phase space and its manipulation, monitoring, comparison, or control. Of course, much of applied physics in general, and accelerator physics in particular, could be described as the judicious manipulation of phase space. However, a major thrust of research in the Wurtele group is to look for novel ways to manipulate details of particle or radiation phase space that have not been previously studied.

Hamiltonian and phase space methods have been central to accelerator physics since its inception, but they offer a unified description for beams of material particles or radiation. For particles, spaces of relevance include the full many-body phase space of the Liouville, Gibbs, or Klimontovitch descriptions, the reduced single-particle phase space of the Vlasov or Boltzmann equations, or the related trace space of Courant-Snyder theory for beams in a paraxial limit, as well as the various spaces in between which can support different levels of inter-particle correlations. For radiation, the familiar ray phase space of geometric optics is generalized by the use of Wigner functions to the wave-kinetic phase space of classical wave optics, while in quantum optics each mode (or cell in the wave-kinetic space) is associated

with an abstract quadrature phase space encoding the probabilistic behavior of the creation and annihilation operators for that mode. Phase space is the natural setting for studying laser-plasma or beam-radiation interactions. Light sources, damping, cooling, conditioning, probes, etc., are all conveniently analyzed in terms of phase space couplings and/or phase space overlaps or exchanges.

Many connections between particle and optical phase spaces, both physical and formal, have been long known and well exploited, while others are relatively under-publicized and offer additional opportunities for cross-fertilization. When dynamics are formulated as transport of relevant quantities in phase space (e.g., energy, linear or angular momentum, action, information, probability, etc.), resemblances and dissimilarities between systems are more readily apparent.

A phase space perspective helps to clarify linear versus nonlinear, classical versus quantum, and stochastic versus deterministic aspects of beam structure and evolution. Phase space offers the natural domain for formal mathematical tools and treatments: operator and Lie-algebraic methods; Gaussian, orthogonal function, and wavelet representations; Heisenberg-Weyl, symplectic, or other group actions; and Wigner, Glauber-Sudarshan, Husimi, and other quasi-distribution functions. Phase space treatments better reveal the importance and interplay between dynamic invariants (conserved quantities), kinematic invariants (Casimirs), and symplectic/geometric (Liouville/Poincaré-Cartan) invariants.

Some more specific remarks regarding our particular investigations are in order. In studying pulse-shaping in the LWFA, the overarching goal was to determine which phase space distributions for the driving laser pulses optimize certain figures-of-merit under reasonable operating constraints. The measures are defined in terms of phase-space observables for the plasma supporting the excited Langmuir wave or for the actual particle bunch that is accelerated. Other phase-space related issues arise as we realize that the Hamiltonian density for the plasma wave in the relevant regime is exactly analogous (just with the roles of kinetic and potential energy reversed) to a single degree-of-freedom nonlinear oscillator. We may therefore leverage our intuitions about the more familiar system. Moreover, I make use of an optimization principle from nonlinear control theory which turns out to involve an additional Hamiltonian structure, different from the original physical Hamiltonian, but whose formal symplectic properties we can exploit.

Autoresonance, a property first noticed in a classical nonlinear oscillator, that leads to phase-locking and action growth, has been exploited by analogy in nonlinear optical systems, quantum systems, and now classical laser-plasma accelerators. We take advantage

of autoresonance effects to improve the robustness and performance of the plasma beat-wave accelerator.

In the development of the maximum-power variational principle for classical synchrotron radiation, we discovered that classical particles literally radiate so as to maximize what can be interpreted as a phase-space overlap between a particle distribution and radiation quasi-distribution functions.

In comparing EIT in atomic vapors and magnetized plasmas, correspondences between the classical plasma effect and quantum mechanical atomic version are made manifest by moving to a Hamiltonian/phase-space picture where collective excitations of action density in the two systems appear most similar.

Transit-time OSC has raised questions of how the particle phase space can be most efficiently coupled to, or overlapped with, the radiation phase space, so that the latter can become an effective entropy sink for the former. As it turns out, the question of quantum mechanical limitations on cooling comes down to the shape and form of the quadrature phase-space quasi-distribution functions for the radiation after emission by a particle beam in a wiggler and after amplification by the lasers. A phase space perspective also reveals interesting analogies between the particle beam in OSC and the radiation beam in a Chirped-Pulse Amplification (CPA) system. In OSC, the particle beam is reversibly stretched by dispersion to allow for a more efficient phase space compression (in the form of cooling), followed by recompression. In CPA, the laser beam is reversibly stretches via diffraction to prevent material damage during the driven increase in phase-space density (via amplification), followed by re-compression.

Phase space themes also suggest new research directions, such possible analogies between BGK modes and solitons in noisy optical media, or fundamental entropic/phase-space limits on a proposed idea to use RBS to improve the temporal coherence of FEL radiation in the SASE regime. Phase space techniques offer an overarching framework in which to study the analogies between the classical physics of high-energy storage rings and ultra-cold annular atomic traps.

A third common thread[1] may be discerned in our predilection for inter-disciplinary and cross-disciplinary questions. We have tended to be most interested in issues at the intersections and interstices of traditional fields or sub-fields. Some might legitimately argue that physics at the margins is marginal for good reason, but we have enjoyed learning

[1] This makes no sense in the context of our weaving metaphor, but no matter....

and applying a diverse mix of physics and mathematics, including classical mechanics and electromagnetism, nonlinear dynamics, statistics, statistical mechanics, and stochastic dynamics, neutral and non-neutral plasma physics, accelerator physics, as well as quantum optics, quantum measurement theory, and semi-classical physics.

We have applied tools from nonlinear control theory to laser-plasma accelerators, ideas originating in gravitational wave detectors to optical stochastic cooling, and notions of quasi-particles familiar from condensed matter physics to magnetized plasma dynamics. We have borrowed ideas and concepts liberally from quantum optics, atomic physics, and condensed matter theory, and other fields, but we hope and expect that, in return, techniques and expertise in beam and plasma physics will increasingly find application in astrophysics, photonics, atomic physics, solid state physics, or even biophysics. Many opportunities for cross-fertilization exist, and we suspect many more await discovery.

Within each of the following chapters, I have attempted to provide self-contained introduction and motivation, development of the models, ideas, and results in greater detail than would be allowed within the page limits of peer-reviewed journals. I discuss novel aspects and limitations of the work and assess additional directions that might form the basis for future research. At the end of each chapter, I acknowledge collaboration, suggestions, and advice. In order to avoid redundancy, I acknowledge here the invaluable assistance and guidance of my advisor, Professor Wurtele, since I take it as understood that essentially all of the doctoral research of a graduate student is done under the expert direction and supervision of the thesis advisor. This underlying sense of the collaborative and collective nature of research is a principle source of our decision to use the first person plural tense throughout most of this thesis.

Many sections may be skipped or skimmed by any not wishing so much detailed information or commentary. I have erred on the side of excess of explanation. My intension is to provide background, derivations, reasoning, and results at a level of detail useful to future students within the field, and specifically those within the Wurtele research group. Of a related editorial nature, I have grown somewhat frustrated with the cryptic communications in the pages of *Physical Review Letters* or other prominent journals which are so terse as to practically negate a central and long-standing goal of scientific communication, namely the reproducibility of results. Finally, I have neglected to heed the advice of H.H. Munro (Saki), who wisely pointed out that "a little inaccuracy sometimes saves a ton of explanation," or William Faulkner, who opined that "there is no such thing as good writing, only good re-writing." Instead, I have allowed my own meandering interests to lead me inevitably to

the brink of imposed deadlines, where I can do no more than invoke the famous apology of Pascal:[2] "I have made this too long because I did not have the time to make it shorter."

[2]In a fitting instance of self-reference, we can offer a rather lengthy story behind the attribution for this oft-repeated quotation or its variations. It is widely and more or less correctly attributed to letter of Blaise Pascal (In *Lettres provinciales* 16, 14 December 1656), but the idea is probably much older, going back to St. Augustine, or maybe even to Cicero. More recent attributions, in various forms and paraphrasing, have included Lord Chesterfield, Voltaire, Baroness de Staël, Thomas Jefferson, Henry David Thoreau, Mark Twain, Oscar Wilde, George Bernard Shaw, or even Albert Einstein.

Chapter 2

Optimal Longitudinal Pulse-Shaping in Laser-Wakefield Accelerators

... the will is infinite, and the execution confined,
... the desire is boundless and the act a slave to limit.

William Shakespeare
Troilus and Cressida, Act III, Scene 2

We cannot live better than in seeking to become still better than we are.

Socrates

Argue for your limitations,
and sure enough they are yours.

Richard Bach

2.1 Introduction

The ability of plasmas to sustain very large electric field gradients, together with the advent of chirped-pulse amplification (CPA) techniques[1, 2], have led to serious interest in the use of an intense laser source to excite plasma wakefields for charged particle acceleration. Several concepts have been proposed for using high-power, ultra-short lasers to drive large amplitude, high phase-velocity Langmuir waves suitable for particle acceleration in plasma (see, e.g., [3], and references therein). These include the laser wakefield accelera-

tor (LWFA)[4], wherein a short pulse, of duration of about one-half plasma period or less, ponderomotively excites a plasma wave with phase velocity equal to the group velocity of the laser pulse. In the self-modulated laser wakefield accelerator (SM-LWFA), modulations spontaneously induced by Raman or related modulational instabilities on a somewhat longer laser pulse envelope can resonantly excite the plasma wave[5]. In the plasma beat-wave accelerator (PBWA)[4], the beat between two long laser pulses drives the plasma wave. In the more recently proposed colliding beam accelerator (CBA)[6], a long pump laser is applied to supply additional energy to the wake while a short seed continues to provide the coupling. Each scheme typically also has various possible regimes of operation, depending on the relative sizes of relevant parameters and the dominant scattering processes or nonlinearities involved.

The further ability to manipulate and control the spatio-temporal profile of laser pulses[7] is rapidly progressing, suggesting opportunities for tailoring or shaping the pulse to enhance performance. Ideally, pulse-shaping can lead to larger field gradients with fixed laser power, or given gradients with less laser power, or perhaps to wakefields or accelerated electron bunches with other desirable features. Here we consider issues of how the longitudinal profiles of lasers should be shaped so as to optimize certain figures-of-merit associated with the wake-generation or particle acceleration. That which can increasingly come under experimental control becomes of increasing theoretical concern.

These investigations were originally inspired by work of Spitkovsky, Chen *et al.*[8, 9, 10, 11] on optimization of the so-called transformer ratio in the linear or nonlinear LWFA, because in the "wake" of their interesting but controversial claims, questions of what should be optimized, and under what constraints, remained unclear (at least to us), the original proofs of their results appeared incomplete, and arguments in favor of their pulse-shapes debatable. Despite remaining differences of opinion, as we have achieved greater understanding, we have also achieved somewhat greater consensus, although we continue to disagree with their conclusions on certain fundamental points.

2.1.1 Motivation

In conjunction with CPA systems, various optical and computational techniques are now emerging to precisely time, shape, chirp, and otherwise control the spectral content (or equivalently, the temporal profile) of high-intensity, short-duration laser pulses, as well as nonlinear spectroscopic technologies such as Frequency-Resolved Optical Grating (FROG)

and Spectral Interferometry for Direct Electric field Reconstruction (SPIDER) to measure and recover (with ever improving resolution, as technology improves) the full laser field (i.e., amplitude and phase) at near-infrared or optical frequencies.

It is then natural to turn to questions of how suitably shaped laser pulses can more efficiently or effectively excite such wakefields, or even what pulse shapes can, in some well-defined sense, optimize chosen features of the driven wakefield or of the ensuing particle acceleration. For certain schemes in certain regimes, various answers to these questions have been advocated, but consensus has sometimes remained elusive, so here we add our own ideas and arguments to the discussion, in the hope of working towards better fundamental understanding of how the laser envelope might be specifically tailored to enhance or even optimize desirable features of the wake or characteristics of the wake generation and particle acceleration processes.

2.1.2 Assumptions and Limitations

For the sake of simplicity of formulation and analytic tractability of results, we confine attention to rather highly simplified models of pulse propagation and plasma response. Throughout, we adopt a cold, single-fluid model of the plasma, wherein the heavier ions are assumed to remain motionless on relevant time-scales, merely providing an inert neutralizing background of positive charge for the more mobile electrons, while electrons are assumed to be initially quiescent, with initial velocity and velocity spread assumed to be small compared to the quiver-velocity induced by the lasers and the high group velocity of the laser. Of course, any time we speak of a plasma as a cold fluid we are venturing somewhat into a state of logical sin. As the temperature of the system approaches zero, its behavior inevitably deviates from what we conventionally regard as a plasma. Typically what we really require, and unless otherwise specified what we really mean, is some effective electron temperature which is large enough to ignore inverse Bremsstrahlung or other collisional effects, two-body correlations, and any quantum mechanical effects in particle dynamics or statistics, while small enough to ignore non-collisional Landau damping of the longitudinal motion and the effects of finite thermal spread on quiver or longitudinal fluid response. In addition, here various plasma instabilities which can disrupt the laser fields or plasma wave and limit the useful interaction time are occasionally addressed but mainly ignored.

We treat the electron dynamical response in either the fully linearized limit, valid as the leading-order truncation of a perturbation expansion in the normalized laser strength,

or in the one-dimensional quasi-static regime, which includes the effects of nonlinearities arising from fluid convection and relativistic kinematics, but determines the fast-scale quiver response through canonical momentum conservation strictly valid in the absence of any transverse variation in field or fluid quantities.

Throughout we assume an underdense plasma (where the plasma density is far below the critical density, or equivalently where the laser carrier frequency lies far above the cutoff) and mostly an eikonal laser field, in which a separation of scales is typically possible between the fast carrier oscillation and associated transverse electron quiver motion, and the slower-scales associated with the ponderomotive forces arising from the laser envelope or the resulting plasma oscillation, although in the one-dimensional geometry explicit averaging of the fluid equations is not actually necessary.

We do not pretend to achieve a fully self-consistent treatment. We consider only the cases of linearized or nonlinear quasi-static plasma response in the presence of *prescribed* laser fields – i.e., no diffraction, dispersion, or depletion effects or back-action on the laser are considered, or other nonlinearities or instabilities that might distort the pulse. By considering prescribed laser fields, the problem of optimal pulse-shaping may be fruitfully approached from the point-of-view of control theory or dynamic programming, where the pulse envelope plays the role of the dynamic control field to be determined.

In order to ignore diffraction, we mostly consider propagation in the one-dimensional limit, although we also briefly consider the case of linear propagation of a paraxial pulse in a matched channel. In the assumed under-dense plasma regime, we allow for a slightly sub-luminal group velocity depending on the background electron density, but otherwise ignore group velocity dispersion, in effect presuming laser pulses characterized by a sufficiently small product of pulse bandwidth and propagation time in a sense made precise below. However, certain optimizations will converge to narrow, impulsive envelopes which may suffer significant dispersion, so some care is in order.

Neglect of depletion and nonlinear distortion effects is perhaps the most problematic. We find ourselves in the curious situation where our assumptions tend to be increasingly invalidated to the extent that our goals are achieved. Obviously, deposition of a significant amount of energy in the wake (or ultimately a particle bunch) requires an equal amount of energy be removed from the laser pulse. The production of a large wake implies a large density gradient, eventually invalidating linear theory and leading to an effective nonlinear refractive index experienced by the beam. High efficiency in transferring energy from pulse

to wake implies that depletion or other back-reaction on the laser must ultimately become important.

To a certain extent, the results developed here are being superseded by recent technological advances: most experimental initiatives in plasma-based electron acceleration now involve either the so-called laser-wakefield accelerator (LWFA) scheme in a strongly nonlinear regime for both plasma and laser, where electrons are ponderomotively expelled from the high-field region, and the short laser pulse experiences significant depletion, or the self-modulated laser-wakefield accelerator (SM-LWFA) scheme, where modulations on a longer laser pulse envelope induced by Raman Backscatter resonantly excite the plasma wave. In either case, the laser driver experiences significant back-action, distortion, or depletion that cannot be entirely neglected.

Nevertheless, while the present results are confined to the case of prescribed, non-evolving laser pulses and quasi-static electron response, we believe they may remain of some potential interest. However, we begin with the simplest models because their relative tractability enables clarification of the logical details and physical principles involved, the results may be sufficiently accurate for certain experimental arrangements, and their consideration will help point the way to an appropriate framework for handling more complete physics. They might provide a useful starting point and framework from which to address the issues of optimal pulse-shaping more broadly and eventually within more realistic models. Our analysis attempts to bring some measure of clarity to the questions of what should be optimized in laser-plasma accelerators, and why — issues over which some disagreement has swirled over the last decade. In making use of the Pontryagin Maximum Principle, we give some exposure to a very useful mathematical tool that is familiar in fields of operations research and control theory, but deserves wider advertising to physicists. We correct some false claims in the existing literature, and provide, we believe for the first time, complete proofs of certain other conjectures. Lastly, the results may be sufficiently accurate to guide certain experimental designs, if perhaps not precisely predict quantitative behavior, especially because, as issues of beam quality and control rather than sheer energy begin to be emphasized in second or third-generation experiments, and multi-pulse or other injection schemes are explored, it may actually prove desirable, at least temporarily, to scale back from the highly nonlinear regimes in order to limit self-injection.

2.1.3 Overview

Laser driven wakefield acceleration (LWFA) schemes can potentially produce large accelerating gradients, raising the question of how to optimize, in some well-defined sense, the exchange of energy from the laser to the particle beam via the plasma. Here we consider the longitudinal shaping of the driving laser pulse, and clarify how the optimal pulse shapes and properties can depend critically on precisely what figure-of-merit is used and what additional constraints are imposed. In particular, we carefully define a number of different transformer ratios that have been invoked in the beam physics literature, relate them to other notions of efficiency and energy gain, and then address the ongoing debate concerning whether and when to maximize a transformer ratio versus the wakefield amplitude.

After a detailed discussion of ideas and opinions regarding various optimization criteria and relevant constraints, we conclude in favor of maximizing the wakefield amplitude, subject to one or more limitations on laser energy, intensity, bandwidth, and duration, rather than the transformer ratio. Closely paralleling earlier work on the PWFA[12, 13], the studies of Spitkovsky, Chen *et al.* [8, 9, 10, 11] focus on pulse-shaping so as to optimize a particular version of the transformer ratio, and it is suggested, claimed, or hinted to various degrees that their solutions purporting to maximize transformer ratio also achieve large accelerating gradients and high efficiency in the transfer of energy from driving source to accelerated beams. We adopt a somewhat different point of view and arrive at rather different conclusions, arguing that: optimization of the acceleration gradient or wake energy on the one hand, or of certain measures of the efficiency of the energy transfer or of driver-beam loading the other, tend to be conflicting rather than compatible goals. Any figure of merit based only on efficiency, or any measure which is largely or entirely insensitive to the absolute magnitude of the accelerating gradient achieved, will tend, without further constraints imposed, to improve only at the expense of the peak gradient produced, or equivalently, at the expense of the interaction length required to achieve a certain amount of energy gain.

Also, we argue that in what still remain essentially preliminary stages of research into and development of LWFA scheme and systems, the primary prerequisite is large energy gains over short distances, so it is reasonable to focus preferentially on the excitation of large wake fields and largely to defer questions of efficiency in the generation of the wakes.

Furthermore, although various parallels can be drawn between the LWFA and PWFA schemes, because the physics which governs the laser-driven and electron-driven wakefield excitation and limit the driver-plasma coupling are analogous in several respects, but do

differ in important details, in ways which perhaps call into question the usefulness of optimizing the transformer ratio in the LWFA.

So we argue for and mostly use the peak longitudinal electric field as the principle figure-of-merit, primarily because it is the possibility of high acceleration gradients that most strongly motivates plasma-based accelerator research in the first place. Additionally, when the acceleration length is limited by diffraction or de-tuning length-scales of the electrons in the laser-driven wake (as is typical), maximizing the peak electric field is equivalent to maximizing the achievable gain in particle energy. The gain actually achieved in particle energy is a more complicated story, which cannot really be answered without tracking particle orbits.

With these consideration in mind, we concentrate on the case of cold, quasi-static, collisionless plasma dynamics without self-consistent back-reaction on the laser driver, because: its relative simplicity and analytic tractability will allow us to focus on important logical and physical aspects of the problem rather than mathematical and computational details; it may be of sufficient accuracy to be of interest in certain short-interaction-length LWFA experiments; and despite or even because of its limitations, it might help point the way to a framework for handling more realistic dynamics in a more complete nonlinear and self-consistent treatment.

Because dynamical evolution of the laser pulse plays an essential role, the laser-plasma acceleration in the self-modulated regime will be discussed only briefly in passing. Improving performance in plasma beat-wave acceleration involves its own unique considerations, which will not be discussed here, although see Chapter 3 for some of our ideas in this direction. So we will mainly treat the case of longitudinal pulse-shaping for the short pulse or multi-pulse LWFA, along with some less detailed analysis of and comparisons to the CPA.

One-dimensional linear and nonlinear theory in a cold, underdense homogeneous plasma can be treated analytically in the case of a prescribed laser envelope; linear dynamics in a simplified channel can also be handled analytically, while fully nonlinear and self-consistent evolution of the laser pulse and wakefield is simulated numerically in a 1D PIC code. Further PIC simulations and fluid simulations in 1D and 2D are ongoing.

Using analytic scalings and some PIC simulations, we illustrate optimal wake generation in the LWFA under constraints of maximum laser energy, intensity, and bandwidth, and compare and contrast these pulses with those optimizing the transformer ratio under similar constraints. We demonstrate that pulse shapes optimizing natural measures of the size of

the wake in the LWFA do differ markedly from those maximizing the transformer ratio, and point out the potential merits and pitfalls of both approaches. We then compare the optimized LWFA to the CBA, finding that while the addition of a suitably de-tuned pump can increase the wake achieved by a single short pulse, the CBA is inferior to a single-pulse or multiple-pulse LWFA of identical total laser energy. Finally, we offer some tentative conclusions and suggest some directions for future research.

2.2 Plasma Wake Dynamics: Cold 1D Analytic Theory

For details and derivations of the quasi-static dynamical model[3] or its linearized limit, see Chapter 3 or the references therein. Here we briefly review the principle assumptions and results.

Except for a brief excursion into the case of a linear two-dimensional channel, we work throughout in one-dimensional geometry, where all particle and field quantities are assumed to vary only in the longitudinal (z) direction. The plasma is assumed infinite, cold, collisionless, underdense, unmagnetized, and initially homogeneous and quiescent, with stationary ions.

More specifically, we assume: $\beta_{\text{th}} \ll \beta_p$, where $\beta_{th} = v_{\text{th}}/c$ is the normalized RMS thermal velocity of the electrons, $\beta_p = \beta_g = v_g/c$ is the normalized phase velocity of the plasma wave, which with our approximations will remain equal to the normalized group velocity of the laser envelope, and c is the speed of light *in vacuo*; $m_e \ll m_i$, where m_e and m_i are the electron and ion rest masses, respectively; $\omega_p \ll \omega_0$, where ω_0 is the characteristic laser carrier frequency, and

$$\omega_p = \sqrt{\frac{4\pi n_0 e^2}{m_e}} \tag{2.1}$$

is the electron plasma frequency in Gaussian units in which e is the magnitude of the electron charge, and n_0 is the background density, which is constant for a homogeneous plasma, or more generally can be considered slowly-varying (compared to the laser wavelength); and $\nu_e \ll \omega_p$, where ν_e is the effective collision (Coulomb scattering) frequency. Some further consistency conditions will be formulated as we proceed.

We suppose that both the laser drive and therefore the resulting wake propagate to the right (in the $+\hat{\boldsymbol{z}}$ direction), i.e., $\beta_p = \beta_g > 0$. To be useful for acceleration of relativistic particles, the induced plasma wave should have a high phase velocity, i.e., $0 < 1 - \beta_p \ll 1$.

2.2.1 Scaled Variables

For convenience, we introduce a number of scaled variables: in addition to the usual relativistically-normalized fluid velocity $\boldsymbol{\beta}$, energy γ, and momentum $\gamma\boldsymbol{\beta}$, we adopt dimensionless coordinates co-moving with the group velocity $v_g = c\beta_g$ of the driver pulse (or equivalently phase velocity of the Langmuir wave), i.e., a scaled time $\tau = \omega_p t$, and a scaled co-moving position $\xi = \omega_p t - k_p z$, with the time t and longitudinal position z measured in CGS units in the rest frame of the plasma, and where we have introduced the linear plasma wavenumber $k_p = \frac{\omega_p}{v_g}$.

We also employ a normalized scalar potential $\phi(\xi,\tau) = \frac{e}{m_e c^2}\Phi(z,t)$ associated with the longitudinal plasma wave, and a normalized vector potential $\boldsymbol{a}(\xi,\tau) = \frac{e}{m_e c^2}\boldsymbol{A}(z,t)$ associated with the laser field, both defined in the Coulomb (i.e., transverse) gauge. In this geometry and gauge, the vector potential is also geometrically transverse, i.e., $\hat{\boldsymbol{z}}\cdot\boldsymbol{a}(\xi,\tau) = 0$.

In these scaled co-moving coordinates, the scaled electric field is equal to the positive gradient of the normalized potential:

$$\mathcal{E}(\xi,\tau) \equiv \frac{\beta_g}{E_0}E_z(z,t) = \frac{\partial}{\partial\xi}\phi(\xi,\tau), \tag{2.2}$$

where $E_0 \equiv \frac{m_e c\omega_p}{e}$ is known as the cold, linear wavebreaking limit. With this choice of coordinates, note that a positive potential gradient $\phi'(\xi) > 0$ corresponds to what we will term a negative or decelerating longitudinal electric field, i.e., to a force on an electron or other negatively-charged particle in the $+\hat{\boldsymbol{\xi}}$ (or $-\hat{\boldsymbol{z}}$) direction, opposite to the direction of laser propagation and of the wake phase advance. Obviously a positron bunch traveling in an electron plasma would be forced oppositely.

2.2.2 Some Conventions for Delta and Step Functions

Throughout our discussions, it will frequently prove useful to consider mathematically convenient if physically unrealistic idealizations involving discontinuous changes in or impulsive contributions to the laser source, i.e., sources involving Heaviside step functions and their derivatives, Dirac delta-functions. The latter can be defined either as a kernel of the valuation functional over some set of sufficiently well-behaved test functions in the theory of distributions, or in the sense of a limit of some parameterized family of piecewise continuous functions (such as Gaussians or Lorentzians, or triangular or square pulses, etc.), where strictly speaking the limit is to be taken after any integrations involving the impulse are performed. In either case, the meaning of the product $\delta(\xi)\, f(\xi)$ is only well defined

when integrated over a non-vanishing interval whose endpoints do not coincide with the singularity at the origin, and when $f(\xi)$ is at least continuous within some neighborhood of the origin. In fact, all of the usual properties of the Dirac delta function follow from the requirement that

$$\int_{-\infty}^{\infty} d\xi\, \delta(\xi) f(\xi) = f(0) \tag{2.3}$$

for any function $f(\xi)$ which is continuous in the neighborhood of the origin. The result is not defined if $f(\xi)$ is discontinuous at $\xi = 0$, or if one of the end points coincides with the origin. But often it is convenient to adopt some additional convention about the "skew" of the delta function to resolve this ambiguity, or in other words, assign a meaning to integrals which terminate at the point of singularity, or involve test functions with an additional discontinuity there.

In particular, we will use the notation $\delta_\mu(\xi) = \frac{d}{d\xi}\Theta_\mu(\xi)$ to denote a one-dimensional Dirac delta function for which we take

$$\int_{-\infty}^{0} d\xi\, \delta(\xi) = \mu \tag{2.4}$$

for some μ such that $0 \le \mu \le 1$, so the corresponding Heaviside step function is assumed to satisfy

$$\Theta_\mu(\xi) = \int_{-\infty}^{\xi} d\xi'\, \delta_\mu(\xi') = \begin{cases} 0 & \text{if } \xi < 0; \\ \mu & \text{if } \xi = 0; \\ 1 & \text{if } \xi > 0. \end{cases} \tag{2.5}$$

Because $\delta(-\xi) = \delta(\xi)$ as normally defined, and

$$\int_{-\xi_1}^{+\xi_1} d\xi\, g(\xi) = 2\int_{-\xi_1}^{0} d\xi\, g(\xi) = 2\int_{0}^{\xi_1} d\xi\, g(\xi) \tag{2.6}$$

for any conventional function $g(\xi) = g(-\xi)$ that is even and integrable, it is often simply presumed that the symmetric choice $\mu = \frac{1}{2}$ is the only consistent one, but this does not follow necessarily from (2.3). This symmetric convention is often the most natural, as in the theory of Fourier series or integrals, where expansions of (periodic or rapidly-decaying) piecewise-continuous functions converge to the average of their one-sided limits at any point of discontinuity, but in the context of initial-value problems where causality is important, the choices $\mu = 0$ or $\mu = 1$ is sometimes more convenient.

For simplicity, we will here use the usual notation $\delta(\xi) = \frac{d}{d\xi}\Theta(\xi)$ without subscripts to denote the symmetric ($\mu = \frac{1}{2}$) convention. Our $\delta_0(\xi)$ corresponds to what Spitkovsky and Chen[9] referred to as "half a delta function."

Speaking of one-sided limits, we will also refer to certain "times" (or values of ξ) strictly before, within, or after, some source is applied or changed, possibly discontinuously. We will use subscripts to indicate whether points should be approached in the sense of a limit from above or below (or before or after). For example, we will use shorthand such as $\phi(\xi_1^+)$ to denote the one-sided limit $\lim_{\epsilon\to 0^+} \phi(\xi_1 + \epsilon)$ from above, or similarly $\phi(\xi_1^-)$ to denote the one-sided limit $\lim_{\epsilon\to 0^-} \phi(\xi_1 + \epsilon) = \lim_{\epsilon\to 0^+} \phi(\xi_1 - \epsilon)$ from below.

With this notation in place, we can deduce that

$$\int_{-\infty}^{0-} d\xi\, \delta_\mu(\xi) = 0, \tag{2.7a}$$

$$\int_{-\infty}^{0+} d\xi\, \delta_\mu(\xi) = 1, \tag{2.7b}$$

and also

$$\int d\xi\, \delta_\mu(\xi) f(\xi) = \mu f(\xi^-) + (1-\mu) f(\xi^+) \tag{2.8}$$

for any function $f(\xi)$ which is piecewise continuous in a neighborhood of the origin.

2.2.3 The Quasi-Static Approximation

In the Quasi-Static Approximation (QSA), we assume that the plasma fluid response is essentially independent of time τ in the co-moving frame, implying that the plasma wave itself moves without appreciable dispersion at a fixed velocity equal to the group velocity $\beta_g c$ of the driving laser. Validity of the QSA also specifically requires that any distortion or depletion of the laser envelope remains negligible during the typical interaction time with any one transverse slice of plasma, but we have already assumed that laser envelope evolution is negligible (except for overall translation) throughout the entire interaction.

In the QSA, all τ derivatives in equations describing the plasma response can then be neglected. If we further assume a sufficiently tenuous plasma, or equivalently a sufficiently high phase-velocity Langmuir wave, i.e., one for which

$$\gamma_g \equiv \left[1 - \beta_g^2\right]^{-1/2} \sim \tfrac{\omega_0}{\omega_p} \gg 1, \tag{2.9}$$

and take the plasma to be initially unperturbed ahead of the at least weakly-localized laser pulse, i.e.,

$$\lim_{\xi\to-\infty}\frac{\partial^\ell}{\partial\xi^\ell}a(\xi,\tau)=\lim_{\xi\to-\infty}\frac{\partial^\ell}{\partial\xi^\ell}\phi(\xi,\tau)=0 \tag{2.10}$$

for all integers $\ell\geq 0$, the quasi-static electron response is governed by the simple equation of motion

$$\frac{\partial^2}{\partial\xi^2}\phi=\frac{1}{2}\left[\frac{1+a^2}{(1+\phi)^2}-1\right], \tag{2.11}$$

describing coherent nonlinear Langmuir waves with ultra-relativistic phase velocities, and where $a^2\equiv|\boldsymbol{a}|^2=\boldsymbol{a}\cdot\boldsymbol{a}$ is referred to as the normalized wave action density, or more often the normalized laser intensity, although it is not really equal to either except in the sense of a slowly-varying envelope approximation. The RHS of (2.11) represents both the electrostatic restoring force and the ponderomotive driving force, with full relativistic and quasi-static hydrodynamic nonlinearities included in a high group-velocity limit. The achievable amplitude of the wave is limited by $\inf_\xi[\phi]>-1$ within the present approximations. It is possible to show that for a freely oscillating wave (i.e., after the drive is turned off), the maximum possible longitudinal field is limited to the so-called cold, nonlinear (or relativistic) wavebreaking limit

$$E_z(\xi)\leq E_{\rm WB}\equiv\sqrt{2(\gamma_g-1)}E_0, \tag{2.12}$$

although if the phase velocity is sufficiently high, such that $\gamma_g\beta_{\rm th}^{1/2}\gg 1$, in a warm fluid theory even rather moderate plasma electron temperatures can reduce the effective wavebreaking field noticeably.

Because the forcing in not purely additive, note carefully the possibility of a nonlinear "synergy" in the effectiveness of the ponderomotive drive. For a laser of given strength a^2, the ponderomotive effects are magnified if applied at points where $(1+\phi)^2$ is already small, or in other words where the potential is as negative as possible.

After a^2 is turned off, the nonlinear Langmuir waves described by (2.11) oscillate freely with a characteristic amplitude and fundamental period, but as this amplitude is made larger, the $\phi'(\xi)$ exhibits an increasingly pronounced "sawtooth" pattern associated with wavefront steepening due to higher harmonic content, and the nonlinear period increases due to relativistic nonlinearities.[1] The waveform can be solved for analytically in terms of Jacobi Elliptic functions, but its exact form will not be needed here.

[1]It is a well known yet still somewhat surprising fact that in the cold, nonrelativistic, 1D fluid theory, neither the convective Eulerian nonlinearity in the momentum equation nor the nonlinearity in the continuity equations leads to changes in the frequency of plasma oscillations, although wave steepening still occurs.

The linear plasma wavelength is $\lambda_p = \frac{2\pi}{k_p}$, but in the nonlinear regime, the wavelength grows with the amplitude, and is given approximately by

$$\lambda_{\mathrm{NL}} \approx \begin{cases} \lambda_p & \text{if } \mathcal{E}_{\max} \ll 1 \\ \frac{2}{\pi}\mathcal{E}_{\max} & \text{if } \mathcal{E}_{\max} \gg 1 \end{cases} \tag{2.13}$$

in the low and high amplitude limits, where $\mathcal{E}_{\max} = \sup_{a(\xi)=0} |\mathcal{E}(\xi)|$ is defined as the maximum amplitude of the scaled electric field while the forcing is turned off (or is otherwise negligible).

Contrary to frequent claims and conventional wisdom, the derivation of the quasi-static dynamical equation (2.11) does not itself actually require any *explicit* averaging to remove the fast carrier oscillations in the laser field. However, in order to ignore consistently any τ derivatives in ϕ throughout its evolution, the τ-dependence in the normalized laser intensity a^2 should either be negligible or tend to average away, at least in its effects on ϕ. For this reason, where a separation of scales is possible, a^2 is often replaced with its average over the fast time-scale associated with the carrier oscillation for analytical consistency and convenience, or in certain numerical simulations where the unwanted high-frequencies can cause accuracy or stability problems.

Specifically, in an underdense, quasi-homogeneous plasma, we can write the normalized vector potential in terms of a constant polarization vector, a fast carrier oscillation, and an envelope modulating the carrier oscillation:

$$\boldsymbol{a}(\xi,\tau) = \mu_a \, \hat{\boldsymbol{\epsilon}} \, \tilde{a}(\xi) e^{i\psi_a(\xi,\tau)} + c.c.. \tag{2.14}$$

Here, the polarization vector $\hat{\boldsymbol{\epsilon}}$ is assumed fixed, and satisfies both $\hat{\boldsymbol{\epsilon}}^* \cdot \hat{\boldsymbol{\epsilon}} = 1$ and $\hat{\boldsymbol{z}} \cdot \hat{\boldsymbol{\epsilon}} = 0$, and specifically may be taken to be, say, $\hat{\boldsymbol{\epsilon}} = \hat{\boldsymbol{x}}$ for linear polarization or $\hat{\boldsymbol{\epsilon}} = \frac{1}{\sqrt{2}}(\hat{\boldsymbol{x}} + i\hat{\boldsymbol{y}})$ for circular polarization; while $\mu_a > 0$ is just an overall positive constant depending on normalization conventions, typically chosen to be $\mu_a = 1$, $\mu_a = \frac{1}{2}$, or $\mu_a = \frac{1}{\sqrt{2}}$. In this chapter we will adopt the latter convention when a definite choice is necessary, since

$$a(\xi,\tau)^2 = 2\mu_a^2 \left|\tilde{a}(\xi)\right|^2 + \mu_a^2 \left[\tilde{a}(\xi)^2 \, e^{2i\psi_a(\xi,\tau)} + \tilde{a}(\xi)^2 \, e^{2i\psi_a(\xi,\tau)}\right] \tag{2.15}$$

in the case of linear polarization, and

$$a(\xi,\tau)^2 = 2\mu_a^2 \left|\tilde{a}(\xi)\right|^2 \tag{2.16}$$

for circular polarization, so in either case setting $2\mu_a^2 = 1$ proves convenient.

The fast carrier phase is given explicitly by

$$\begin{aligned}\psi_a(\xi,\tau) = k_0 z - \omega_p t &= -\tfrac{k_0 v_g}{\omega_p}\xi + \left(\tfrac{k_0 v_g}{\omega_p} - \tfrac{\omega_0}{\omega_p}\right)\tau \\ &= -\tfrac{\omega_0}{\omega_p}\left(1 - \tfrac{\omega_p^2}{\omega_0^2}\right)\xi - \tfrac{\omega_p}{\omega_0}\tau \approx -\tfrac{\omega_0}{\omega_p}\xi + O\left(\tfrac{\omega_p^2}{\omega_0^2}\right),\end{aligned} \tag{2.17}$$

expressed in terms of the central laser frequency ω_0 and wavenumber k_0. Typically, these are assumed to satisfy a dispersion relation like

$$\omega_0^2 = \tfrac{\omega_p^2}{\gamma_\perp} + c^2 k_0^2, \tag{2.18}$$

where $\gamma_\perp$ is the effective relativistic kinematic factor associated with the rapid transverse quiver motion of the electrons in the laser field, and is given by

$$\gamma_\perp = \left[1 + \langle a^2 \rangle\right]^{1/2} = \left[1 + |a|^2\right]^{1/2}, \tag{2.19}$$

although for linear polarization it is normally replaced with the averaged version

$$\gamma_\perp \to \bar{\gamma}_\perp = \left[1 + \langle a^2 \rangle\right]^{1/2} = \left[1 + |\tilde{a}|^2\right]^{1/2} \tag{2.20}$$

when the transverse scale of variation of the laser envelope is also much larger than the laser wavelength $\lambda_0 = \frac{2\pi}{k_0}$, and where $\langle a^2 \rangle$ is defined as the normalized laser intensity averaged over the carrier oscillation. The corresponding normalized laser group velocity is then

$$\beta_g = \tfrac{1}{c}\tfrac{\partial}{\partial k}\omega\big|_{k=k_0} = \tfrac{ck_0}{\omega_0} = \sqrt{1 - \tfrac{1}{\gamma_\perp}\tfrac{\omega_p^2}{\omega_0^2}} \approx 1 - \tfrac{1}{2}\tfrac{1}{\gamma_\perp}\tfrac{\omega_p^2}{\omega_0^2}, \tag{2.21}$$

corresponding to $\gamma_g \approx \sqrt{\gamma_\perp}\frac{\omega_0}{\omega_p}$. In the regime (enumerated below) where dispersion may be neglected, note that this carrier phase is approximately independent of τ, as is the scalar laser envelope $\tilde{a}(\xi)$.

These naive expressions for phase and group velocities are expected to be valid in the linear or at most the moderately nonlinear regimes, and for sufficiently slowly-varying envelopes. Higher-order corrections (in $\frac{\omega_p}{\omega_0}$) to the nonlinear laser group velocity and Langmuir phase velocity have been investigated[2] by a number of authors[14, 15, 16, 17]. Diffraction or other transverse effects can also reduce the longitudinal group velocity, while certain nonlinear propagation effects neglected here can lead to distortions in the driving pulse, although the wake phase velocity still tends to still follow the effective velocity of the peak of a unimodal laser envelope even if re-shaping occurs. However, if a sufficiently large plasma wave is excited, the density perturbations can lead to local enhancements in focusing or

[2]In personal communications with C. Schroeder, we have also learned of his ongoing research into corrections for shorter or more intense pulses, including some transverse effects.

de-focusing and which can then feed back on the plasma wave to further affect its phase velocity.

In frequency-space (conjugate to un-scaled time t), the width of the laser envelope $\tilde{a}$ is characterized by some finite frequency bandwidth $\Delta\omega > 0$ centered near $\omega = 0$, and in wavenumber space (conjugate to the un-scaled position z) by a corresponding spatial bandwidth $\Delta k \approx \frac{\Delta\omega}{v_g}$ centered near $k \approx 0$. For definiteness we will consider $\Delta\omega$ as an RMS rather than FWHM measure, since the former is more sensitive to secondary bumps due to "ringing" that might be present as a results of sharp temporal features in the profile. Assuming the laser is transform-limited, the corresponding temporal duration of the laser pulse is $\Delta t_a \sim \Delta\omega^{-1}$, corresponding to the pulse length $\Delta Z_a \approx v_f \Delta t_a$ in physical space or $\Delta\xi_a = \omega_p \Delta t_a \sim \frac{\omega_p}{\Delta\Omega}$ in scaled co-moving coordinates. Conventionally in the LWFA, the bandwidth is assumed to satisfy the eikonal condition $\Delta\omega \lesssim O(\omega_p) \ll \omega_0$. In actual experimental systems, this has been true and will likely remain so for the immediate future, although continued technological improvements may eventually push ultra-short pulses to lengths of only a few (or one) wavelength, while in seeking optimal pulse-shapes analytically, we will also consider, with certain reservations, shorter envelopes which violate this eikonal ordering. If possible, it is usually simplest to employ circular polarization, for which $a^2(\xi) = |\tilde{a}(\xi)|^2$ exactly without any further averaging or assumptions about the carrier phase or its effective dispersion relation or about the slowness of the envelope.

If L_a is the distance over which particles are ultimately accelerated, then neglect of collisional effects in laser propagation, wake production, and particle acceleration requires $\frac{\nu_e L_a}{c} \ll 1$ as well as $\frac{\omega_p}{\omega_0} \ll \frac{\Delta\omega}{\nu_e}$ in addition to the previously-mentioned condition $\nu_e \ll \omega_p$.

Neglect of transverse effects in the laser itself implicitly requires that $k_0\sigma_a \ll 1$, where σ_a is taken as the RMS laser spot size. (Even if channel or relativistic and/or ponderomotive guiding is imposed, this ordering is still necessary in order to consistently neglect any longitudinal components of the laser electric field, as we have done here.) In order to ignore the effects of transverse laser structure on the plasma, i.e., employ results deduced from transverse canonical momentum conservation, we also assume that $\frac{ca}{\omega_0\gamma_\perp} \ll \sigma_a$, so that transverse excursions of plasma electrons while quivering in the laser field are small, and individual electrons do not experience the transverse inhomogeneities in the laser field. Re-arranging, this is seen to be roughly equivalent to $k_0\sigma_a \ll a[1 + \frac{1}{2}a^2]^{-1}$, which essentially imposes the same limitations as those above for $a^2 \sim O(1)$, or even weaker restrictions for very strong ($a^2 \gg 1$) or very weak ($a^2 \ll 1$) lasers. A careful treatment reveals that

the reduction of the plasma fluid response to 1D requires the somewhat stronger condition $\sigma_a > \lambda_p$.

If $a(\xi)^2$ and $\phi(\xi)$ can both be regarded to be functions of ξ only, then (2.11) can be regarded as a second-order ordinary differential equation (ODE) in the independent variable ξ for the dependent variable $\phi(\xi)$,

$$\phi''(\xi) = \tfrac{1}{2}\left[\frac{1+a(\xi)^2}{(1+\phi(\xi))^2} - 1\right], \tag{2.22}$$

with "initial" conditions

$$\lim_{\xi\to-\infty}\phi(\xi) = \lim_{\xi\to-\infty}\phi'(\xi) = 0, \tag{2.23}$$

and therefore the first integral

$$\phi'(\xi^+)^2 + \frac{\phi(\xi)^2}{1+\phi(\xi)} = \int_{-\infty}^{\xi^+} d\xi' \, \frac{\phi'(\xi')}{(1+\phi(\xi'))^2}\, a^2(\xi'), \tag{2.24}$$

where primes denote differentiation with respect to ξ, i.e., $\phi'(\xi) \equiv \frac{d}{d\xi}\phi(\xi)$ and $\phi''(\xi) = \frac{d^2}{d\xi^2}\phi(\xi)$. In fact, for any initial conditions this system is canonical, and derivable from the Hamiltonian

$$\mathcal{H}(\phi, p; \xi) = \tfrac{1}{2}p^2 + \tfrac{1}{2}\frac{\phi^2}{1+\phi} + \tfrac{1}{2}\frac{a^2}{1+\phi}, \tag{2.25}$$

where ϕ is taken as the generalized coordinate, and $p \equiv \phi'$ is the conjugate momentum. If we are to allow the source to contain Dirac delta function terms, then the simplest way to maintain consistency with the conservation law is to adopt the $\mu = \frac{1}{2}$ convention.

Note that if the source eventually turns off, in the sense that $\lim_{\xi\to\infty} a(\xi)^2 = 0$, then the maximum wake amplitude $\mathcal{E}_+$ "behind" the laser pulse can be deduced from the first integral to be

$$\mathcal{E}_+ = \lim_{\xi\to\infty}\left[\phi'(\xi)^2 + \frac{\phi(\xi)^2}{1+\phi(\xi)}\right]^{1/2} = \int_{-\infty}^{\infty} d\xi' \, \frac{\phi'(\xi')}{(1+\phi(\xi'))^2}\, a^2(\xi'), \tag{2.26}$$

which can be used instead of somehow trying to locate the actual position of this maximum, especially if $a(\xi)$ is only weakly localized.

It may at first appear puzzling that a^2 rather than something like $\frac{\partial}{\partial z}a^2$ appears directly in the wake dynamics, as we expect the longitudinal ponderomotive force $F_{\rm pf}(\xi)$ to be proportional to the gradient of the laser intensity, as it arises from Lorentz forces depending on the transverse magnetic field and the electron quiver velocity induced along the direction of the crossed electric field, forces that do not average away over fast carrier time-scale only

in regions where the field envelope varies spatially. However, the wake equation describes the dynamics of the potential rather than the field, so in effect a spatial integration has already been performed. Also, by assuming an initially quiescent plasma both longitudinally and transversely, we are also presuming that the laser field vanishes at sufficiently remote times, so the QSA wake equation "knows" that the normalized laser intensity must have risen to its current value from zero.

2.2.4 Linearized Limit

If we can assume $|\phi(\xi)| \ll 1$, we may perform Taylor expansions of the RHS of (2.33) in ϕ and rearrange to obtain the linearized plasma response:

$$\frac{d^2}{d\xi^2}\phi + \phi = s(\xi), \tag{2.27}$$

where the source, or driving term is given by $s(\xi) = \frac{1}{2}a^2$ and represents the ponderomotive effects of the laser envelope on the plasma, which have decoupled from the intrinsic space-charge restoring forces in the linear regime. Alternatively, but equivalently, we can obtain the linear dynamics directly from the fluid equations by assuming the normalized laser strength is small, i.e., $a \ll 1$, and then performing a formal perturbation expansion of all fluid quantities in powers of a, yielding

$$\frac{\partial^2}{\partial t^2}\phi + \omega_p^2\phi = \frac{1}{2}\omega_p^2 a^2 \tag{2.28}$$

to lowest nontrivial order. If a is non-evolving in the co-moving frame, i.e., $a = a(\xi)$, then by use of the chain rule this reduces to the scaled co-moving form (2.27). In the cold linear 1D regime, plasma waves of any amplitude below the cold, linear wavebreaking limit E_0 (or $\phi_0' \equiv \beta_g$ in scaled units) can be supported.

Given a prescribed profile for the laser envelope, and still assuming the plasma is quiescent ahead of the laser pulse, the resulting normalized wake potential can be written in terms of a convolution of the ponderomotive source term $s(\xi)$ with an appropriate Green function:

$$\begin{aligned}\phi(\xi) &= \int_{-\infty}^{\infty} d\xi'\, \Theta(\xi - \xi') \sin(\xi - \xi') s(\xi') \\ &= \int_{-\infty}^{\xi} d\xi'\, \sin(\xi - \xi') s(\xi'),\end{aligned} \tag{2.29}$$

and upon differentiation, the scaled force on an electron in this potential may then be written as

$$\phi'(\xi) = \int_{-\infty}^{\xi} d\xi' \, \cos(\xi - \xi') s(\xi'). \tag{2.30}$$

The first integral may be written as

$$\tfrac{1}{2}\phi'(\xi)^2 + \tfrac{1}{2}\phi(\xi)^2 = \int_{-\infty}^{\xi} d\xi' \, \phi'(\xi') s(\xi'), \tag{2.31}$$

where the RHS is proportional to the total work done on the plasma by the ponderomotive drive from the moment it turns on, while the LHS can be shown to be proportional to the total kinetic energy of electron fluid motion plus electrostatic potential energy in the plasma. Expanding in powers of ϕ, the Hamiltonian generating the dynamics simplifies to

$$\mathcal{H}(\phi, p; \xi) = \tfrac{1}{2}p^2 + \tfrac{1}{2}\phi^2 + s(\xi)\,[1 - \phi] + \cdots \tag{2.32}$$

which, as might be expected, reduces at lowest nontrivial order to that of a forced simple harmonic oscillator (SHO). In the wake frame, if the source is turned off by $\xi = \xi_1$, then for $\xi > \xi_1$, the wakefield will simply oscillate sinusoidally at frequency ω_p and wavelength λ_p, with normalized amplitude $\mathcal{E}_{\max} = \sqrt{\phi(\xi_1)^2 + \phi'(\xi_1^+)}$.

2.2.5 Mechanical Oscillator Analogy

These wake dynamics are therefore analogous to those of a mechanical oscillator with external forcing but no damping, where the co-moving coordinate ξ is analogous to time, $\phi(\xi)$ corresponds to to the oscillator displacement from equilibrium, $\phi'(\xi)$ is analogous to the oscillator velocity, and the ponderomotive drive corresponds to the external forcing. Under our assumed initial/boundary conditions, the analogous oscillator is initially at rest at the point of equilibrium (which is taken to be the origin). In the linear regime, the restoring force is harmonic, while in the nonlinear quasi-static regime, the restoring force is nonlinear, and the overall external forcing is also explicitly dependent on "position" in addition to "time." Since $a(\xi)^2 \geq 0$, the external forcing can "push" in one direction only (towards positive ϕ). Continuing the analogy, $\frac{1}{2}\frac{\phi(\xi)^2}{1+\phi(\xi)}$ corresponds to the potential energy of the mechanical oscillator, although using the quasi-static fluid equations, it can be shown to be proportional to the density of relativistic kinetic energy of electrons in the plasma wave. Conversely, the quantity $\frac{1}{2}\phi'(\xi)^2$, equal to the mechanical oscillator kinetic energy, physically is proportional to the electrostatic field energy density of the plasma wave (or

equivalently, to the potential energy of plasma electrons). The first integral is analogous to the mechanical work supplied by the drive to the oscillator, i.e., to the total (kinetic plus potential) energy acquired, if the system begins at rest. The total "envelope energy" of the driving pulse, proportional to $\int d\xi\, a^2(\xi)$, is analogous to the total impulse of the external forcing applied to the mechanical oscillator.

2.3 Laser-Driven versus Particle-Driven Wakefield Accelerators

As is well known, the plasma dynamics in the LWFA are similar to those of the particle-driven wakefield accelerator (PWFA), only with with the ponderomotive force from the laser replacing a source term proportional to the charge density of the drive beam, even though the specific mechanisms of wake excitation by the laser and particle perturbations are quite different.

The similarities arise from the fact that the driving laser field can also be interpreted as a particle beam, only one consisting of photons rather than electrons, while the differences are due to the different nature of the luminous and material particles. In the case of an eikonal (narrow-band) laser in an at most slowly-varying medium, the Planck-Einstein relations suggest that each photon can be taken to have momentum $\hbar \boldsymbol{k}$ proportional to its wavevector $\boldsymbol{k}$ near the central wavevector $k_0 \hat{\boldsymbol{z}}$ and energy $\hbar\omega$ proportional to its frequency $\omega = \omega(k)$ near the central frequency ω_0, assumed to satisfy the effective dispersion relation (2.18). The velocity of each of these quasi-particles then lies near the central group velocity v_g. The normalized laser intensity $a(\xi)^2$ is taken as proportional to the density[3] of photons.

This is analogous in many respects to the case of the PWFA, where the wake is excited by a collimated, quasi-monoenergetic electron drive beam of particles (typically electrons, but positron or other beams are also possible) of rest mass m_d, charge q_d, and with density

[3]In the combined coherent, quasi-monochromatic, eikonal, and paraxial regimes, this interpretation is largely unproblematic. Subtleties emerge in more general cases. For example, the expected photon density for a nearly monochromatic laser, as revealed by ideal normally-ordered photon-counting experiments, will be proportional to the quantum mechanical expectation value $\langle \boldsymbol{A}^-(\boldsymbol{x},t) \cdot \boldsymbol{A}^+(\boldsymbol{x},t) \rangle$, where the annihilation operator $\boldsymbol{A}^+(\boldsymbol{x},t)$ is related to the positive-frequency part (analytic signal) of the Coulomb-gauge vector potential operator, $\boldsymbol{A}^-(\boldsymbol{x},t)$ is its Hermitian conjugate, and the brackets denote quantum mechanical and statistical averaging. However, this is not in general quite the same as the classical normalized field strength $a(\xi)^2 \propto \left| \langle \boldsymbol{A}^+(\boldsymbol{x},t) + \boldsymbol{A}^-(\boldsymbol{x},t) \rangle \right|^2$. Also, well-known difficulties are encountered when one attempts to define a consistent position operator for photons or any other massless Bosons. Whether and how the second quantization of the electromagnetic field can in a sense be reverse-engineered, and either the vector potential $\boldsymbol{A}(\boldsymbol{x},t)$, the Silberstein vector $\boldsymbol{E}(\boldsymbol{x},t) + i\boldsymbol{B}(\boldsymbol{x},t)$, or related fields consistently interpreted as "the wave function of a photon," are topics of current investigation and debate in the literature.

$n_d(\boldsymbol{x}, t)$, moving longitudinally with average per-particle kinetic momentum p_d, average per-particle relativistic energy $\varepsilon_d = \sqrt{m_d c^2 + c^2 p_d^2} \equiv \gamma_d m_d c^2$, and with characteristic velocity $v_d = c\sqrt{\frac{cp_d}{\varepsilon_d}}$. This is reminiscent of the laser-driven case, if we interpret the dispersive effects in the plasma associated with the cutoff by imagining each photon to have acquired an effective "rest mass" given by $m_{\rm ph} \equiv \frac{\hbar\omega_p}{\sqrt{\gamma_\perp c^2}}$, so that the central photon energy can be written $\hbar\omega_0 = \gamma_g m_{\rm ph} c^2$.

In fact in the QSA with prescribed forcing by charged particle and/or optical driving pulses, the plasma response can be generalized to

$$\tfrac{\partial^2}{\partial \xi^2}\phi = -\,\mathrm{sgn}[q_d]\,\frac{n_d}{n_0} + \tfrac{1}{2}\left[\frac{1+a^2}{(1+\phi)^2} - 1\right]. \tag{2.33}$$

In either the photon-driven or electron-driven scenario, assuming a cold neutral plasma with stationary ions, the driving-particle density can act as a source to shock-excite a Langmuir wave which then oscillates behind the driving beam. However, the specific details of the particle dynamics and wake excitation are somewhat different in the two cases. The most obvious difference is that, even without any self-consistent feedback included, the forcing in the case of an applied laser is not purely additive, but also depends nonlinearly on the wake potential, which can be traced back to the transverse quiver induced by the laser.

Also, we tend to think of massive particles as sluggish in comparison to photons, but in plasma-based accelerator applications, both driving and driven electrons can outpace the photons. In a PWFA, the driving electrons (or positrons) have a rest mass energy $m_e c^2 \sim O(0.510\ \mathrm{MeV})$ and can be highly relativistic, with $\gamma_d = O(10^2)$ or $O(10^3)$ or even higher, whereas in a LWFA, the driving photons typically have energies $\hbar\omega \approx O(1\ \mathrm{eV})$ (corresponding to $\lambda_0 \sim O(1\ \mu\mathrm{m})$ in an underdense plasma) and therefore effective relativistic factors of only $\gamma_g \approx \sqrt{\gamma_\perp}\frac{\omega_p}{\omega_0} \sim O(10)$, so *individual* photons are typically much less energetic, both absolutely and relative to their own effective "rest mass," than are the individual electrons.

Of course both longitudinal and transverse dynamics depend strongly on the effective relativistic kinematic factor of the driving particles. In terms of the normalized energy spread in the drive beam, the relative longitudinal velocity spread is given approximately by $\frac{\delta\beta}{\beta} \sim \frac{4}{\gamma^2}\frac{\delta\gamma}{\gamma}$, so dispersive effects are expected to be rather more pronounced in the case of photons. With greater effective inertia, particles with larger γ are generally stiffer or more resistant to perturbations in transverse velocity as well.

For relativistic electrons in the driving beam, the de Broglie wavelength $\lambda_{\rm dB}$ is typically extremely small compared to inter-particle spacings ($\sim n_d^{-1/3}$) and transverse or longitu-

dinal beam dimensions, even in the average rest frame, and betatron actions associated with the transverse motion are typically quite large compared to $\hbar$, so the electrons can be treated completely classically. With quantum mechanical effects negligible for any foreseeable electron drive beam, the initial longitudinal velocity spread is essentially independent of the pulse length or profile. However, despite having acquired an effective mass, the photons in the LWFA are still more wave-like than particle-like in nature. For a longitudinally coherent laser pulse, the photon density profile and spread in photon momenta are related via the Heisenberg complementarity of the Fourier (or Wigner) transform; e.g., a more narrowly-peaked laser profile requires a correspondingly larger bandwidth. Since the plasma dispersion relation has cubic and higher terms in the laser wavenumber k, this uncertainty principle implies that some group velocity dispersion is unavoidable as the pulse duration is decreased. On the other hand, electrons, unlike photons, directly repel each other, so maintaining a small pulse-length does become more challenging at high electron densities.

For the electron beam, any any transverse evolution in the beam envelope is determined by the total fields, including self-fields and those from external beam optics, and the initial classical transverse emittances ϵ_x, ϵ_y characterizing spread in the classical phase space distribution function, which swamp any quantum effects from a finite de Broglie wavelength. Transverse evolution of the beam envelope is dominated by geometric and mechanical, rather than diffractive, effects. Outside the plasma, the radial electric and magnetic self-fields which could cause expansion tend to cancel to $O(\gamma^{-2}) \ll 1$ for a symmetric beam, whereas inside the plasma, the drive beam tends to repel like charges, which then acts to partially neutralize the beam, and the un-balanced magnetic forces can cause plasma lensing,[4] but these self-focusing effects can be compensated if necessary with external solenoid or quadrupole magnetic lenses, or better, exploited and incorporated into the lattice design. In typical applications, with suitable beam optics, the inverse betatron wavenumber κ_β^{-1} determining the characteristic longitudinal length-scale over which the electron envelope undergoes transverse oscillations can be made much longer than the length L_0 of the plasma section.

In the LWFA, the laser pulse is typically strongly focused and maintains a high degree of transverse coherence, so diffractive effects are much more important than in the

[4]In more realistic 3D geometries, at high currents the particle beam in a plasma is also subject to various hosing or other transverse instabilities, while a sufficiently intense laser spot is subject to filamentation, hosing, or other instabilities as well. More or less out of analytic necessity and the hope that they may be avoided in practical applications by suitable operating regimes, such effects are ignored here.

particle-driven case. In the absence of self-focusing (possibly only for sufficiently intense pulses) or suitable density channeling, the Raleigh range $Z_{\rm R}$ (the photonic analog of κ_β^{-1}) is typically much shorter than L_0, as we will see below. Contrary to the electron-driven case, in the nearly-coherent optical case the wave-like, or quantum contributions dominate the emittance.

Also, in the PWFA, the driving particles are of course electrically charged, so the drive beam acts as a charge and current perturbation which can both directly excite electromagnetic fields and directly respond to them. Neglecting the small effects from large-angle collisions, the total number of electrons in the drive beam is conserved as it travels through the plasma, and the particle beam loses kinetic energy as longitudinal electric fields decelerate the individual driving electrons, while this energy is, ideally, deposited in the plasma oscillation persisting behind the density perturbation. In the one-dimensional, electrostatic limit, the gradient of the force on plasma electrons is proportional to the total induced charge density perturbation, via Gauss's law: $\frac{\partial}{\partial z} F_z \sim \delta n$, where δn is just equal to the beam density n_d in the linear regime but may include contributions from additional nonlinear "blow-out" effects for sufficiently high drive currents.

Roughly speaking, and despite the complication of direct self-forces, an electron driving beam of very short duration can be made with small dispersion initially, and even as it loses energy and possibly accumulates energy spread while in the plasma, it can propagate with relatively little distortion of its relative shape, just with a small overall decrease in average velocity.

In the LWFA, the photons do not respond directly to the field, but rather are a field, or specifically crossed electric and magnetic fields, which act directly on the plasma electrons to produce rapid transverse quiver motion, as well as the typically more slowly-varying longitudinal ponderomotive force which is responsible for the wake excitation. To leading order, the ponderomotive force is proportional to the gradient of the laser photon density: $F_z \propto \frac{\partial}{\partial z} a^2$, although additional nonlinear effects will also emerge at sufficiently large intensities.

While the driving photons do not directly respond to the ensuing longitudinal electric fields, they are subjected, via nonlinear feedback, to changes in the background plasma density and fluid velocity; and while photons need not be conserved in general, absorption and large-angle scattering cross-sections in typical underdense laser-plasma interactions are typically small, so that photon number (or equivalently, wave action) is approximately conserved, and energy is primarily transferred from the laser to the plasma wave through

coherent frequency red-shifting, or so-called *photon deceleration*, of all photons in the drive pulse, which to leading order is just proportional to the longitudinal electric field E_z as it would be for charged particles.

Phase-space transport equations describing beam propagation (based, say, on the Wigner function) will be of similar form for the electron or radiation beams. If we do not need to keep track of the phase of the laser pulse explicitly, or of higher-order phase space moments in either case, spatial envelope equations associated with the photon or particle particle density will appear even more similar in mathematical structure, although typically with very different parameters reflecting the characteristic energy and length-scales.

While we should exploit these similarities whenever possible, we should also remain mindful of important qualitative and quantitative differences. Technologies needed to produce, manipulate, or control either electron beams or radiation beams of given density, energy, energy spread, shape, etc. are quite different, and these differences effect ultimate design and performance for accelerators. Conditions for or considerations of optimality for the PWFA do not necessarily transfer to the case of the LWFA, even if we take seriously the corpuscular nature of light.

2.4 Limitations on Acceleration

As we have seen, neglect of diffraction, dispersion, depletion or distortion effects in the laser propagation can at most be valid only within certain regimes. With other parameters fixed, the effects of each of these processes on the subsequent electron dynamics will remain small for propagation distances significantly less than certain characteristic scale-lengths beyond which effective acceleration will be disrupted. Energy gain of any accelerated particle of mass m and charge q is limited to $mc^2\Delta\gamma_a \leq |q| \sup[E_+(z,t)]L_a$, where $E_+(\boldsymbol{z},t) = \Theta_0\left(qE_z(z,t)\right)|E_z(z,t)|$ is the magnitude of the *accelerating* field; and L_a is the effective acceleration length, limited to be the smallest of: the overall length L_0 of the plasma as an "accelerating structure;" the optical diffraction length, which for unguided pulses is given approximately by the Raleigh range, i.e.,

$$L_{\rm diff} \approx Z_{\rm R}\pi\tfrac{\sigma_a^2}{\lambda_0} \sim \pi\lambda_p\tfrac{\omega_0}{\omega_p}\left(1-\sqrt{\gamma_\perp}\tfrac{\omega_p^2}{\omega_0^2}\right) \tag{2.34}$$

for unguided pulses (assuming $k_p\sigma_a \sim O(1)$), while typically $L_{\rm diff} \sim O\left(10Z_R\right)$ or more for short pulses in a channel or for sufficiently long and intense pulses subject to relativistic

and/or ponderomotive self-guiding;[5] the optical dispersion length

$$L_{\rm disp} \sim \lambda_p \frac{\omega_0}{\Delta\omega} \frac{\omega_0^2}{\omega_p^2} \left(1 + \frac{\omega_p^2}{\omega_0^2}\right) \left(1 - \tfrac{1}{2}\frac{\omega_p^2}{\omega_0^2}\right) \left(1 - \tfrac{1}{4}\frac{\Delta\omega^2}{\omega_0^2}\right), \tag{2.35}$$

beyond which the overall length of the pulse envelope is significantly broadened (compared to λ_p) due to group velocity dispersion; the so-called de-tuning or de-phasing length, estimated as the length over which a highly-relativistic particle will slip ahead of the accelerating phase of the wakefield, approximately given by

$$L_{\rm det} \sim \lambda_p \frac{\omega_0^2}{\omega_p^2} \left(1 + \tfrac{1}{2}\frac{\omega_p^2}{\omega_0^2}\right) \tag{2.36}$$

in the linear case ($a_0^2 \ll 1$), and

$$L_{\rm det} \sim \lambda_p \frac{\omega_0^2}{\omega_p^2} \left(1 + \tfrac{1}{2}\frac{\omega_p^2}{\omega_0^2}\right) \frac{2a_0^2}{\pi} \tag{2.37}$$

or

$$L_{\rm det} \sim \lambda_p \frac{\omega_0^2}{\omega_p^2} \left(1 + \tfrac{1}{2}\frac{\omega_p^2}{\omega_0^2}\right)^2 \tfrac{1}{2}\mathcal{E}_+ \tag{2.38}$$

in the highly nonlinear regime ($a_0^2 \gg 1$), where here $a_0^2 = \sup\left[a(\xi)^2\right]$ is taken as the peak normalized laser intensity and $\mathcal{E}_+ = \frac{\beta_g}{E_0}\sup[E_+(z,t)]$ is the magnitude of the maximum scaled accelerating wakefield amplitude behind the laser; the pump depletion length, over which the pump deposits a significant fraction of its energy in the wakefield, estimated as

$$L_{\rm pd} \sim \begin{cases} \lambda_p \frac{\omega_0^2}{\omega_p^2} \frac{1}{a_0^2} & \text{if } a_0^2 \ll 1 \\ \lambda_p \frac{\omega_0^2}{\omega_p^2} \frac{a_0}{3\pi} & \text{if } a_0^2 \gg 1 \end{cases}, \tag{2.39}$$

assuming for simplicity a "resonant" square pulse drive; and finally the growth-length $L_{\rm inst}$ for the dominant (fastest-growing) laser-plasma instability, typically Raman Back-Scatter (RBS). In intermediate regimes, estimates are typically smoothly interpolated between the limiting cases, without much justification, but rather for lack of any better theoretical guidance.

A typical underdense parameter regime for the LWFA in a laser-ionized capillary or gas-jet plasma might be $\lambda_0 = \frac{2\pi}{k_0} \sim O\left(1\ \mu\text{m}\right)$, $L_0 \sim O\left(5\text{ mm}\right)$, $\frac{\omega_p}{\omega_0} \sim O(10^{-1})$, $\Delta\omega \sim O\left(\omega_p\right)$, and $a_0 \sim O\left(10^{-3}\right)$ in the linear regime or $a_0 \sim O\left(1\right)$ in a nonlinear regime. In such typical cases,

$$L_{\rm diff} \ll L_{\rm det} < L_{\rm disp} \leq L_{\rm pd} \leq L_{\rm inst} < L_0 \tag{2.40}$$

[5]Correct treatment of the case of relativistic self-guiding is somewhat complicated, because the head of the pulse may not be strongly guided. Such effects will be ignored in in our one-dimensional treatment here.

without guiding by a channel or other means, while usually

$$L_{\rm det} < L_{\rm disp} \leq L_{\rm pd} \leq L_{\rm diff} \leq L_{\rm inst} < L_0 \tag{2.41}$$

in a suitable plasma channel.

Note that overall energetic efficiency in accelerating a single bunch will always be limited to less than unity if $L_a \leq L_{\rm det} < L_{\rm pd}$.

Based on these simple scaling laws and 1D XOOPIC Particle-In-Cell (PIC) simulations, various length-scales for a "resonant" (FWHM equal to a half-plasma period) Gaussian laser pulse with $\sigma_a \sim \lambda_p$ and $\frac{\omega_p}{\omega_0} = 0.1$ are roughly:

a_0	$L_{\rm diff}$ (theory)	$L_{\rm det}$ (theory)	$L_{\rm pd}$ (theory)	$L_{\rm pd}$ (PIC)	$L_{\rm disp}$ (theory)	$L_{\rm disp}$ (PIC)
0.001	$3.1 \cdot 10^1\,\lambda_p$	$1.0 \cdot 10^2\,\lambda_p$	$1.0 \cdot 10^8\,\lambda_p$	$5.2 \cdot 10^5\,\lambda_p$	$1.0 \cdot 10^3\,\lambda_p$	$4.5 \cdot 10^2\,\lambda_p$
0.30	$3.1 \cdot 10^1\,\lambda_p$	$1.0 \cdot 10^2\,\lambda_p$	$1.1 \cdot 10^3\,\lambda_p$	$4.3 \cdot 10^4\,\lambda_p$	$1.0 \cdot 10^3\,\lambda_p$	$1.3 \cdot 10^3\,\lambda_p$
1.00	$3.1 \cdot 10^1\,\lambda_p$	$1.0 \cdot 10^2\,\lambda_p$	$1.0 \cdot 10^3\,\lambda_p$	$6.2 \cdot 10^2\,\lambda_p$	$1.0 \cdot 10^3\,\lambda_p$	$3.8 \cdot 10^2\,\lambda_p$

Obviously 1D simulations cannot capture diffractive effects, and simulations of the laser and background plasma alone without the accelerated beam cannot independently determine de-phasing lengths, so no numerical estimates for the those categories are presented. Reasonable agreement is seen between theoretical scalings and PIC-based numerical estimates for the dispersion lengths, while correspondence between the pump depletion lengths is rather poor for the lower values of a^2. But in any case, it would seem that acceleration will be limited by diffraction without guiding and by de-tuning if diffraction is overcome.

As mentioned, effective diffraction lengths can be increased by the use of pre-formed density channels[5] or possibly the additional of carefully-phased plasma waves, or by nonlinear self-guiding effects for sufficiently intense pulses, or possibly the deliberate tailoring of the pulse shape[3]. Dispersive effects can be reduced by resorting to longer pulses or operating at lower densities, while nonlinear effects at very high laser intensities might also be able to be exploited. De-tuning lengths can also be increased by working at lower densities if possible, or adding a static transverse magnetic field, as in the surfatron, which however continuously deflects the bunch from a linear trajectory, so is not very practical for laser-plasma applications. Tapered channels might also be used to gradually shift the plasma wave phase velocity and extend the de-tuning length[18, 19]. Otherwise multiple plasma stages might be used with carefully delayed drive pulses to achieve greater gain than allowed by de-tuning in a single wave. Growth rates and saturation levels of various instabilities will also depend on pulse length as well as plasma densities and temperatures.

Typically growth rates are safely longer than other relevant length-scales for pulses of length $\Delta z \lesssim \lambda_p$. Generally (although not always), growth rates tend to grow with density, but saturation levels might decrease. Warmer temperatures can increase the amount of initial noise from which Raman instabilities start to grow, but tend to inhibit subsequent growth.

Some Limitations on the Predicted Limitations

The above scalings laws should be regarded as providing rough estimates, but determining more accurate values might require solution of the equations of motion for the laser, the plasma, and possibly the injected electron beam, ideally with at least some transverse effects included. In almost all cases, diffraction will limit the useful acceleration in the absence of guiding, but here in our simplified one-dimensional analysis we will mostly ignore such effects without explicitly specifying how they are to be counteracted, except for some basic analysis of the linear channel.

The usual ordering $L_{\text{det}} \ll L_{\text{disp}}$ is calculated assuming an underdense plasma and a near-"resonant" laser pulse where $\Delta\omega \sim O\left(\frac{1}{2}\omega_p\right)$, but shorter pulses (including some of the optimal solutions found below) exceed the resonant bandwidth. But according to the scaling laws, the detuning length remains smaller than the dispersion length provided

$$\frac{\Delta\omega}{\omega_0} \lesssim \begin{cases} 1 & \text{if } a_0^2 \ll 1 \\ \frac{\pi}{2}\frac{1}{a_0^2} & \text{if } a_0^2 \gg 1 \end{cases}, \tag{2.42}$$

so for weak laser pulses, dispersive effects should remain essentially negligible over distances relevant to particle acceleration down to pulse lengths containing only a few optical wavelengths, but will become increasingly important for more intense pulses because larger-amplitude wakes have longer nonlinear plasma wavelengths and therefore longer detuning lengths over which particles could be accelerated. For given envelope energy, narrow intense pulses converging to Dirac delta functions eventually violate this assumption, because $a_0^2 \propto \Delta\omega \to \infty$ so that $\frac{L_{\text{disp}}}{L_{\text{det}}} \propto \Delta\omega^{-2} \to 0$. However, in this limit we also expect $L_{\text{pd}} \propto a_0^{-2} \propto L_{\text{disp}}$ even into the nonlinear regime, so that $\frac{L_{\text{pd}}}{L_{\text{det}}} \to 0$ as well, raising the possibility that sufficiently weak impulse-like sources might be able to deplete before they disperse. Unfortunately, only consistent evolution of the coupled laser and plasma dynamics can really answer this question.

The scaling for the detuning length is better expressed in terms of the peak wakefield $\mathcal{E}_+$ behind the pulse than in peak laser strength, because the former is more directly related to the nonlinear plasma wavelength λ_{NL}, and practically the entire *raison d'être* for this

investigation resides in the fact that wakes with very different values for $\mathcal{E}_+$ can be excited by appropriately-shaped laser pulses with the comparable a_0^2.

Likewise, the simple scaling law used above for the depletion length ignores the essential fact that laser pulses of the same peak intensity that produce larger wakes must deplete faster. One expects a close relationship between maximizing the energy in the wake or the maximum amplitude of the wake and minimizing the actual depletion length.

Of course, without any back-action on the laser source actually included in the dynamics, no depletion effects can appear explicitly, but rather they are inferred through energy conservation and knowledge of the wake field. Obviously this approach is not entirely self-consistent, its accuracy cannot be determined without appealing to a more consistent one, and the approximations involved are only expected to worsen as the depletion length decreases. Nevertheless, it offers the logical first step on the way to a more systematic and self-consistent analysis.

With any scattering and absorption cross sections contributing to wake generation remaining negligible[6] for paraxial lasers of realistic intensity traveling through underdense plasmas, the energy lost to the wake must manifest predominately in the form of frequency down-shift of the laser, as mentioned previously. From quasi-static energy conservation, the corresponding rate of photon deceleration can be determined to be approximately

$$\frac{\partial}{\partial \xi}\left[\frac{\omega}{\omega_p}\right] \approx -\frac{1}{2}\frac{1}{\gamma_\perp}\frac{\omega_p}{\omega_0}\kappa(\xi) \tag{2.43}$$

in scaled units, where, following the literature, we define the wake-dependent part of the deceleration as

$$\kappa(\xi) \equiv -\frac{\partial}{\partial \xi}\frac{1}{1+\phi(\xi)} = \frac{\phi'(\xi)}{[1+\phi(\xi)]^2}. \tag{2.44}$$

for all points ξ within the support of $a(\xi)$.

In terms of this photon deceleration, the depletion length can then be estimated as the distance over which the maximally decelerated slice of the laser pulse redshifts down from the carrier frequency ω_0 to an effective cutoff frequency $\frac{\omega_p}{\tilde{\gamma}_\perp}$:

$$L_{\mathrm{pd}} \sim \frac{\tilde{\gamma}_\perp \frac{\omega_0^2}{\omega_p^2} - 1}{k_p \sup\limits_{a(\xi)\neq 0}[\kappa(\xi)]}. \tag{2.45}$$

To estimate $\tilde{\gamma}_\perp$, one approach is probably to use the value $\tilde{\gamma}_\perp = \gamma_\perp[a(\tilde{\xi})] = 1+\left\langle a(\tilde{\xi})^2 \right\rangle$ at the position $\xi = \tilde{\xi}$ corresponding to the maximum of $\kappa(\xi)$, although this obviously ignores the

[6]Raman or other scattering processes can remove additional energy from the laser, but within the QSA will not affect the relative book-keeping between laser and high-phase velocity wake energy of relevance here.

fact that the separation-of-scales implicit in the derivation of $\gamma_\perp$ breaks down as $\omega_0 \to \omega_p$. The other extreme is to just use $\tilde{\gamma}_\perp = 1$. For small a^2, it does not much matter, but for large a^2 we would need a better theory of nonlinear pulse propagation near the cutoff. In any case, the depletion length is expected to be inversely proportional to the rate of photon deceleration.

2.5 Analytic Optimization Techniques

At least when back-action on the drive laser is ignored, questions of optimal pulse-shaping can be most fruitfully approached from the point of view of dynamical control theory, quite familiar to mechanical and industrial engineers but somewhat less so to physicists.

In the *linear* regime, standard Calculus of Variations can be used, along with familiar function-space inequalities (Cauchy-Schwarz, Holder, etc.), as well as arguments based on the mechanical oscillator analogy. In the *nonlinear* regime (or *a forteriori* the linear regime as well), certain analytic results can be obtained by employing *Pontryagin's Maximum Principle*[20, 21], which converts a functional optimization problem in so-called Bolzano form into an analogous Hamiltonian dynamical problem and a scalar function optimization.

Specifically, suppose we seek to minimize the cost functional

$$\mathcal{J}[\boldsymbol{m}] = \int_{t_0}^{t_1} dt\, F\big(\boldsymbol{x}(t), \boldsymbol{m}(t)\big), \tag{2.46}$$

in which t is the evolution parameter, or independent variable (referred to for simplicity as "time," although other interpretations are of course possible), $\boldsymbol{x}(t) \in \mathbb{R}^n$ is the n-dimensional state vector defined over some interval $[t_0\,,\, t_1] \subseteq \mathbb{R}$, and $\boldsymbol{m}(t) \in \mathcal{M}[t] \subseteq \mathbb{R}^r$ is the piecewise-continuous r-dimensional control vector, assumed to be confined to some parametric family of compact sets $\mathcal{M}[t]$, and $F\colon \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ is a sufficiently well-behaved scalar-valued kernel or "rate-of-cost" function, subject to: the first-order evolution conditions:

$$\tfrac{d}{dt}\boldsymbol{x}(t) = \boldsymbol{q}\big(\boldsymbol{x}(t), \boldsymbol{m}\big); \tag{2.47}$$

for some autonomous flow field $\boldsymbol{q}\colon \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^n$; the equality constraints

$$\int_{t_0}^{t_1} dt\, \boldsymbol{g}\left(\boldsymbol{x}(t), \boldsymbol{m}(t)\right) = \boldsymbol{K} \tag{2.48}$$

for some vector function $\boldsymbol{g}\colon \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^r$ and vector of constants $\boldsymbol{K} \in \mathbb{R}^r$; and two-point boundary conditions

$$\boldsymbol{S}\big(\boldsymbol{x}(t_0), t_0; \boldsymbol{x}(t_1), t_1\big) = \boldsymbol{0} \tag{2.49}$$

for some function $\boldsymbol{S}\colon \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \to \mathbb{R}^{n_S}$ for which the evolution remains well-posed.

If we then define an effective Hamiltonian

$$H\left(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{m}\right) = -F\left(\boldsymbol{x}, \boldsymbol{m}\right) + \boldsymbol{\delta} \cdot \boldsymbol{g}\left(\boldsymbol{x}, \boldsymbol{m}\right) + \boldsymbol{\lambda} \cdot \boldsymbol{q}\left(\boldsymbol{x}, \boldsymbol{m}\right), \tag{2.50}$$

where $\boldsymbol{\lambda}(t) \in \mathcal{R}^n$ is the so-called co-state vector, and $\boldsymbol{\delta} \in \mathcal{R}^p$ is a vector of constant Lagrange multipliers, then necessary conditions for a constrained optimum $\tilde{\boldsymbol{m}}(t)$ are:

$$\tfrac{d}{dt}\boldsymbol{x} = \quad \tfrac{\partial}{\partial \boldsymbol{\lambda}} H, \tag{2.51a}$$

$$\tfrac{d}{dt}\boldsymbol{\lambda} = -\tfrac{\partial}{\partial \boldsymbol{x}} H, \tag{2.51b}$$

$$\tfrac{d}{dt} H = \quad \tfrac{\partial}{\partial t} H = 0, \tag{2.51c}$$

$$\tfrac{d}{dt}\boldsymbol{\delta} = \quad \boldsymbol{0}, \tag{2.51d}$$

for almost all $t \in [t_0\,,\, t_1]$, together with

$$\boldsymbol{S} = \boldsymbol{0}, \tag{2.52a}$$

$$\int_{t_0}^{t_1} dt\, \boldsymbol{g} = \boldsymbol{K}, \tag{2.52b}$$

$$\left[-H\delta t + \boldsymbol{\lambda} \cdot \delta \boldsymbol{x}\right]\Big|_{t_0}^{t_1} = 0, \tag{2.52c}$$

and finally

$$H\left(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \tilde{\boldsymbol{m}}\right) \geq H\left(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{m}\right) \tag{2.53}$$

for all $m(t) \in \mathcal{M}[t]$ and almost all $t \in [t_0\,,\, t_1]$.

Note that the vector field $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{m})$ can be conservative and generated by some Hamiltonian, or might be dissipative or otherwise non-Hamiltonian. A non-autonomous flow can be made formally autonomous by the usual device of augmenting the state variables. If the flow is symplectic, the associated dynamical Hamiltonian $\mathcal{H}$ need bear no obvious relationship to the effective Hamiltonian H which emerges in the optimization procedure.

We shall see that many of our optimizations involving quasi-static LWFA dynamics with a prescribed laser field can be formulated precisely in this form, with t replaced by ξ, the control field $\boldsymbol{m}(t)$ taken as the normalized laser intensity $a(\xi)^2$ (not the laser field $a(\xi)$ itself), and the state vector trajectory $\boldsymbol{x}(t)$ taken to be the pair $(\phi(\xi)\,,\, \phi'(\xi))$, where the vector field $\boldsymbol{q}$ is equivalent to the QSA wake equations for the normalized potential

and field, the initial conditions $\boldsymbol{S}$ ensure quiescence prior to the arrival of the laser field, $\boldsymbol{g}$ represents the kernel of any integrated constraints imposed on the laser field, and F is proportional to the kernel of the integrated figure-of-merit to be optimized.

Note that typical LWFA optimizations with prescribed laser field do not exhaust the full generality of the Pontryagin formalism: although nonlinear in the potential, the flow field $\boldsymbol{q}(\phi, \phi', a^2)$ is actually linear in the control field $a(\xi)^2$, the constraints embodied in $\boldsymbol{g}$ are usually independent of the wake variables (ϕ, ϕ') and depend only on the normalized laser intensity $a^2(\xi)$, and the often (but not always) the cost kernel F is independent of $a(\xi)^2$, depending only on the wake potential $\phi(\xi)$ and/or field $\phi'(\xi)$, or at most depending on the laser field only through $\Theta\left(a(\xi)^2\right)$. It is essential that we neglect any τ-dependence in the laser envelope due to dispersion, depletion, distortion, etc., so that the ponderomotive "control field" remains a function of a single independent variable. What is not immediately obvious, but what is for us one of the more useful aspects of the Pontryagin formalism, is that because of these simplifications, in many cases of interest certain features of the optimal drive envelope $a(\xi)^2$ can be deduced without explicitly solving for the exact trajectory $\phi(\xi)$ of the corresponding wake potential.

2.6 What to Optimize, What to Vary, and What to Constrain?

In asserting or assessing claims of "the best" pulse-shape, we should of course keep in mind that the answer will depend, sometimes quite sensitively, on precisely what figure-of merit is optimized, on the range or set of possibilities (parameterized or not) over which the laser profile is allowed to vary, and what other constraints are imposed or how they imposed, or what remaining quantities and parameters are bounded or held fixed. Some of the debate on the topic of LWFA pulse-shaping in the literature and at conferences has arisen because investigators were really attempting to answer different questions, but not explicitly specifying enough of their assumptions as to make this evident.

The ultimate aim of laser-plasma-based acceleration would seem to be the reproducible production of accelerated electron bunches, of sufficiently high energy and with acceptably low energy spread and transverse emittance, at an acceptably high repetition rate, by using an experimentally accessible laser source of achievable intensity, energy, and longitudinal

and transverse profile, bandwidth, within plasma of achievable and reproducible density, temperature, physical extent, and degree of homogeneity.

Along the way, direct or indirect goals include high energy gain over short distances, enhancement of other desirable post-acceleration beam parameters (high current, low emittance, low energy spread, etc.) ease of structuring, injecting, or extracting accelerated beam, and overall efficiency of energy transfer between laser driver and plasma and between plasma and accelerated beam.

2.6.1 Possible Figures-of-Merit

In attempting to quantify these goals, possible figures-of merit include:

- "gain-like" quantities: energy gain per bunch or per particle, peak longitudinal electric field of wake, RMS field amplitude, wakefield energy, either measured on-axis or transversely integrated or averaged and possibly weighted by some assumed transverse shape of accelerated beam;
- "efficiency-like" quantities: efficiency (proportion) of energy transfer between driver and wakefield, efficiency of energy transfer between driver and accelerated beam, driving-beam loading, various transformer ratios;
- "driver back-reaction" measures: uniformity of driver particle deceleration, resistance to drive beam distortion or laser instabilities that could impede further coupling to the wake;
- other "beam quality" measures: peak or RMS beam current, energy spread in accelerated beam, transverse emittance of accelerated beam;
- "wake quality" measures: width, number, shape, or suitability of wakefield buckets, uniformity of accelerating fields, ease of or tolerances on injection;
- "robustness" measures: resistance of desirable wakefield or beam properties to degradation in the face of jitter, fluctuations, or shot-to-variations.

In the literature, figures-of-merit are also referred to variously as performance measures or metrics, objective functions, utility functions, or (negative) cost functions. Of course, to ensure a well-posed optimization problem, any of these choices would have to be defined more precisely in terms of the physical fields or particle distribution functions.

Various closely-related possibilities may nevertheless turn out to be *inequivalent* with respect to the corresponding optimal driver-shapes, depending for example, on whether one considers the amount of acceleration or energy transfer achievable in principle or actually achieved, on whether the energies delivered to the accelerated beam or just to the plasma are considered, or whether the energies delivered by the driving beam or acquired by the accelerating beam are measured per unit length or over the whole accelerating structure, or measured per particle or for the beams as a whole.

2.6.2 Possible Constraints

Natural constraints on the driving laser pulse include prescribed values (equality constraints) of, or limits (inequality constraints) on, such quantities or parameters as:

- total laser energy, envelope energy, or action, etc.;
- total energy or kinetic energy per-particle;
- peak laser intensity, normalized intensity, or action density, etc.;
- "gross" (first-on to last off) laser pulse duration;
- "net" pulse duration (time over which the intensity is nonzero, or more generally exceeds some small threshold);
- RMS duration, weighted by energy, action, or related measure;
- average intensity or action (of either a gross or net variety);
- number of sub-pulses allowed in a pulse train;
- limitations on the bandwidth or on the spectral content or temporal profile more broadly;
- initial accelerated bunch properties (energy, current, emittance, current, etc.),

any of which must also be quantified precisely before they can be imposed, and how they are quantified can matter greatly.

For example, in general very different optima can be obtained if we limit what we have called gross or net pulse duration, or if we constrain the pulse duration in terms of absolute bounds or, say, action-weighted or intensity-weighted RMS values, or through more general

conditions on the laser bandwidth. When needed, we will mostly use strict bounds for mathematical convenience, but also briefly work directly with the spectral transfer functions in the linear regime.

Attempting to minimize or constrain absolute versus relative energy spread, or normalized versus unnormalized emittance, might be expected to lead to very different solutions. The latter quantities are usually but not always of more interest in beam physics applications.

In order to determine a meaningful optimal pulse-shape, one must first choose a primary *scalar* figure-of-merit, such as one of those above, or perhaps some compound measure consisting of some weighted combination of them, and impose certain equality or inequality constraints on one or more of the drive beam parameters above, as well as perhaps prescribe minimally-acceptable values for certain other secondary figures-of-merit. Different combinations will lead to different optimal pulse shapes, and of course excessive or inconsistent constraints can lead to no solution at all.

One essential but implicit mathematical constraint on the drive pulse in the case of the LWFA is nonnegativity: obviously $a(\xi)^2 \geq 0$. This might sound too trivial to warrant mentioning, but it actually imposes rather severe restrictions, and has important consequences for optimal shaping.

Energy, Envelope Energy, Action

A comment is in order on our use of energetic or related constraints on the driving laser source. Usually we will constrain the integral

$$U_a = \tfrac{1}{2}\int d\xi\, a(\xi)^2 \tag{2.54}$$

which is variously referred to as the scaled envelope energy or the scaled wave action. In the case of eikonal fields and lowest-order (i.e., purely transverse) paraxial polarization, U_a is approximately proportional to both the EM energy and the action, but is proportional to neither if the laser envelope varies insufficiently slowly either longitudinally or transversely and the paraxial or envelope approximations break down. In standard units and lab-frame coordinates, this is just proportional to

$$U_a \propto \int_{-\infty}^{\infty} dz\ |A(z,t)|^2 \tag{2.55}$$

for any time t, since we are assuming a prescribed driver traveling at constant group velocity v_g. Assuming a constant plasma density, and ignoring any nonlinear dielectric response[7] the transverse EM energy associated with the laser is instead proportional to

$$\begin{aligned}\mathcal{H}_{\rm EM} &\propto \int_{-\infty}^{\infty} dz\ |\boldsymbol{E}_\perp(z,t)|^2 + |\boldsymbol{B}(z,t)|^2 \\ &= \int_{-\infty}^{\infty} dz\ \left|\tfrac{1}{c}\tfrac{\partial}{\partial t}\boldsymbol{A}(z,t)\right|^2 + |\boldsymbol{\nabla}\times\boldsymbol{A}(z,t)|^2\,.\end{aligned} \tag{2.56}$$

For simplicity, let us assume a fixed circular polarization. (Other cases are similar, but with additional terms which complicate the expressions without changing our essential argument). In Fourier-space (conjugate to z), we define $a^+(k;t)$ as the analytic signal (positive-frequency part) of $a(z,t)$, which can be determined in principle by taking Hilbert transforms of $\boldsymbol{A}(z,t)$ or Fourier transforms of $\boldsymbol{E}(z,t) \pm i\boldsymbol{B}(z,t)$. Up to an overall constant of proportionality, $a^+(k;t)$ is the classical analog of what would become after canonical quantization the annihilation operator associated with photons of the chosen helicity and momentum $\hbar k$.

What we have called the scaled envelope energy is proportional to

$$U_a(t) \propto \int dk\ \tfrac{1}{\omega(k)}\left|a^+(k;t)\right|^2. \tag{2.57}$$

where $\omega(k)$ is the EM dispersion relation, while the full EM energy is proportional to

$$\mathcal{H}_{\rm EM}(t) \propto \int dk\,\omega(k)\left|a^+(k;t)\right|^2, \tag{2.58}$$

and if decide instead that the wave action (i.e., number of photons) rather than energy is a more natural physical currency, we find

$$\mathcal{J}_{\rm EM}(t) \propto \int dk\ \left|a^+(k;t)\right|^2, \tag{2.59}$$

so each of these differs from the others by a power of the optical frequency $\omega(k)$ inside the integral. If we can invoke a slowly-varying envelope approximation and set $\omega(k) \approx \omega_0$ everywhere $|a^+(k;t)|^2$ is appreciable, as will typically be the case in realistic pulses, then the three different quantities are all approximately proportional.

However, even when we consider impulses, or other laser envelopes with very short or sharp features which may invalidate the eikonal approximation, we will for the most part

[7] Alternatively, once can refer the energy to some early time where the pulse is assumed to be launched in vacuum, and assume that reflection losses are made negligible by sufficiently gradual tapering of the density at the boundaries. Then the dielectric (linear or nonlinear) inside the bulk of the plasma is immaterial.

continue to constrain U_a rather than $\mathcal{H}_{\text{EM}}$ and continue to refer to it as a scaled envelope energy. Likewise, we will refer to $a(\xi)^2$ itself as the normalized laser intensity or laser strength, even though the actual intensity is instead proportional to $\left|\frac{\partial}{\partial\xi}\boldsymbol{a}(\xi,\tau)\right|^2$.

This is in part a desire to follow previous conventions in the literature, but mostly out of sheer convenience. It is far simpler to treat $s(\xi) = \frac{1}{2}a(\xi)^2$ as the exogenous control field over which to optimize, rather than the $\boldsymbol{a}(\xi,\tau)$. Then constraints on U_a involve no additional derivatives, and in fact are then linear in the control field, which greatly simplifies the mathematics. For pulses of duration significantly less than the plasma period, i.e., $\Delta\xi_d \ll 1$, the actual energy may differ significantly from the so-called envelope energy, and the effects of this discrepancy for short pulses on our various optimizations has begun to be investigated numerically.

Pulse Durations

Pulse durations can also be specified in various ways, sometimes with profound effects on the optimized pulses.

The "gross" pulse duration (in scaled units) can be defined as

$$\Delta\xi_g = \sup\left\{\xi\colon a(\xi)^2 > 0\right\} - \inf\left\{\xi\colon a(\xi)^2 > 0\right\}, \tag{2.60}$$

although if we seek to impose an upper bound on the duration, it is easier to specify

$$\left\{\xi\colon a(\xi)^2 > 0\right\} \subseteq [\xi_0\,,\,\xi_0 + T] \tag{2.61}$$

for some $\xi_0 \in \mathbb{R}$ and $T \geq 0$. In order to constrain the "net" pulse duration, we can use

$$\Delta\xi_n = \int d\xi\,\Theta_0\left(a(\xi)^2 - \epsilon\right), \tag{2.62}$$

where $\epsilon \geq 0$ is some optional threshold value, although we will take $\epsilon = 0$. The RMS duration may be defined as

$$\Delta\xi_{\text{RMS}} = \left[\left\langle\xi^2\right\rangle - \left\langle\xi\right\rangle^2\right]^{1/2} \tag{2.63}$$

where in this context we define

$$\left\langle\xi^n\right\rangle = \frac{\int d\xi\,\xi^n\,a(\xi)^2}{\int d\xi\,a(\xi)^2} \tag{2.64}$$

Instead, spectral rather than temporal constraints might be invoked, related by Fourier conjugacy.

2.6.3 Additional Parameters and Optimizations

Here we investigate the choice of optimal *longitudinal* laser envelopes, under various constraints of laser energy, intensity, and/or pulse duration, while fixing other laser parameters, including the central laser carrier frequency ω_0, and essentially all of the relevant plasma parameters, such as the plasma frequency ω_p and overall length L_0.

More practically, one might also consider *simultaneously* determining, along with the longitudinal laser profile, the optimal choices for the laser frequency ω_0 and laser spot-size $\sigma_\perp$, as well as the plasma density profile n_0, including the overall average density and any possible channel density gradient.

Such considerations largely lie outside the scope of our investigations here. We will just suggest by example that the issues are not entirely trivial. For example, for a given set of laser parameters (frequency ω_0, normalized peak intensity a_0^2, bandwidth $\Delta\omega$, etc.), in a homogeneous plasma, one might naturally examine the dependence on the plasma density n_0, or equivalently, plasma frequency ω_p. The physical electric field is related to the scaled wakefield by $E_z(z,t) = \frac{E_0}{\beta_g}\phi'(\xi)$, where E_0 increases linearly with ω_p and β_g decreases weakly with ω_p in the underdense regime, suggesting that larger gradients can be achieved in denser plasmas at fixed laser strength. But increasing the density tends to decrease the characteristic diffraction, dispersion, depletion, de-phasing, and instability length-scales, and of course increases the collision cross sections and eventually invalidates the assumptions of an underdense plasma. Also, according to the simple scaling predictions, the product $\max\big[\mathcal{E}_z(\xi)\big]L_{\rm det}$ should instead actually decrease with increasing density throughout most of the underdense regime. If, rather than considering fixed laser energy, we operate in the nonlinear regime but somehow fix the ratio $\frac{E_+}{E_{\rm WB}}$, then energy gain in a single-stage limited by detuning can also be enhanced by moving to lower densities. However, as ω_p decreases, the collisionless skin depth and nonlinear plasma wavelength increases, as does the required interaction length within the homogeneous plasma. In all these various considerations, the optimum may depend on factors such as whether the energy required to ionize the plasma is included in the budget.

Increasing the density may increase the peak field, but by decreasing the plasma wavelength, will also make phase-matching requirements in the injection process more severe, and otherwise deteriorate the energy spread for fixed bunch size or reduce the bunch-length that can experience something close to the maximum gradient and be accelerated with an acceptable degree of uniformity. In general one might expect to achieve optimal particle ac-

celeration consistent with the various technological constraints at some intermediate plasma density, but the specific scalings must be examined carefully in individual cases.

2.6.4 Transformer Ratios and their Interpretations

Despite frequently-encountered references to "the" transformer ratio in the beam physics literature, a number of different transformer ratios can be defined, depending on whether total relativistic energy, kinetic energy, change in kinetic energy, or rates of change (gradients) of kinetic energy of the drive and accelerated particles are involved, and whether maxima or averages are to be considered, and whether unscaled or scaled energies (by the rest mass or effective rest mass) are used, and perhaps other variations. These options leave at least 256 permutations for ratios, not all of which are truly distinct in their meaning or particularly transparent in their interpretation or even useful in practice, but by our count, at least three distinct definitions for transformer ratios have been proposed in the beam physics literature, and at least two more might be of interest.

First is the ratio of the maximum kinetic energy gain per accelerated particle to the initial average total relativistic energy per drive particle:

$$R_1 = \frac{m_a c^2 \max[\Delta\gamma_a]}{m_d c^2 \langle\gamma_d\rangle_{t=t_0}}, \tag{2.65}$$

where here averages are taken over all particles in the relevant beam, and t_0 is some time before the laser deposits energy into the plasma. Assuming ideal injection of the accelerated electrons, this simplifies in the case of the LWFA to

$$R_1 = 2\pi \frac{\omega_p}{\omega_0} \frac{\max[|E_+(\xi)|\, L_a]}{\lambda_c E_0}, \tag{2.66}$$

where $\lambda_c = \frac{h}{m_e c}$ is the electron Compton wavelength, and L_a is here regarded as the effective acceleration length as determined self-consistently from the energy gain (not necessarily exactly following any of the particular scaling laws mentioned above). For a fixed value of the average driving particle energy (just proportional to the carrier frequency ω_0 in the case of the LWFA), maximizing R_1 is therefore just equivalent to maximizing the energy gain itself, or in practice the *best-case* gain if actual trajectories are not followed but rather ideal injection and acceleration is assumed. In situations the acceleration length L_a is limited either to L_{diff} or L_{det}, the ratio R_1 will just be a monotonically increasing function of the (scaled) peak accelerating wakefield amplitude $\max\big[|\mathcal{E}_+(\xi)|\big]$. Rather than introducing the transformer ratio R_1, it is probably simpler to speak directly of the wake amplitude or energy gain.

A more commonly-encountered definition is the ratio of maximal kinetic energy gain per accelerated particle to the average kinetic energy lost per drive particle:

$$R_2 = \frac{m_a c^2 \max[\Delta\gamma_a]}{m_d c^2 \left|\langle \Delta\gamma_d \rangle\right|}. \tag{2.67}$$

For the LWFA, again assuming ideal electron injection, this can be written as

$$R_2 = 2\pi \tfrac{\omega_p}{\omega_0} \frac{\max\left[|E_+|\, L_a\right] \bar{L}_d}{E_0 \lambda_e \min\left[\bar{L}_d, L_0\right]}, \tag{2.68}$$

where is an average depletion length averaged over all photons in the drive pulse. This ratio coincides with R_1, and is proportional to the best-case (with optimal injection) energy gain, if $\bar{L}_d < L_0$, but otherwise is difficult to interpret.

Instead, in most treatments of pulse-shaping in either the PWFA or the LWFA, still another definition of transformer ratio has been used, namely the ratio of the maximal rate (per unit distance) of normalized kinetic energy gain per accelerated particle to the maximal rate of normalized kinetic energy lost per drive particle:

$$R_3 = \frac{\max[\frac{\partial}{\partial z}\gamma_a]}{\max\left[\left|\frac{\partial}{\partial z}\gamma_d\right|\right]}. \tag{2.69}$$

In the LWFA, this becomes

$$R_3 = \frac{k_p \max\left[|\mathcal{E}_+(\xi)|\right]}{\frac{1}{2} k_p \frac{\omega_p}{\sqrt{\tilde{\gamma}_\perp}\omega_0} \max\limits_{a(\xi)\neq 0}[\kappa(\xi)]} = \frac{2\sqrt{\tilde{\gamma}_\perp}\omega_0}{\omega_p} \frac{\max\left[|\mathcal{E}_+(\xi)|\right]}{\max\limits_{a(\xi)\neq 0}[\kappa(\xi)]} \propto \sqrt{\tilde{\gamma}_\perp} \max[|\mathcal{E}_+(\xi)|] k_p L_{\rm pd}, \tag{2.70}$$

which in the linear limit ($a^2 \ll 1$) simplifies further to

$$R_3 \propto \frac{\max\left[|\mathcal{E}_+(\xi)|\right]}{\max\limits_{a\neq 0}\left[|\mathcal{E}_-(\xi)|\right]}, \tag{2.71}$$

where $\mathcal{E}_-(\xi) = \Theta_0\left(-q_a \mathcal{E}(\xi)\right) |\mathcal{E}(\xi)|$ is the (scaled) magnitude of the decelerating field.

In either the linear or nonlinear regime, maximizing R_3 is therefore equivalent to maximizing the best-case energy gain *if and only if* $L_a \propto L_{pd}$ for essentially all particles in the accelerated bunch. Again, however, in essentially all achievable regimes of interest for the LWFA, either with or without a channel, recall that $L_{\rm det} < L_{\rm pd}$, so $L_a < L_{\rm pd}$ strictly, and therefore acceleration over the full depletion length tends not to occur, and in realistic cases there is no reason to expect any direct connection between a high transformer ratio R_3 and a large energy gain for the target particles (either in the best-case or on average). This will be discussed further below when we analyze the pulse shapes purporting to maximize R_3.

Without further assumptions that are debatable or even doubtful, a large value of the transformer ratio R_3 does not seem directly related to a large wakefield amplitude or large energy transfer from pulse to wake or from wake to accelerated bunch.

Due to bunch-length limitation, or limited tolerances over initial energy spread or overall phase control, or possibly subsequent phase-slippage or beam-loading in the wake, particles in the accelerated bunch may not experience something close to the maximum possible field over the entire acceleration length. In this even, it might be better to use an average rather than maximal energy gain in the numerator. While more defensible physically, this tends to be much more complicated mathematically, and is usually avoided.

Although it does not seem to be defined as such, we can also introduce the ratio of average energy gained per accelerated particle to the average energy lost per drive particle:

$$R_4 = \frac{\langle \Delta\gamma_a \rangle}{\langle \Delta\gamma_d \rangle}. \tag{2.72}$$

But if N_a is the number of particles in the accelerated bunch and N_d is the number of particles in the drive beam, then by definition this is just

$$R_4 = \tfrac{N_d}{N_a}\eta', \tag{2.73}$$

where η' measures the fraction of energy lost by the drive beam that actually goes into particle acceleration. Similarly we can define introduce the ratio of average energy gained per accelerated particle to the average total energy per drive particle:

$$R_5 = \frac{\langle \Delta\gamma_a \rangle}{\langle \gamma_d \rangle} = \tfrac{N_d}{N_a}\eta, \tag{2.74}$$

where η is the overall efficiency, meaning the actual fraction of the total energy in the drive that is successfully transferred to the accelerated bunch.

One might also envision definitions involving averages of ratios, rather than ratios of averages, but we will not pursue such ideas or various other possibilities further.

2.6.5 More on Competing Figures-of-Merit: Critiques and Apologias

Those not involved or interested in the long-running debate concerning what to optimize in the LWFA can certainly skim or skip this sub-section.

In considerations of pulse-shaping in LWFA (or PWFA) schemes, one must first decide what attribute or weighted combination of attributes of the wake generation and particle

acceleration processes one most desires optimized, and what further constraints limit the set of possible solutions. Any number of attributes might be of interest, but are not always or usually compatible goals in an optimization framework. The over-arching goal is of course to accelerate a beam of sufficiently collimated and quasi-mono-energetic particles to relativistic energies over a relatively short distance, without introducing excessive degradation of other beam properties. Thus it seems natural to us to adopt as a *primary* figure of merit some measure of the energy transfer from driving beam to the particles in the accelerated beam. To simplify the problem and defer issues regarding the beam injection, wake and beam quality, and beam-loading, will consider some measure of the energy transfer from driving beam to the plasma, or of the size of the resulting wakefield, a somewhat more indirect but far more convenient metric.

Many other quantities might also be of interest, which under suitable constraints might be independent of, complementary to, or in conflict with the primary goal of acceleration to various degrees. For example, despite previous claims in the literature that a single pulse shape can yield the maximum possible efficiency of laser-plasma coupling, and the maximum possible wakefield for a given total laser power, the peak magnitude of the wake and the efficiency of energy transfer to the wakefield tend to be conflicting goals. Under most natural constraints, optimization of the transformer ratio R_3 or the peak wakefield amplitude do not lead to the same pulse-shapes.

High energy gain and low energy spread had been suspected to be rather incompatible in practice, but more recently, the theoretical prediction and apparently the experimental verification of a highly-nonlinear "bubble" regime have demonstrated that reasonably collimated and mono-energetic bunches can be self-injected (i.e., trapped into the Langmuir wave directly out of the background plasma) and accelerated to O(Gev) energies. Whether this can be extended to the case of higher-currents through a controlled injection process is a subject of current study.

Although issues of driver beam loading and overall efficiency will ultimately be of some concern in any practical multi-stage plasma accelerator design, interest in and comparative advantages of the LWFA primarily lie in the possibility of large electric fields and large energy gain over short distances, compared to conventional RF-cavity accelerators, so it would see natural that this should be reflected in the *primary* figure-of-merit used to choose laser pulse shapes., especially in proof-of-principle experiments and research-and-development-oriented stages of first or second-generation laboratory laser-plasma systems.

The goals of acceleration and high accelerating gradients simply seem more fundamen-

tal, logically prior to constraints regarding overall or per-particle efficiency in acceleration. Unless sufficiently high energies and acceptable beam quality can be produced over relatively short distances, the question of efficiency will largely be moot, because the plasma accelerator will be of little practical interest in comparison to standard RF structures. Only after the possibility and range of accelerating gradients are established do considerations of efficiency really enter, informing the physical and economic analysis of how much one must be willing to pay (energetically, and ultimately financially) for a given final beam energy.

While any practical accelerator design should not needlessly waste an appreciable fraction of the driving beam energy, if performance and reproducibility continue to improve to the point where imagining "production models" is possible, construction and secondary operating and maintenance costs of compact (if never quite "table-top") LWFA systems might be so much less than traditional synchrotron designs that a lower wall-power efficiency might be happily tolerated.

As achieved energies continue to make impressive experimental gains, we do not dismiss or deny the increasing importance of accelerated beam injection, beam loading, and beam quality issues, but only suggest that they are mostly deferred until a better understanding of the competing issue in the wakefield generation itself are better understood. Besides, issues of injection will likely depend sensitively on the specific schemes designs, which constant a active and rapidly-evolving area of experimental and theoretical research.

We contend that wakefield strength or energy gain will likely remain the primary figure-of-merit of interest, beam quality issues (current, energy spread, and transverse emittance) will increasingly emerge as important secondary goals, followed by ease or controllability of injection as tertiary goals, and only then will overall efficiency emerge as a quaternary goal. Per-particle efficiency or quantities like the R_3 transformer ratio will become relevant to the LWFA only to the extent that acceleration can be sustained over distances approaching the full depletion length of the drive beam. Eventually, in any specific situation one might try to optimize some combined or compound performance measure balancing the competing demands of the device, including large acceleration in a compact space, high efficiency in the acceleration, ease of structuring, injecting, and extracting the particle beam, etc., or explore how changing bounds on certain features or parameters will alter the bounds on others.

An obvious requirement of any figure-of-merit (or constraint) is that the physics to which it alludes should actually be included in the dynamical model used. Otherwise some indirect proxy can be employed, with some degree of caution depending on the extent to which we

might expect it to be correlated with the original quantity of interest, or else some other kind of measure altogether must instead be chosen. For example, we may be most interested in the peak or average energy gain of the (externally or self) injected particles, but unless we actually track the beam distribution function or at least some representative sample of particle trajectories, we must instead rely on the wakefield amplitude, or a predicted gain estimated as a product of the field strength and some estimated acceleration length, either of which which ignores specifics of the injection and subtleties of slippage, as well as beam-loading issues, where the accelerating particles absorb energy from the wake and locally decrease its amplitude.

When speaking of the energy transferred from the laser pulse to the wake or from the wake to the injected particles, without accounting for the back-action on the laser or the wake, respectively, physically and logically we slip into something of a state of sin, at least outside of the weakly perturbative regime (the transcendence of which is often the ultimate aim of pulse-shaping in the first place). The transformer ratio is necessarily connected to notions of loading and depletion of the driver beam, whereas in the analytic work, essentially no back-reaction on the driver is even considered, and the photon deceleration is estimated non-self-consistently from the wake response via inferred energy conservation.

Actually, with the transformer ratio R_3, our sin may actually be cardinal rather than venial. In analogy with the PWFA, it has been suggested that true merit of the pulse shape maximizing R_3 may not be in the large transformer ratio *per se*, but in the uniform photon deceleration over the bulk of the pulse, which tends to minimize additional dispersive effects as well as Raman self-modulation or other instabilities which might otherwise distort the pulse shape and inhibit further coupling to the wake as the pulse continues to propagate.

We can only be suspicious of appealing to so much physics not incorporated into the dynamical model for justification of the optimization framework adopted or the optimal solutions found. While maximizing the transformer ratio for the LWFA may tend to extend the distance or duration over which energy can be effectively extracted from the driver before excessive beam degradation occurs, this question should be studied within a dynamical framework which actually allows for nontrivial evolution and back-reaction on the laser driver, including the parametric laser-plasma instabilities and any significant diffraction, dispersion, and depletion effects. Compared to the case of the electron-driven PWFA, in the typical LWFA regime we are less able to justify any assumption that the laser driver can lose a significant portion of its energy without greatly affecting its relative shape or its group velocity.

Besides, this *ex post facto* justification really amounts to an argument for figure-of-merit different from a maximal transformer ratio, namely maximizing the uniformity (or minimizing the variability) of drive particle deceleration, or minimizing the rate or amount of drive beam distortion quantified in some manner, or perhaps just maximization of the energy transfer from laser to wake. If any of these really is the desired figure-of-merit, then it should be invoked directly, but in a model rich enough to describe it. Even within the non-self-consistent model, we will see that the proposed transformer ratio solution leads to a prediction of uniform deceleration throughout most but not all of the driver beam. The impulsive precursor does not experience the same photon deceleration as the subsequent nonlinear ramp, and in fact the deceleration factor $\kappa(\xi)$ changes extremely rapidly at the head of the pulse, leaving open the possibility of other shapes which lead to still less overall differential deceleration or distortion when averaged over all photons in the pulse. One can imagine that if, due to short or sharp features, the driving pulse experiences significant group-velocity dispersion, a pulse shape intentionally leading to some differential photon deceleration, resulting in a sort of longitudinal focusing tending to counteract the dispersion, would experience less overall distortion than one exhibiting uniform deceleration. It is an interesting but as yet unanswered questions as to whether there is something resembling a "self-similar" pulse that retains its relative shape and/or efficacy as it uniformly slows while exciting a wake.

We care about a large value of the ratio R_3 insofar as we are interested in a large value of its numerator, and possibly interested in small values of its denominator, if that really does turn out to minimize pulse distortion. But the ratio in the absence of a large wake can also be made large by making the numerator small but the denominator relatively smaller. Mathematically speaking, maximization of a ratio is equivalent to the constrained maximization of its numerator subject to a suitable constraint on its denominator, namely fixing it at the value that it would have assume in the unconstrained maximization of the ratio. In the case of the R_3 transformer ratio, unless acceleration over the full depletion length can somehow be achieved, it remains unclear to us why we should be particularly concerned about these quantities in this particular combination or why we should think to maximize the wakefield at a given value of the depletion length rather than some other length-scale of greater relevance to the actual LWFA operation or that that can be more naturally fixed or measured experimentally. In any case, rather than insisting on the transformer ration, we can instead just include both the numerator (proportional to maximum wakefield) and

denominator (proportional to predicted photon deceleration, or inverse depletion length), separately on the menu of possible figures-of-merit and constraints.

If we do seek to optimize overall efficiency of energy transfer in a wakefield excitation scheme (either laser-driven or particle-driven), maximizing efficiency without allowance for the absolute magnitude of the energy transfer is problematic. It is obvious that in the absence of any other constraints the problem of maximizing efficiency is rather ill-defined, for it is a trivial matter to minimize any wasted energy by supplying no energy in the source and delivering no energy to the wake. At minimum then, a lower bound on the total driver beam energy must be specified, as one would of course expect. But even this is not sufficient, since it is seems possible to adiabatically transfer energy to the wake with essentially an arbitrarily high efficiency, by using an arbitrarily long driving source interacting with the plasma over a sufficiently long distance. If the total energy in the driving beam is bounded, then this implies that the resulting wake will either possess low peak gradient, low phase velocity, or both, making it unsuitable for accelerator applications. In addition, other constraints must be imposed, such as the maximum duration of the drive beam or the minimum amplitude of the wake, to ensure that one avoids converging to an "optimal" solution which transfers little useful energy, but with high efficiency. It might be argued that such degenerate or spurious cases are obvious and easy to avoid with suitable constraints, but we will see that the transformer ratio itself also has something of this character.

Why the Transformer Ratio?

What then is the appeal of the transformer ratio, or what was the reasoning leading to it? We suspect that the motivation for maximizing the transformer ratio R_3 in the LWFA arose out of taking somewhat too seriously the close but not perfect analogy to the PWFA – being too easily seduced by the similarities without sufficient mindfulness of the important remaining differences.

Although often described in accelerator applications as a measure of "efficiency of coupling" between energy source and accelerating structure (in our case, between the drive pulse and the plasma) the transformer ratio is *not* really a pure measure of efficiency in the strict sense used above, so perhaps here we should refer to it as figure-of-merit reflecting the "efficacy" of laser-plasma coupling: it is meant to offer a *relative* measure of the *per-particle*

rate of energy transfer between the driving source and the accelerated beam under ideal conditions.

Prior to its use with the LWFA, the transformer ratio was introduced in the context of the PWFA, where typically the drive beam and the accelerated beam consist of the same species of charged particles (usually electrons), and the intended meaning of the transformer ratio is more apparent, as is its analogy to the conventional step-up ratio of transformers in electrical circuit theory (except that we are primarily concerned with the kinetic rather than potential energy of the particles). If acceleration at nearly the maximum rate can be sustained over the depletion length, a high transformer ratio ($R_3 >> 1$) indicates that a comparatively large number of *relatively* low-energy (but still relativistic) particles in the driving beam can accelerate a smaller number of (already relativistic) particles in the driven beam to higher energies. Conversely, a low value ($R_3 \ll 1$) suggests that we may not gain much at all from attempting to transfer energy form the drive to accelerated beams.

For typical parameter regimes, because the unit energy cost of producing a drive photon at the energy corresponding to the carrier frequency is so much smaller than the accelerating an electron in the drive beam to some at least moderately relativistic energy, in any LWFA scheme whatsoever which produces appreciable energy gain (as compared to the electron rest mass), the number of drive photons will by necessity exceed by an enormous factor the number of accelerated electrons, without any particular effort to deliberately increase the transformer ratio. And In contrast with the LWFA, in the PBWA with sufficiently high drive-beam energy, the depletion length $L_{\rm pd}$ and de-tuning length $L_{\rm det}$ tend to be closer in magnitude, and smaller than the either the dispersion length $L_{\rm disp}$ assuming sufficiently low longitudinal emittance, or the betatron wavelength $\frac{2\pi}{\kappa_\beta}$ (analogous to the optical diffraction length) for suitable beam optics and reasonable transverse emittance.

So in comparison, the transformer ratio (R_3) will be less relevant in the case of the LWFA because the driving and accelerated particles consist of very different species of very different energies in the LWFA, vastly different technologies are needed for creating the driving beams in the two cases, and the characteristic length and time-scales limiting the energy transfer also tend to be very different.

In the PWFA as well as the LWFA, it has been argued that the true appeal of shapes purporting to maximize R_3 is the side-benefit that the drive particles in a pulse so optimized tend to all slow down at the same rate. If the driving beam is non-dispersive initially, then such uniform deceleration will maintain the small velocity spread and hence tend to preserve the relative profile of the driving source, i.e, the driving beam tends to slow down

without changing shape, so that the beam can propagate further through the plasma and deliver more energy before it degrades too much to be effective for excitation. Uniform deceleration also suppresses instabilities (two-stream in the case of particle beam, Raman Self-Modulation in the case of the laser driver) which rapidly distort the beam shape and tend to re-arrange energy within the driving beam or excite unwanted modes rather than transfer energy to the plasma wave. It might be desirable to limit any distortion of the driver which would otherwise impede coupling, but such feedback has not been included even in analyses of the PWFA, and cannot easily be included in analytic treatments, so this claim is largely unjustified, and such back-reaction on the driver is expected to be quite different in the two schemes, as are the natures of the instabilities supposedly suppressed.

While this connection between the transformer ratio may turn out to be approximately true, we again stress that different figures-of-merit measure different characteristics, and their optimization will in general lead to different beam shapes; one must decide what one wants to optimize most.

Wakefield Amplitude, Wakefield Energy, and Particle Energy Gain

Unlike the transformer ratios R_2, R_3, R_4, or R_5, both $\max\big[|E_+|\big]\,L_{\rm diff}$ and $\max\big[|E_+|\big]\,L_{\rm pd}$ are in fact *monotonically-increasing* functions of the peak accelerating wakefield $\max\big[|E_+|\big]$ according to the above scaling laws, so in the linear or nonlinear, channeled or unguided cases, maximizing the peak wakefield will typically be equivalent to maximizing the *best-case* energy gain per particle. Admittedly, this does not necessarily mean that the actual average or typical energy gain per accelerated particle will also be maximized, because these simple scaling laws ignore the fact that the drive pulse is not characterized by a single peak intensity a_0^2 but by an entire envelope or photon distribution, and these simple estimates for work done by the wake ignore the fact that electrons in the bunch, with their own distribution of initial velocities and phases, will not all experience something close to the maximal wakefield over their entire respective intervals of acceleration, due to injection phase-slippage and beam loading effects

However, in the absence of detailed knowledge of the initial distribution of the electron beam, and of the corresponding probable electron trajectories in the wake, all depending on particulars of the injection process, self-consistent beam-loading, and other aspects of the specific experimental configuration, it would seem that the best we can do is strive for a large amplitude wake (assuming it also maintains reasonable coherence) and hope that a

suitable injection scheme may take advantage of it by producing a sufficiently short bunch of moderate charge at the right phase and velocity.

In analyses of the linear channel and of explicit bandwidth limitations, it is far more convenient to work with wake energy rather than peak wake amplitude. Even in the linear regime, in comparing wakefields there is no strict relationship between energy and peak amplitude, in that a wake with larger energy does not necessarily also exhibit a larger maximum. But in the linear regime with prescribed (i.e., non-depleting) driving fields, in either the homogeneous or channeled cases, it can be shown that maximizing peak wakefield amplitude or total wakefield energy under the same constraints on the envelope energy or duration of the drive pulse are equivalent. In nonlinear regimes, the wake energy and peak wake amplitude may not necessarily be optimized simultaneously. Since we typically envision an injection of a single electron bunch into the largest bucket just behind the drive, it is really the wake amplitude at and near the peak which remains of primary interest whether the goal is extent or overall efficiency of acceleration. If multiple buckets could be filled, then total wake energy might be of more direct interest.

2.7 Optimizations in the LWFA

Various optimizations and comparisons for the case of the LWFA are collected here, as always assuming a prescribed laser and quasi-static plasma response in the linear, channeled linear, or nonlinear regimes. Results for the nonlinear cases could be applied *a forteriori* to the corresponding linear cases as well, but we have presented the results more or less in the historical order in which they were developed, believing the demonstrations and results specialized to the linear regime remain physically informative even as technological improvements are nudging them towards obsolescence.

For each case, the optimal solutions will first be summarized and discussed, and then in some cases further details and demonstrations will be provided for the more interested reader, and in the hopes of helping to resolve certain lingering technical arguments in the field. Most of the work is analytic, but some preliminary numerical simulations[8] have also been performed.

[8]The PIC simulations summarized in the sequel were primarily carried out by R.R. Lindberg, using 1D XOOPIC[22].

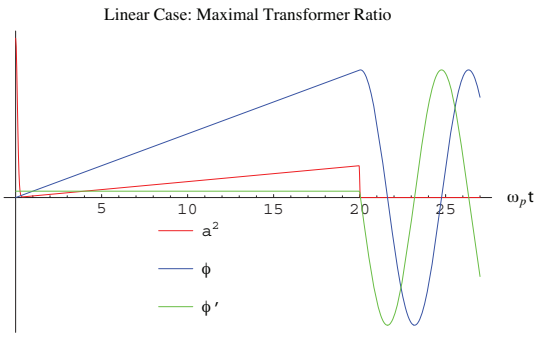

Figure 2.1. An example of the impulse-plus-ramp laser envelope maximizing the transformer ratio R_3 in the linear regime, along with the scaled potential and electric field that it excites.

2.7.1 Linear Regime: Maximizing the Transformer Ratio

In the linear regime, the laser pulse shape maximizing the *transformer ratio* R_3 subject to constraints on the *maximum gross driver pulse duration*:

$$\big\{\xi\colon a^2(\xi) > 0\big\} \subseteq [\xi_0\,,\,\xi_0 + T] \tag{2.75}$$

for some bound $T \geq 0$ (in scaled units) and some constant $\xi_0 \in \mathbb{R}$ that just determines the moment of pulse activation, and *total pulse envelope energy*:

$$\tfrac{1}{2}\int d\xi\, a^2(\xi) = U, \tag{2.76}$$

for some constant U, is the well known "impulse-plus-ramp" solution:

$$a^2(\xi) = 2\tilde{\epsilon}\big[\delta(\xi - \xi_0) + (\xi - \xi_0)\Theta\left(T - \xi + \xi_0\right)\Theta\left(\xi - \xi_0\right)\big], \tag{2.77}$$

where we have defined the normalization constant $\tilde{\epsilon} = \frac{U}{1+\frac{1}{2}T^2}$. A typical profile of this form and the resulting wakefield are shown in Fig. 2.1.

Despite inadequacies in published proofs, it can be demonstrated that this shape is indeed the *unique* optimum, irrespective of additional assumptions about the monotonicity of $\phi(\xi)$ within the driver pulse. Such an assumption was made in the original derivation, and while plausible, and in fact true, it may be deduced as an obvious consequence of the optimization shape rather than presumed unnecessarily as an *ansatz*. What we believe is the first complete proof for the LWFA case, inspired by the mechanical oscillator analogy, is presented below.

Note that for this parameterized family of pulses, the transformer ratio itself satisfies

$$R_3 \propto \sqrt{1 + T^2}, \tag{2.78}$$

and is independent of then envelope energy U and the absolute magnitude of the wakefield achieved, so that while specifying the energy is necessary to pick out a unique solution, it does not actually effect the value of the figure-of-merit. As a result, the driver energy U must be specified exactly rather than by an upper bound. Otherwise the same transformer ratio could be achieved with arbitrary small but non-zero driver energy in the limit as $U \to 0^+$.

Imposition of the upper bound on the pulse duration is also essential. As $T \to \infty$, it is easy to see that at any fixed energy U,

$$\max\big[|\mathcal{E}_-(\xi)|\big] \propto \tfrac{1}{2}\tilde{\epsilon} \to 0, \tag{2.79}$$

and

$$\max\big[|\mathcal{E}_+(\xi)|\big] \propto \tfrac{1}{2}\frac{\sqrt{1+T^2}}{U\left(1+\frac{1}{2}T^2\right)} \to 0 \tag{2.80}$$

but $R_3 \propto \sqrt{1+T^2} \to \infty$. That is, as the pulse duration is allowed to increase at a given pulse energy, the maximal accelerating and decelerating fields both become vanishingly small, but the accelerating field shrinks more slowly, so that their ratio diverges. From the transformer ratio perspective, without further constraints on the pulse duration, the most effective procedure is to drive the wake arbitrarily slowly and weakly for arbitrarily long times, leading to a large value for R_3 but low amplitude wakefields unsuitable for the LWFA.

This is reminiscent of energy transfer in conventional thermodynamic systems. In fact one can show that the maximum wakefield amplitude achieved by maximizing R_3 for any finite nonzero T is always less than what could be achieved instead by just putting all of the available envelope energy into the impulsive precursor alone, i.e., enforcing $T = 0$:

$$\frac{\mathcal{E}_+[T;U]}{\mathcal{E}_+[0;U]} = \frac{\sqrt{1+T^2}}{1+\frac{1}{2}T^2} \leq 1, \tag{2.81}$$

where this ratio approaches unity from above as $T \to 0^+$, but scales like $\sim \frac{2}{T} \to 0$ as $T \to \infty$. For given drive envelope energy, peak wakefield and transformer ratio are in fact inversely related.

Some PIC simulations were performed, replacing the unphysical impulsive precursor with an experimentally realizable Gaussian calculated to produce the desired maximum value of $\phi'(\xi) \ll 1$ somewhere behind the precursor, at which point the ramp is activated. Results were in excellent agreement with the predictions of linear theory, and are omitted, but they did confirm that the robustness of this pulse shape may be of some concern. The

value of the transformer ratio achieved degrades quickly as either the phase mismatch or strength mismatch between precursor and ramp increase.

Details for Optimizing the LWFA Transformer Ratio in the Linear Regime

Although the family of shapes optimizing the transformer ratio, consisting of an impulse followed immediately by a truncated ramp, has been discussed for both the LWFA[8, 9] and the PWFA[12], we present here a physically-motivated derivation, based on the oscillator analogy. The answer revealed by this argument has been verified by a more direct but less illuminating variational calculation. Also, our demonstration does not rely on certain plausible but unproved *a priori* assumptions of earlier derivations, which left open potential loopholes. In previous work[8, 9], for example, it was argued but not proven that the optimal solution must correspond to a monotonically increasing potential within the driver beam, the idea being that it would be inefficient for the electrostatic potential to rise, decline, then have to regain lost ground. While the optimized solution under reasonable constraints does turn out to correspond to a potential which is strictly increasing within the support of the driving beam, the demonstration is far from trivial, and it is simpler and safer to conclude this as a consequence of the optimal solution, rather than assume it as a requirement.

In the linearized LWFA theory, the transformer ratio associated with a driver source profile $s(\xi)$ can be defined as

$$R = R_3 = r_0 \frac{|E_+[s(\xi)]|}{|E_-[s(\xi)]|} = r_0 \frac{|\mathcal{E}_+[s(\xi)]|}{|\mathcal{E}_-[s(\xi)]|}, \tag{2.82}$$

where the numerator is the maximum magnitude of the accelerating gradient $\phi'(\xi)$ behind the pulse, the denominator is the maximum (more precisely, the supremum) of the magnitude of the decelerating gradient *strictly* inside the driving pulse (i.e., anywhere that $s(\xi) > 0$), both regarded as functionals of the entire drive profile $s(\xi)$, and r_0 is just a dimensionless normalization factor, given by

$$r_0 = \left(1 + \frac{\omega_0}{\omega_p}\right) \frac{m_e c^2}{\hbar \omega_0} \tag{2.83}$$

for the linear, laser-driven case applied to acceleration of electrons. (An equivalent definition applies to the linear particle-driven-case, except $r_0 = \frac{m_a}{m_d}$, the ratio of rest energies of accelerated and driver particles.)

The well known Fundamental Theorem of Beam Loading[12] implies an upper bound for the transformer ratio R in the case of a *symmetric* driving source, but does not restrict

R for more general profiles. In fact, strictly speaking, maximization of the transformer ratio within this dynamical model, with no further conditions or constraints imposed on the source, is not even well-defined; arbitrarily large values of R can always be achieved, using sufficiently long pulses. As a ratio of two potential energy gradients, R imposes no restriction on the duration of the driving pulse. It is implicitly assumed by Chen, *et al.*[8, 9] that they actually seek the pulse-shape of *specified maximum duration*, (i.e., with a prescribed bound on the size of the support of $s(\xi)$) that optimizes this transformer ratio. This is a natural constraint, since the goal is a short accelerating structure, and our dynamical description of the wake generation has really presumed that the plasma is much longer than the laser pulse. But even with such a bound on the total duration of forcing, the value of R, as a ratio of field gradients evaluated at two different points, will in the linear theory obviously be completely independent of any overall scaling of the laser amplitude, so in order to uniquely determine an optimal solution we must, in addition to the maximum pulse duration, also specify an additional energetic constraint – in this case an equality constraint on the envelope energy. It would seem natural to specify the total energy of the source, or equivalently (once the duration is also fixed) the average intensity of the source. Another possible constraint might be the *peak* normalized intensity of the source, which would necessarily lead to a different solution, which is not pursued further. After the constrained optimum is found, the dependence of the resulting wake on the pulse duration and either intensity or energy of the driving pulse can then be examined. Of course, to fully specify a solution, a definite start time ξ_0 for the pulse must also be specified.

To enforce the constraint on the maximal pulse duration and remove this trivial degeneracy in the optima arising from overall translation invariance, here we choose $\xi_0 = 0$ for convenience but without loss of any real generality, and suppose the laser envelope is non-zero *at most* in $0 \leq \xi \leq T$, vanishing elsewhere, for some prescribed maximal (scaled) duration $T \equiv k_p L_{\max} \geq 0$, where $L_{\max}$ is the maximal allowed pulse length in conventional units of distance. *A priori*, the laser field $a(\xi)$ may or may not vanish inside this interval, but we assume it definitely vanishes outside this range. We will also specify either the drive pulse envelope energy

$$\int_{-\infty}^{\infty} d\xi \, s(\xi) = \int_{0}^{T} d\xi \, s(\xi) = U, \tag{2.84}$$

or the average normalized pulse intensity,

$$\frac{1}{T}\int_{0}^{T} d\xi \, s(\xi) d\xi = \bar{s} \equiv \frac{U}{T}. \tag{2.85}$$

We now consider maximizing R, or equivalently but more conveniently R^2, over trial driver functions $s(\xi)$ which strictly vanish outside the interval $[0\,,\,T]$, but are nonnegative and almost everywhere piecewise continuous inside the interval, but possibly include a discrete sum of positive impulses, i.e., some finite number of Dirac delta-function terms with positive weights. Such impulsive contributions are unphysical, but are useful idealizations of narrow but intense peaks in the laser envelope. Finite discontinuities are also unphysical but useful idealizations of rapidly rising or dropping field envelopes.

For this class of generalized trial functions, it follows from the integral representations for the linear wake potential and gradient that the (normalized) potential $\phi(\xi)$ itself will be continuous everywhere, while the gradient $\phi'(\xi)$ will be continuous except at any point of application of an impulse, where the gradient will undergo an instantaneous positive jump. Note, however, that in order to treat an impulse consistently as the limiting case of a continuous, narrow, but intense peak in the source, we must consider the maximum gradient achieved anywhere inside the impulse to be that value obtained immediately *after* the impulse (in the sense of a limit from above), not the average of the values immediately before and after, which would perhaps the more natural value to assign to the gradient exactly at the instant of the impulse.

Also note that because the analogous mechanical oscillator system starts off at rest from equilibrium ($\phi(0) = \phi'(0) = 0$) and can only be forced in the positive direction ($s(\xi) \geq 0$), any possible negative "displacement" (regions where $\phi(\xi) < 0$) can occur only after the oscillator has first moved initially in the positive direction, encountered a turning point, and passed back through the equilibrium point with negative gradient. Because we can always add singleton points of *finite* forcing without affecting the wake dynamics at all, we may further assume without any loss of generality that the support of any trial function $s(\xi)$ is a *closed* subset of $[0\,,\,T]$; i.e., $s(\xi)$ may be considered as strictly positive at any point where the forcing is switching on or off discontinuously. Any proposed optimal source $s(\xi)$ must then first turn on at some point t_i and then permanently turn off at some point t_f, (still in scaled coordinates, despite the notation), where $0 \leq t_i \leq t_f \leq T$. The maximum accelerating gradient behind the pulse is then simply equal to the amplitude of the sinusoidal wake oscillation for $\xi \geq t_f^+$, and the square of this amplitude, corresponding to the total

energy of the analogous oscillator system, can be written as:

$$\begin{aligned}\mathcal{E}_+^2 \equiv \mathcal{E}^2(t_f^+) &= \phi'(t_f^+)^2 + \phi(t_f)^2 \\ &= \phi'(t_f^+)^2 + \left[\int_{t_i}^{t_f} d\xi\, \phi'(\xi)\right]^2,\end{aligned} \tag{2.86}$$

or after some manipulation, as:

$$\mathcal{E}_+^2 = 2\int_{t_i}^{t_f} d\xi\, \left[\phi''(\xi) + \phi(\xi)\right] \phi'(\xi) = 2\int_{t_i}^{t_f} d\xi\, s(\xi)\phi'(\xi), \tag{2.87}$$

which is analogous to the total *external* work (i.e., including the source but excluding the restoring force) performed on the oscillator.

Over the allowed class of pulse shapes, the supremal decelerating field functional $\mathcal{E}_-[\,]$ will be quite complicated in its functional form, but its actual value $\mathcal{E}_- \equiv |\mathcal{E}_-[s(\xi)]| \geq 0$, when evaluated at an optimal pulse shape $s(\xi)$ consistent with the given energy and pulse-duration constraints, is just some definite, nonnegative number, and so R^2 can be written as:

$$R^2 = r_0^2 \frac{\phi'(t_f^+)^2 + \phi(t_f)^2}{\mathcal{E}_-^2} = r_0^2 \frac{\phi'(t_f^+)^2 + \left[\int_{t_i}^{t_f} d\xi\, \phi'(\xi)\right]^2}{\mathcal{E}_-^2}. \tag{2.88}$$

Because the oscillator must first move in the positive direction before it can swing negative, it follows that if $\mathcal{E}_-^2 = 0$, then in fact $\phi'(\xi) = 0$ everywhere, implying no forcing was applied and no wake was produced, so this ratio is in fact well-defined for all cases of interest, for which $\mathcal{E}_- > 0$ however small, and we can imagine maximizing R^2 by maximizing $\mathcal{E}_+^2$ for $\mathcal{E}_-^2$ fixed at its *as yet unknown* optimizing value. Since $s(\xi) > 0$ by assumption at $\xi = t_f$, it follows by definition that $\phi'(t_f) \leq \phi'(t_f^+) \leq \mathcal{E}_-$. However, if $\phi'(t_f) < 0$, then from it is clear that the positive forcing present is actually removing, rather than adding, energy to the wake, at this instant. We can infer that any optimal shape must have $\phi'(t_f) \geq 0$, since otherwise more energy could have been delivered to the wake simply by stopping the forcing earlier, ideally at or prior to the most recent positive turning point. It therefore follows that for any optimal pulse shape, $0 \leq \phi'(t_f^+) \leq \mathcal{E}_-$; but since we can always apply an impulse at or before t_f to kick the gradient up to $\mathcal{E}_-$, without affecting the potential, we can in fact always achieve $\phi'(t_f^+) = \mathcal{E}_-$ exactly, which is therefore the optimal value.

Our goal then turns to maximizing the "potential energy" contribution $\phi(t_f)^2$. Again, because positive forcing at points of negative gradient implies that the oscillator can only lose and never gain energy while $\phi'(\xi) < 0$, it follows that the value of $\phi(\xi)$ achieved at

any negative turning point occurring during the driving interval, and therefore at any point whatsoever where the potential is negative, can never exceed in absolute value that level reached at the most recently encountered positive turning point. We infer that for any optimal pulse shape, $\phi(t_f) \geq 0$, for otherwise a larger potential and wake energy could have been reached simply by forcing harder at or before the most recent positive turning point, in order to prevent the potential from ever becoming negative at $\xi = t_f$.

Next, suppose for the moment that the source $s(\xi)$ is of connected support, i.e., it consists of a single pulse of some shape, strictly positive throughout the interval $[t_i\,,\,t_f]$ but zero everywhere else, and does not switch off during any intermediate interval and then switch back on later. It then follows that $0 \leq \phi(t_f) = \int_{t_i}^{t_f} d\xi\, \phi'(\xi) \leq \int_0^T d\xi\, \mathcal{E}_- = \mathcal{E}_- T$, so in the case of a connected pulse we see that R^2 can become no larger than

$$R^2 \leq r_0^2 \frac{\mathcal{E}_-^2 + \mathcal{E}_-^2 T^2}{\mathcal{E}_-^2} = r_0^2(1 + T^2). \tag{2.89}$$

To actually achieve this upper bound, the system evidently must be forced in such a way that $\phi'(\xi) = \mathcal{E}_-$ almost everywhere within $[0, T]$. It is easy to see that this can happen only by initially delivering an impulse $\mathcal{E}_-\delta(\xi)$ at $\xi = 0$ to instantaneously kick the gradient $\phi'(\xi)$ up to $\mathcal{E}_-$ at $\xi = 0^+$, and then immediately applying a linear external force $\mathcal{E}_-\xi$ throughout $0 < \xi \leq T$ to exactly balance the (internal) restoring force $-\phi(\xi)$, thereby maintaining a constant gradient at the maximally-allowed value $\mathcal{E}_-$, while allowing $\phi(\xi)$ to grow linearly, until the forcing is abruptly turned off at $\xi = T$, after which the wake rings freely, with squared magnitude $\mathcal{E}_+^2 = R^2 E_-^2$.

It now remains to consider the possibility of an overall driver shape consisting of two or more disconnected sub-pulses. Intuitively, we expect such a driver shape to be sub-optimal, since an interval with no forcing amounts to a wasted opportunity for adding energy to the wake, but it actually requires a somewhat involved argument to demonstrate this carefully. We argue by contradiction. Let the impulse-plus-ramp source, found above under the assumption of a source of connected support, now be denoted by $s_1(\xi)$. We have seen how this source results in a potential which grows linearly inside the driving pulse, i.e., $\phi_1(\xi) = \mathcal{E}_-\xi$ for $\xi \in [0, T]$, as shown in Fig. 2.1. Suppose a disconnected pulse shape $s_2(\xi)$ with some pattern of intermittent forcing remains consistent with the constraints on duration and the maximum decelerating gradient but leads to an equally good or better wake potential solution $\phi_2(\xi)$. The situation would have to look qualitatively something like that depicted in Fig. 2.2. The new solution must result in a wake with $\phi_2'(T_+) = \mathcal{E}_-$ but $\phi_2(T) \geq \phi_1(T) = \mathcal{E}_- T$. But because of the assumed initial conditions and bound

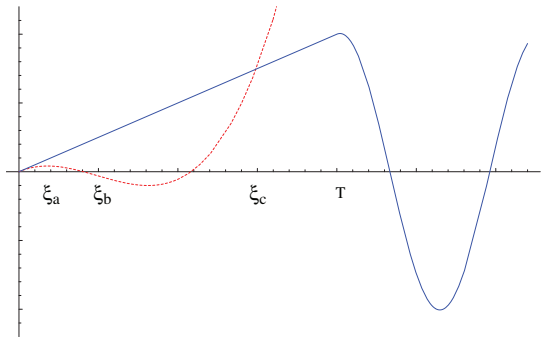

Figure 2.2. The potential due to the impulse-plus-ramp solution, together with an ostensibly superior solution which is in fact physically impossible. The various labeled points are explained in the text.

on the gradient during forcing, no potential can grow any faster than $\phi_1(\xi)$ as $\xi \to 0$, nor indeed anywhere where any forcing is present. The proposed $\phi_2(\xi)$ cannot therefore initially outgrow $\phi_1(\xi)$, and in fact must at some point begin to fall below $\phi_1(\xi)$, since by construction $s_2(\xi)$ switches off at some intermediate point, allowing the restoring force to pull the potential back toward the origin. But if by $\xi = T$ it is to match or exceed $\phi_1(\xi)$, then $\phi_2(\xi)$ must at some later point cross $\phi_1(\xi)$ from below, and we denote the first such instant by ξ_c. So $\phi_2(\xi_c) = \phi_1(\xi_c)$, while $\phi_2'(\xi_c^-) \geq \phi_1'(\xi_c)$ and hence $\mathcal{E}_2^2(\xi_c^-) \geq \mathcal{E}_1^2(\xi_c)$.

We can also infer that forcing is definitely not being applied in some interval immediately prior to this point ξ_c where $\phi_2(\xi)$ catches up to $\phi_1(\xi)$, because the latter potential is growing at the maximum rate possible with forcing present, and to catch up and then surpass $\phi_1(\xi)$ our solution $\phi_2(\xi)$ obviously must be growing faster (i.e., possess a larger positive slope, which is forbidden by the constraint on maximum gradient in the presence of forcing.) So $s_2(\xi_c^-) = 0$, and indeed must have been fixed at zero since some earlier instant, say ξ_b, where it was last nonzero. Since no external work is being done by $s_2(\xi)$ for $\xi_b < \xi < \xi_c$, it follows that $E_1^2(\xi_c^-) = E_1^2(\xi_b^+)$. By construction, $\phi_2(\xi_b) \leq \phi_1(\xi_b) < \phi_1(\xi_c)$. Because of the initial conditions and the nonnegativity of the forcing, no potential can swing negative more than it has previously moved positive, so in fact we have

$$|\phi_2(\xi_b)| \leq \max_{0 \leq \xi \leq \xi_b} \left[\phi_2(\xi)\right]) \leq \max_{0 \leq \xi \leq \xi_b} \left[\phi_1(\xi)\right] = \phi_1(\xi_b) < \phi(\xi_c), \tag{2.90}$$

which, combined with our knowledge that

$$\mathcal{E}_2^2(\xi_b^+) \geq \mathcal{E}_1^2(\xi_c^-), \tag{2.91}$$

implies that

$$\left|\phi_2'(\xi_b^+)\right| > \left|\phi_2'(\xi_c^-)\right| \geq \mathcal{E}_-. \tag{2.92}$$

Since $s_2(\xi_b) > 0$ by construction, $\phi_2'(\xi_b^+) < \mathcal{E}_-$, so it must be the case that $\phi_2'(\xi_b^+) < -\mathcal{E}_- < 0$.

This negative potential gradient at ξ_b, together with the knowledge that $\phi'(\xi_c) > \mathcal{E}_-$, in fact implies that, between ξ_b and ξ_c, the potential must pass through a negative turning point. The negative gradient also implies that sometime prior to ξ_b, the potential passed through a positive turning point, most recently at some point ξ_a, such that $\phi_2'(\xi_a) = 0$ while $\phi'(\xi) < 0$ for $\xi_a < \xi \leq \xi_b$, during some of which time we know forcing must have been present. Since positive forcing at points of negative gradient removes energy, it follows that the system can have no more energy at ξ_b than at ξ_a and in fact, we find

$$\mathcal{E}_2^2(\xi_c^-) = \mathcal{E}_2^2(\xi_b^+) < \mathcal{E}^2(\xi_a) = \phi_2^2(\xi_a) \tag{2.93}$$

and

$$\phi_2^2(\xi_a) < \phi_1^2(\xi_a) < \phi_1^2(\xi_c) < \mathcal{E}_1^2(\xi_c), \tag{2.94}$$

but

$$\mathcal{E}_1^2(\xi_c) \geq \mathcal{E}_2^2(\xi_c^-), \tag{2.95}$$

so we have arrived at a contradiction, and we can finally conclude by *reductio ad absurdem* that the globally optimal shape must be of simply-connected support as expected, and is indeed given by the solution found above:

$$\begin{aligned} s(\xi) = \tfrac{1}{2}\left|a(\xi)\right|^2 &= \Theta_0(\xi)\Theta_1(T-\xi)\mathcal{E}_-\left\{\delta_0(\xi)+\xi\right\} \\ &= \mathcal{E}_-\delta(\xi) + \mathcal{E}_-\xi\Theta(\xi)\Theta(T-\xi) \end{aligned} \tag{2.96}$$

for any choice of $T > 0$, or

$$s(\xi) = \mathcal{E}_-\tfrac{d}{d\xi}\Theta_0(\xi) = \mathcal{E}_-\delta_0(\xi), \tag{2.97}$$

for the degenerate case $T = 0$, and where the number $\mathcal{E}_- \geq 0$ must now be determined by the duration T and the remaining energetic or intensity constraint on $s(\xi)$. In terms of a given envelope energy $U = T\bar{s}$ in the pulse, a simple calculation reveals that $\mathcal{E}_-$ is given by

$$\mathcal{E}_- = \frac{U}{1+\frac{1}{2}T^2}. \tag{2.98}$$

This laser drive leads to a wake potential given by

$$\phi(\xi) = \Theta(\xi)\Theta(T-\xi)\,\mathcal{E}_-\xi + \Theta(\xi)\left[1-\Theta(T-\xi)\right]\left\{\mathcal{E}_-\sin(\xi-T) + \mathcal{E}_-T\cos(\xi-T)\right\}, \tag{2.99}$$

a maximized transformer ratio

$$R = r_0\sqrt{1+T^2}, \tag{2.100}$$

and a peak accelerating wake amplitude

$$\mathcal{E}_+ = r_0 \frac{\sqrt{1+T^2}U}{1+\frac{1}{2}T^2} \tag{2.101}$$

behind the drive pulse.

2.7.2 Linear Regime: Maximizing the Wakefield Amplitude

Instead of concentrating on the transformer ratio, consider the more obvious maximization of the *peak accelerating wake amplitude* $\max \mathcal{E}_+ = \left[|\mathcal{E}_+(\xi)|\right]$ in the linear regime, subject only to a constraint on the *maximum envelope energy*, i.e.,

$$\tfrac{1}{2}\int d\xi\, a^2(\xi) \le U, \tag{2.102}$$

for some constant upper bound $U \ge 0$. This leads to optimal pulse shapes of the impulsive comb form:

$$a^2(\xi) = \sum_{j=1}^{N_a} \alpha_j^2\, \delta\left(\xi - \xi_0 - 2\pi n_j\right) \tag{2.103}$$

for some integer $N_a \ge 1$, integers $n_j \in \mathbb{Z}$, $j = 1, \ldots, N_a$, an overall real temporal offset $\xi_0 \in \mathbb{R}$, and positive constants $\alpha_j > 0$ satisfying the normalization condition

$$\tfrac{1}{2}\sum_{j=1}^{N_a} \alpha_j^2 = U; \tag{2.104}$$

that is, either a single impulse or a series of impulses spaced at integral multiples of the linear plasma period $\mathcal{T}_p = \frac{2\pi}{\omega_p}$. As expected, the optimal solution makes use of all of the driver energy assumed available.

Although Dirac delta functions have infinite intensity and infinite bandwidth and hence are obviously unphysical, this result indicates that, for bounded driver energy, the linear wake response continues to increase *monotonically* towards an asymptotic upper bound as the pulse width decreases, even for narrow pulses shorter than the "resonant" duration $\frac{1}{2}\mathcal{T}_p$, at least as long as group velocity dispersion is ignored, notwithstanding conventional wisdom which presumed that a resonant width would always be optimal.

However, the expected "resonance" maximum does emerge around a pulse width of $\Delta\xi \sim O\left(\frac{1}{2}\mathcal{T}_p\right)$ if the *peak normalized envelope intensity* is constrained rather than total energy, because then the energy available for wake excitation decreases as the pulse width narrows.

Specifically, with an upper bound imposed on the normalized peak pulse intensity, i.e.,

$$a^2(\xi) \leq a_0^2 \tag{2.105}$$

and an upper bound imposed on the total "gross" pulse duration from "first-on" to "last-off," i.e.,

$$\sup\left\{\xi \mid a^2(\xi) > 0\right\} - \inf\left\{\xi \mid a^2(\xi) > 0\right\} \leq T, \tag{2.106}$$

for some constant $T > 0$, the optimal pulse shape can be written as

$$a^2(\xi) = \begin{cases} a_0^2 & \text{if } \xi_0 \leq \xi \leq \xi_0 + T \text{ and } \phi'(\xi) \geq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{2.107}$$

for some arbitrary start time determined by $\xi_0 \in \mathbb{R}$. In the linear case, the plasma wave remains purely sinusoidal with frequency ω_p and wavenumber k_p, so this solution obviously consists of a series of square pulses, each except possibly the last[9] of exactly the resonant width $\frac{1}{2}\mathcal{T}_p$, and with their abrupt onsets separated by exactly one full linear plasma period $\mathcal{T}_p$. Driving at any intermediate values of intensity, or for any longer than a half-period within any constituent sub-pulse (or for any shorter period for any but possibly the last sub-pulse before turn-off), is definitely sub-optimal. This is an example of the well-known "bang-bang" theorem from linear control theory, which specifies conditions under which an all-or-nothing solution is indeed best. More will be said when we extend these results to the the more general nonlinear case.

If in addition to peak intensity, instead of limiting the gross pulse duration T, the total envelope energy is instead constrained by $\frac{1}{2}\int d\xi\, a^2(\xi) \leq U$, then no global optimum actually exists: with the allowed energy apportioned into an ever larger number N_a of ever thinner square pulses of intensity a_0^2, spaced exactly in resonance with ω_p, the excited wake behind the pulse can in principle asymptotically approach that which would be excited by a single delta-function source $a^2(\xi) \to \frac{1}{2}U\,\delta(\xi - \xi_0)$ belonging to the family of optimal solutions found above in the absence of any intensity constraints, leading in principle to $\max\left[|\mathcal{E}_+(\xi)|\right] \to \frac{1}{2}U$ as $N_a \to \infty$. Similar reasoning applies to the case where the intensity and the "net" rather than "gross" pulse duration is constrained.

Of course, bandwidth restrictions and dispersion will ultimately limit how abruptly the drive may be turned on or off, or how narrow any of its constituent sub-pulses may become, and very precise timing of the change-points where the drive switches on or off may not be technologically feasible.

[9] Of course, "last" can mean the first if the constraints allow time for part of one pulse, i.e.if $T < \mathcal{T}_p$.

Numerical Comparisons

In order to assess the validity of the linear analytic theory, and simultaneously of our use of eikonal-based definitions for energy and intensity, as well as to confirm the important differences in behavior expected when constraining intensity or energy, we performed a number of 1D PIC-based numerical simulations,[10] and typical results for the linear regime are summarized in Fig. 2.3, where we plot the plasma response to Gaussian laser envelopes of varying width (measured here by a standard deviation σ in physical units, or the corresponding $\tilde{\sigma} = \omega_p \sigma$ in scaled units), while keeping either the peak intensity or actual energy fixed.

That is, we have used a Gaussian drive envelope of the form $a^2(\xi) = a_0^2 \exp(-\xi^2/\tilde{\sigma}^2)$, corresponding to an actual scaled electric field

$$\mathcal{E}_x(\xi) = a_0 \tfrac{\omega_0}{\omega_p} e^{-\frac{\xi^2}{2\tilde{\sigma}^2}} \left[\cos\left(\tfrac{\omega_0}{\omega_p}(\xi - \xi_0) \right) - \xi \tfrac{\omega_p}{\omega_0 \tilde{\sigma}^2} \sin\left(\tfrac{\omega_0}{\omega_p}(\xi - \xi_0) \right) \right]. \tag{2.108}$$

For definiteness, we choose a relative carrier phase such that $\xi_0 = 0$, and then the constant a_0 is determined by the relevant constraint. The peak instantaneous intensity, i.e., absolute maximum of $\mathcal{E}(\xi)^2$, is cumbersome to locate exactly, but should always be close to $\mathcal{E}_x(0)^2 = \frac{\omega_0^2}{\omega_p^2} a_0^2$ for $\xi_0 = 0$ even as $\tilde{\sigma} \to 0$, so instead of trying to find a precise value which is usually not experimentally known anyway, we will continue to just fix a_0. To instead fix the scaled drive energy $\mathcal{H}_d = \int d\xi \, \mathcal{E}(\xi)^2$, we choose a_0 according to

$$a_0^2 = \tfrac{1}{\sqrt{\pi}} \tfrac{\omega_p^2}{\omega_0^2} \frac{4 \mathcal{H}_d \tilde{\sigma}}{1 + 2\omega_0^2 \sigma^2 + e^{-\omega_0^2 \sigma^2}}, \tag{2.109}$$

rather than fixing the corresponding envelope energy U based on the eikonal assumption, or envelope approximation, that $\mathcal{E}(\xi) \sim \frac{\omega_0}{\omega_p} a(\xi)$, which would lead to $a_0^2 = \frac{2}{\sqrt{\pi}} \frac{U}{\tilde{\sigma}}$.

At fixed peak intensity, the plasma response is exactly as expected. Long, weak pulses allow the plasma wave to oscillate inside the drive pulse, which can then remove energy from the wake in regions where $\phi'(\xi) < 0$, so the overall response behind the pulse tends to average away as $\omega_p \sigma \to \infty$. Very short pulses of fixed peak a^2 contain very little energy with which to excite the wake, so $\mathcal{E} \to 0$ as $\omega_p \sigma \to 0$ as well. The plasma response peaks near a "resonant" width corresponding to $\omega_p \sigma \approx 1.45$, or FWHM $\approx 0.54 \ \lambda_p$, where the

[10] Here, and below, PIC simulations were performed using 1D XOOPIC with the following physical and numerical parameters, unless otherwise specified: the linearly-polarized laser was launched from vacuum with wavelength $\lambda = 1 \ \mu$m into a quiescent plasma with density rising to about $n_0 \approx 10^{19}$ cm^{-3} corresponding to $\frac{\omega_p}{\omega_0} = 10$. Generally a grid with 32 cells per wavelength was used, with 16 macro-particles per cell. The simulation region included an initial vacuum section of length $64 \, \lambda$ to minimize particle loss at the boundaries (important in the nonlinear regime), then linearly ramped the density up to its bulk value over $16 \, \lambda$ to avoid reflective losses of the laser, and finally included homogeneous plasma over a length of about $240 \, \lambda$, or $24 \lambda_p$.

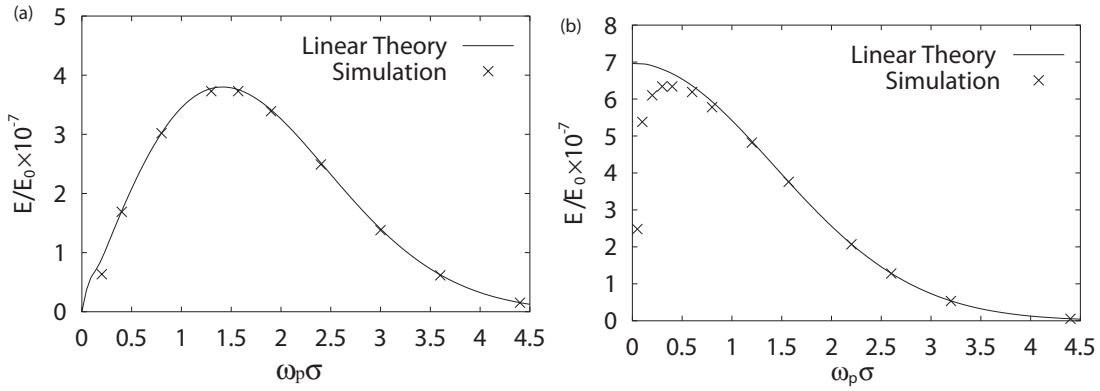

Figure 2.3. Linear plasma response E/E_0 for Gaussian pulses of fixed (a) normalized intensity and (b) vacuum electromagnetic energy. Theory solves for the linearized (harmonic) plasma response using analytic integration of the Green function, constraining either U or a_0. Numerical simulations were performed for $\omega_0/\omega_p = 10$ and (a) $a = 0.001$, (b) $a = 0.001$ for $\omega_p\sigma \sim 1.5$, coinciding with the former pulse at the approximate point of its peak performance. For short pulses, $\sigma \sim O(\lambda)$, the envelope approximation breaks down, as seen in (b).

intrinsic advantages of shorter pulses are balanced with the decreasing amount of total energy available.

For fixed energy, the numerical results follow the linear theory until very short pulse lengths ($\omega_p\sigma \lesssim 0.6$, or equivalently $\sigma \lesssim 2\lambda_0$) are reached. In small part, this may be due to group velocity dispersion, where initially Gaussian pulses of short duration σ do not remain so. However, it mostly results from a breakdown in the eikonal assumption in determining the pulse energy constraint. A source $s(\xi) = \frac{1}{2}a(\xi)^2$ that would converge to a delta function as $\sigma \to 0$ if U were actually fixed instead has its amplitude a_0 going to zero for fixed $\mathcal{H}_d$ according to (2.109).

While this breakdown in the envelope approximation may be cause for some alarm, its effects only manifest for extremely narrow pulses where $\sigma \sim O(\lambda_0)$, and our general conclusion hold over a wide range of pulse lengths. For a given amount of energy, the plasma response continues to increase as the pulse narrows below the resonant duration $\frac{1}{2}T_p$, down to about $\sigma \sim O(2\,\lambda_0)$. In practice, narrower pulses become increasingly difficult to achieve experimentally and suffer increasingly egregious dispersion, making them unsuitable for LWFA applications.

Details and Proofs for Maximizing Peak Wakefield Amplitude

Now we seek to maximize the peak magnitude $|\mathcal{E}_+|$ of the (accelerating) wakefield, subject to either a constraint on envelope energy or constraints on duration and normalized intensity.

Fixed Envelope Energy

First we maximize the peak magnitude with an upper bound on the envelope energy of the driving beam: $\int_{-\infty}^{\infty} d\xi\, s(\xi) \leq U$. In order to facilitate comparison with the maximal-transformer-ratio solution, for now we still neglect any other, potentially important, constraints, such as restrictions on frequency content of the driving pulse due to laser bandwidth limitations. Because the wake response is here assumed linear, it is clear that we should make use all of the pulse energy available; if a proposed optimal pulse profile contains less than the allowed energy, we can create a larger wake just by scaling up the pulse amplitude everywhere without changing its relative shape. Hence we can take the pulse-energy limitation to be an equality, rather than inequality, constraint, which will be more convenient mathematically. By assumption, the wakefield is zero before the driver turns on, so obviously the absolute field maximum cannot occur ahead of the driving laser pulse. Since we are here ignoring bandwidth restrictions on the driving pulse and can hence turn the pulse on or off as rapidly as we please, it is also evident that the optimal wake generation need never achieve the globally maximum gradient strictly before the end of the driving pulse proper, for if this were to occur, a wake at least as large and generally larger could be generated strictly behind the pulse, and with less driver energy, by simply truncating the driving pulse at the instant of this gradient maximum. Thus we can infer that the optimal laser pulse shape will vanish after some point, say $\xi = t_f$ (in scaled coordinates, despite appearances), will have envelope energy equal to the maximal allowed energy U, and will maximize $\mathcal{E}_+ {t_f^+}^2 = \phi'(t_f^+)^2 + \phi(t_f)^2$ consistent with these constraints. Maximization of the squared amplitude of the post-pulse wake in the form

$$\mathcal{E}_+^2 = \int_{-\infty}^{t_f} d\xi\, s(\xi)\phi'(\xi) = \int_{-\infty}^{t_f} d\xi \int_{-\infty}^{\xi} d\xi\, s(\xi)s(\xi')\cos(\xi - \xi'), \tag{2.110}$$

such that $\int_{-\infty}^{t_f} d\xi\, s(\xi) = U$, results, after some manipulation, in the variational Euler-Lagrange equation:

$$s(\xi) \left\{ \int_{-\infty}^{t_f} d\xi' \, \left[\cos(\xi - \xi') s(\xi') \right] - \lambda \right\} = 0, \tag{2.111}$$

where λ is a constant Lagrange multiplier whose value is chosen so that both this stationarity condition (2.111) and the energy constraint are satisfied. We must take care to find a true maximum, rather than some other stationary solution, and also ensure that the important positivity constraint $s(\xi) \geq 0$ on the source is satisfied, which has not yet been explicitly imposed. Integrating the stationarity condition, re-arranging, and imposing the energy constraint, we can write the multiplier λ formally as:

$$\begin{aligned} \lambda &= \frac{\int_{-\infty}^{t_f} d\xi \int_{-\infty}^{t_f} d\xi' \, \cos(\xi - \xi') s(\xi) s(\xi')}{\int_{-\infty}^{t_f} d\xi s(\xi)} \\ &= 2U^{-1} \int_{-\infty}^{t_f} d\xi \int_{-\infty}^{t_f} d\xi' \, \Theta(\xi - \xi') \cos(\xi - \xi') s(\xi) s(\xi') = \frac{\mathcal{E}_+^2}{U}, \end{aligned} \tag{2.112}$$

where $\mathcal{E}_+^2$ is the actual value of (scaled) wake "oscillator energy" (squared-magnitude) achieved. The stationarity condition demands that almost everywhere either $s(\xi) = 0$ or $\int_{-\infty}^{t_f} d\xi' \, \cos(\xi - \xi') s(\xi') = \lambda$. The latter condition can be re-written as:

$$\cos(\xi - t_f) \int_{-\infty}^{t_f} d\xi' \, \cos(t_f - \xi') s(\xi') + \sin(\xi - t_f) \int_{-\infty}^{t_f} d\xi' \, \sin(t_f - \xi') s(\xi') = \lambda, \tag{2.113}$$

or

$$\cos(\xi - t_f) \phi'(t_f) + \sin(\xi - t_f) \phi(t_f) - \lambda = 0. \tag{2.114}$$

Because the functions $\sin(\xi)$, $\cos(\xi)$, and 1 are linearly independent as vectors in any $\mathcal{L}^2[\xi_1, \xi_2]$ Lebesgue space (for $-\infty \leq \xi_1 < \xi_2 \leq \infty$) , the latter term can vanish over some interval $\xi \in (\xi_1\,,\, \xi_2)$ only if $\phi(t_f) = \phi'(t_f) = \lambda = 0$, i.e., only in the uninteresting case in which no forcing is applied and no wake is produced. So in fact $s(\xi) = 0$ almost everywhere, implying that the energy in the driving pulse is concentrated in a discrete set of impulses, i.e., $s(\xi)$ can be taken to be a sum a Dirac delta-functions:

$$s(\xi) = \sum_i s_i \, \delta(\xi - t_i), \tag{2.115}$$

where we can assume $t_i \leq t_f$, and such that the real coefficients s_i, $i = 1, 2, 3, \ldots$ must satisfy the energy constraint $\sum_i s_i = U$ as well as the nonnegativity requirements $s_j \geq 0$ and the stationarity conditions $\lambda = \sum_i s_i \cos(t_i - t_j)$ for each impulse j. Since $\cos(t_i - tj) \leq 1$, and both $\mathcal{E}_+^2 \geq 0$ and $U \geq 0$, the stationarity conditions imply

$$\mathcal{E}_+^2 = U \sum_i s_i \cos(t_i - t_j) \leq U^2, \tag{2.116}$$

with equality if and only if $\cos(t_i - t_j) = 1$ for all impulses i and j. We deduce that the maximal wake response $\mathcal{E}_+^2 = U^2$ is produced by either a single delta-function driving pulse:

$$s(\xi) = U\,\delta(\xi - t_f) \tag{2.117}$$

or as a series of delta-functions spaced at integral multiples of the plasma frequency:

$$s(\xi) = \sum_i s_i\,\delta(\xi - t_f + 2\pi n_i), \tag{2.118}$$

where the $n_i \in \mathbb{Z}$ are all integers and $\sum_i s_i = U$. Of course, as in the transformer ratio case, any appearance of Dirac delta-functions in the driver envelope is unphysical, strictly speaking. Any impulse would have infinite bandwidth, which cannot be achieved with actual laser amplifiers, as well as spectral content at arbitrarily low frequencies below ω_p, which cannot be supported in the plasma. It would also induce an infinite quiver response, and suffer maximal dispersion and therefore distortion and broadening during propagation in the plasma, both of which would invalidate the simple linear dynamics used here. It is not clear what it would mean to take the square root of the delta function to determine $a(\xi)$ from $s(\xi)$. But the result does reveal that for fixed driver energy, and within the linear electrostatic model, the wake amplitude increases toward its asymptotic maximum as the driving pulse narrows in width and correspondingly grows in peak intensity, provided dispersion is neglected.

Fixed Normalized Intensity

Instead of bounding the envelope energy, we turn to the case where we bound instead the normalized intensity: $a(\xi)^2 \leq a_0^2$. To avoid a divergent wake response over arbitrary long times, and to obtain a physically realizable pulse, we must also bound the pulse duration in some way. Here we choose an upper bound T on the total "gross" pulse duration (in scaled units).

It is fairly simple to deduce that the optimal solution must be of the specific bang-bang form described above. From the form (2.31) for the first integral in the linear case, it is clear

that for any necessarily nonnegative ponderomotive source $s(\xi)$, the wake energy density is increased for any forcing applied during intervals[11] in which $\phi'(\xi) > 0$. Conversely, any forcing applied at points where $\phi'(\xi) < 0$ only serves to remove energy from the wake. In the mechanical oscillator analogy, we should only force the oscillator when it is already swinging in the direction we can push, and never perform any negative work on the system to slow it down.

Because the wave is assumed to remain linear, no amount of earlier forcing can change the periodicity with which the critical points of $\phi(\xi)$ occur. Also, any previous forcing applied during advantageous intervals only increases the subsequent wake energy and therefore increases the average value of $\phi'(\xi)$ that would occur, in the absence of any subsequent forcing during the next propitious interval. Thus when properly timed, previous forcing never decreases the opportunities for subsequent forcing nor the efficacy of subsequent properly-timed forcing, and in general previous forcing enhances the work that can be delivered by appropriately-timed subsequent forcing.

Therefore, to maximize the work performed, and the resulting wake amplitude, we should force the system as strongly as allowed during any and all allowed half-periods (or a fraction thereof, if the duration constraint forces the pulse to be terminated) where $\phi'(\xi > 0)$, and then turn the forcing completely off for the other half-periods.

2.7.3 Wake in a Channel

In the LWFA scheme, we have seen that of all aspects of the dynamics of and the back-reaction on the laser drive, diffractive effects typically operate most quickly to distort the beam shape and disrupt effective beam-wake coupling, unless they are counteracted by relativistic and/or ponderomotive self-focusing in an intense beam or by suitable guiding in a density channel. Therefore it makes sense to next consider the effects of a channel counteracting diffraction, before considering any nonlinear effects in the plasma interaction or pulse propagation. Since in the absence of guiding acceleration will otherwise be severely limited by diffraction, all current LWFA experiments use density channels to guide the laser drivers, produced either through hydrodynamic expansion of the plasma induced by heating lasers, or by photo-ionization inside capillaries, although schemes using dynamical density perturbations consisting of additional, suitably-phased plasma waves have also been proposed.

[11] These are the same intervals over which electrons would be decelerated.

Many generalizations (or shall we say complications) are possible, but we will confine attention to the simplest extension of our model, that of linear wake dynamics driven by a matched beam in an effectively two-dimensional density channel, with translational symmetry in the longitudinal direction and either translational (i.e., slab geometry) or cylindrical symmetry in the transverse direction. Commonly studied examples are either the parabolic or hollow (i.e., step-function) transverse profile in either slab or cylindrical geometry.

In a cold but collisionless homogeneous plasma, recall that the Green function for the wakefield $\phi'(\xi)$ in the linear regime is purely sinusoidal: $G(\xi, \xi') = \Theta(\xi - \xi') \cos(\xi - \xi')$. In a symmetric channel, the excited eigenmodes need not be purely electrostatic, but can have a high-phase velocity longitudinal electric field suitable for electron acceleration. For an eikonal drive laser of fixed carrier frequency, wavenumber, and transverse profile, the Green function will be separable in longitudinal and transverse coordinates, and may generally be written as an infinite sum over complex exponentials in ξ multiplying trigonometric, hyperbolic, Bessel, Hermite, or other hypergeometric functions in the (scaled) transverse coordinates $\boldsymbol{r} = k_p \left[x\hat{\boldsymbol{x}} + y\hat{\boldsymbol{y}} \right]$, although the precise form of the Green function will not be essential to our arguments here. Leakage out of the channel can occur, so the response generally decays behind the excitation. Numerical evidence based on fluid simulations[12] suggests that at any given transverse position $\boldsymbol{r}$ inside the channel, the Green function can be reasonably approximated by a sum of a *small* number of decaying exponentials and/or damped sinusoids:

$$G(\xi, \xi'; \boldsymbol{r}) \propto \Theta(\xi - \xi') \left[e^{\beta_1(\boldsymbol{r})(\xi - \xi')} + e^{\beta_2(\boldsymbol{r})(\xi - \xi')} + \cdots \right], \tag{2.119}$$

where $\beta_j(\boldsymbol{r}) \in \mathbb{C}$ are continuous in $\boldsymbol{r}$, with $\mathrm{Re}[\beta_j] < 0$ strictly if $\beta_j \neq 0$ for $j = 1, 2, 3, \ldots$. An example is depicted graphically in Fig. 2.4.

For any such Green function, the pulse shape maximizing peak wake amplitude or total wake energy, either on-axis, or weighted by some transverse beam profile, for fixed driver energy U, can also be shown to be a single impulse: $a^2(\xi) = U\, \delta(\xi - \xi_0)$.

For given energy, because of the decay in the Green function, multiple pulses can never perform as well as a single pulse, even if optimally timed, unlike the homogeneous case.

In the nonlinear regime, or with a tapered channel designed to extend the de-phasing length, the solution for a channel is not analytically tractable, but one expects a balance between the decay of the plasma response, favoring depositing the energy impulsively in

[12] Such simulations have been performed by B.A. Shadwick, and discussed in personal communication.

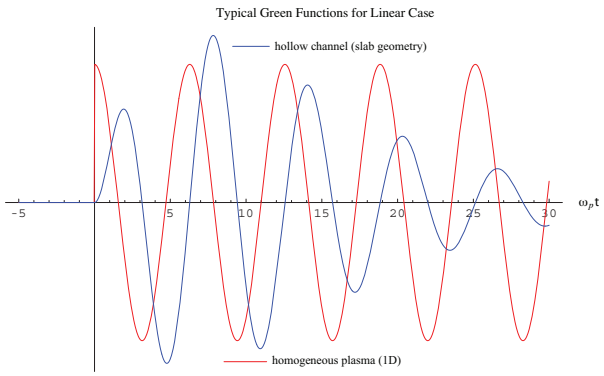

Figure 2.4. An example of the Green function for longitudinal Electric field response in a hollow plasma channel of transverse cross-section somewhat larger than a plasma wavelength, with the Green function for the homogenous plasma shown for comparison.

a short pulse, and the synergistic effects of the nonlinearity, favoring an arbitrarily large number of carefully timed pulses, leading perhaps to an optimal shape with some finite number of appropriately phased pulses. This is left as an open problem.

Details for the Linear Channel

Now we consider a somewhat more involved model, that of linearized plasma dynamics in a plasma channel of some prescribed density cross-section, which is assumed translationally invariant in the longitudinal direction, and in either slab or cylindrical geometry transversely. We imagine driving the plasma channel by a polarized laser pulse of prescribed polarization and transverse and longitudinal structure. The polarization and transverse envelope are assumed to be matched sufficiently well to the channel so that leakage or betatron-type oscillation of the laser profile can be neglected. We assume the characteristic channel cross section and matched spot size σ_a are sufficiently large compared to the optical wavelength λ_0 so that the propagation is essentially paraxial, and the longitudinal component of the laser field remains small. Longitudinally, we will also continue to ignore dispersion and depletion effects as well. The laser pulse will excite a plasma response which in general will not be an electrostatic mode, but can have a longitudinal electric field component suitable for acceleration, but which will decay behind the laser pulse due to leakage out of the channel. For a narrow-band laser around a fixed carrier frequency, we are thus in effect assuming that the laser envelope is separable in the moving coordinate ξ and the transverse position $\boldsymbol{r}$, so that the (possibly complex) eikonal amplitude of the laser may be written as

$$\boldsymbol{a}(\xi,\tau;\boldsymbol{r}) = \mu_a \hat{\boldsymbol{\epsilon}} f_\perp(\boldsymbol{r}) a(\xi) e^{i[-k_0 v_g \xi + (v_g k_0 - \omega_0)\tau]} + c.c., \tag{2.120}$$

where μ_a is a normalization constant, $\hat{\boldsymbol{\epsilon}}$ is the polarization vector (assumed to be either linear in slab geometry or circular in cylindrical geometry), $f_\perp(\boldsymbol{r}_\perp)$ is the relative transverse profile, taken to have unit modulus without loss of generality, and $a(\xi)$ is the longitudinal profile.

For a fixed plasma channel structure, and with a separable eikonal laser envelope, fixed polarization $\hat{\boldsymbol{\epsilon}}$, given carrier phase, and a fixed transverse profile $f_\perp(\boldsymbol{r}_\perp)$ assumed matched to the channel, we consider shaping only the longitudinal profile $a(\xi)$ to optimize some aspect of the wake generation process. Under such assumptions, the wake response in linear theory is essentially independent of τ in the co-moving frame, and can be written at any transverse position $\boldsymbol{r}$ in terms of a one-dimensional convolution:

$$E_z(\xi, \boldsymbol{r}) \propto \mathcal{E}(\xi; \boldsymbol{r}) = \int_{-\infty}^{\infty} d\xi'\, s(\xi') G(\xi - \xi'; \boldsymbol{r}), \tag{2.121}$$

where the driving source may be defined as $s(\xi) \equiv \frac{1}{2} |a(\xi)|^2$ similar to the homogeneous case.

The precise shape of the longitudinal Green function, or impulse response $G(\xi; \boldsymbol{r})$ depends implicitly on the details of the plasma channel, laser polarization, and transverse laser envelope; such responses have been determined numerically for hollow and parabolic channels in slab or cylindrical geometry a typical example was shown in Fig. 2.4. Generically, such Green functions will share some important general properties: because we assume that no perturbation exists ahead of the pulse, the Green function will be causal, vanishing for $\xi < 0$ but real and analytic for $\xi > 0$; because of leakage of energy out of the channel, the Green function at any fixed transverse location inside the channel will exhibit damping for sufficiently large $\xi > 0$, and in practice will look very much like a single exponentially-decaying sinusoidal oscillation, except for some small but noticeable differences largely confined to the very first oscillation, features which depend on details of the plasma channel, the transverse mode, and the transverse position. As stated above, the form of the Green function $G(\xi; \boldsymbol{r})$ at any given transverse position can typically be well approximated by a sum (2.119) of a small number of complex exponentials.

Maximization of the transformer ratio R_3 in this situation would require detailed knowledge of the plasma response $G(\xi; \boldsymbol{r})$ However, it is intuitively plausible that, because the plasma response is roughly similar to that in the one-dimensional linear electrostatic (homogenous) case, apart from some decay, we would expect the optimized pulse shape to

appear qualitatively similar, with an initial kick followed by a rising ramp (not linear, in general), which then sharply cuts off.

Maximization of the peak wake amplitude, subject to a bound on envelope energy, is also expected to lead to an impulsive laser envelope, as before, and in fact we will now explicitly prove this result, because it turns out that the consequences of pulse-shaping so as to maximize wake response in this model of a channel are largely independent of most specific details and precise form of the longitudinal Green functions. Because of the decay of the wake within the channel, the electrostatic energy will be finite, so it is more convenient mathematically to optimize some measure of this wake energy rather than the wake amplitude at some point or the peak wake amplitude behind the source. We have argued above that maximizing wake energy, rather than wake amplitude, without further constraints might easily lead to solutions with long wakes of low peak gradient unsuitable for LWFA applications, but because in this case the characteristic decay time of the plasma response behind the laser drive is in fact independent of the longitudinal duration of the laser, it turns out that one can establish that maximizing the total wake energy is equivalent to maximizing the amplitude of the wake immediately behind the source in the linear regime, for a broad class of decaying Green functions. The proof is straightforward but surprisingly tedious, and will be omitted.

We thus proceed to maximize the weighted-energy-functional

$$\mathscr{U}_w \equiv \mathscr{U}_w[a(\xi)] = \iint d^2\boldsymbol{r}\, w(\boldsymbol{r})\, u(\boldsymbol{r}), \tag{2.122}$$

where

$$u(\boldsymbol{r}) = \int_{-\infty}^{\infty} d\xi\, \mathcal{E}^2(\xi;\boldsymbol{r}) \tag{2.123}$$

is proportional to the energy density associated with a wake "filament" at a fixed transverse position $\boldsymbol{r}$. The function $w(\boldsymbol{r}) \geq 0$, assumed to be nonnegative, scales the transverse contributions of wake energy, to preferentially weight the contributions at different transverse positions, if desired. If $w(\boldsymbol{r})$ is constant, then H_w measures the total energy of the longitudinal electric field throughout the plasma, while if $u(\boldsymbol{r}) \propto \delta^2(\boldsymbol{r})$, then $\mathscr{U}_w$ measures the energy-density on axis only, which may be of more interest in accelerator applications; more general functional forms for $w(\boldsymbol{r})$ can interpolate between these extremes, preferentially but not exclusively weighting the contribution of the field on-axis, for example. As a natural constraint on feasible solutions, we first bound the laser envelope energy, which for a fixed transverse envelope $f(\boldsymbol{r})$ is just proportional to $\int_{-\infty}^{\infty} d\xi\, s(\xi)$. In this simple case, we will find

that that the *same* longitudinal laser profile separately maximizes each filament's contribution $u(\boldsymbol{r})$, so the optimal answer is actually independent of the specific weight function $w(\boldsymbol{r})$ used, as well as the transverse structure. We will later address and incorporate bandwidth limitations, i.e., constraints on the spectral content of the laser. Since we still assume a linear wake response, we can again conclude that the optimal driving pulse will contain all the energy allowed to it, so that an upper bound imposed on laser energy is never slack, and can be replaced with an equivalent equality constraint. Suppressing for the moment all $\boldsymbol{r}$-dependence in our notation for convenience, we first seek to maximize the filament-energy

$$u = \int_{-\infty}^{\infty} d\xi\, \mathcal{E}^2(\xi) = \int_{-\infty}^{\infty} d\xi \int_{-\infty}^{\infty} d\xi'\, G(\xi - \xi') s(\xi') \int_{-\infty}^{\infty} d\xi''\, G(\xi - \xi'') s(\xi''), \tag{2.124}$$

subject to the constraint that

$$\int_{-\infty}^{\infty} d\xi\, s(\xi) = U. \tag{2.125}$$

Rearranging the order of integration, we have

$$u = \int_{-\infty}^{\infty} d\xi' \int_{-\infty}^{\infty} d\xi''\, k(\xi', \xi'') s(\xi') s(\xi''), \tag{2.126}$$

where we have defined an energy-response kernel $k(\xi', \xi'')$ given by

$$k(\xi', \xi'') = \int_{-\infty}^{\infty} d\xi\, G(\xi - \xi') G(\xi - \xi'') = \int_{-\infty}^{\infty} d\xi\, G(\xi) G(\xi + [\xi' - \xi'']), \tag{2.127}$$

and where, to obtain the second expression immediately above, we have made a change of variables without re-naming, i.e., set $\xi \leftarrow \xi + \xi'$. We see that the kernel $k(\xi', \xi'')$ can be interpreted as the *autocorrelation function* of the translation-invariant Green function response $G(\xi)$, and must be real and symmetric, and also positive semi-definite when regarded as the kernel of an integral operator, but not necessarily nonnegative in its function values. Because of overall translation invariance, the kernel is a function of the relative coordinate $\Delta \equiv (\xi' - \xi'')$ only, and in fact is an even function of Δ because of the symmetry. Since $G(\xi)$ must decay as $\xi \to +\infty$, it cannot be periodic or even measure-periodic, i.e., if $\Delta \neq 0$ then there exist no complex constants μ, λ such that $\mu G(\xi) = \lambda G(\xi + \Delta)$ almost everywhere. It follows from the Schwarz-Holder inequality that $k(\Delta)$ achieves its unique global maximum at zero lag; i.e., $k(\Delta) < k(0)$ if $\Delta \neq 0$.

Making use of the causality in $G(\xi)$ and the even parity of $k(\Delta)$, we can also write $k(\Delta) = k(|\Delta|) = \int_0^{\infty} d\xi\, G(\xi) G(\xi + |\Delta|)$. With another change of variables, we see that the

filament energy u can be written as

$$u = \int_{-\infty}^{\infty} d\Delta\, k(\Delta) C(\Delta), \tag{2.128}$$

where $C(\Delta)$ is the autocorrelation of the source function $s(\xi)$:

$$C(\Delta) = \int_{-\infty}^{\infty} d\xi\, s(\xi) s(\xi + \Delta), \tag{2.129}$$

which must be real, even, and nonnegative. Because the energy constraint implies that the laser envelope $s(\xi)$ must have a finite but non-zero measure, it too cannot be measure-periodic, so its autocorrelation also satisfies $C(\Delta) < C(0)$ whenever $\Delta \neq 0$. Exploiting parity, we need only integrate over positive lags if convenient, i.e., $u = 2\int_0^\infty d\Delta\, k(\Delta) C(\Delta)$.

Using Parseval's theorem and the Convolution theorem, we can also conveniently express this energy response in "frequency" space, i.e., in Fourier space conjugate to ξ (not t):

$$u = \int_{-\infty}^{\infty} d\omega\, K(\omega) \left|S(\omega)\right|^2 = 2\int_{0}^{\infty} d\omega\, K(\omega) \left|S((\omega)\right|^2, \tag{2.130}$$

where $K(\omega)$ is the Fourier transform of $k(\Delta)$, and must be real and even because $k(\Delta)$ is, and $S(\omega)$ is the Fourier transform of $s(\xi)$, so that its modulus $|S(\omega)|$ is real, nonnegative, and also even because $s(\Delta)$ is real. Because $s(\xi)$ is also nonnegative, $S(\omega)$ must necessarily satisfy certain other conditions; in fact, from Bochner's theorem we may conclude that $S(\omega)$ must itself be interpretable as the autocorrelation function of some other function. In this specific case we know the latter is just proportional to the Fourier transform $A(\omega)$ of $a(\xi)$, and in particular this additional structure implies that $|S(\omega)| \leq |S(0)|$.

The constrained optimization of any particular filament energy u proceeds much like the previous case of optimizing the wake amplitude in a homogeneous plasma. The ξ-space variational Euler-Lagrange equation, a necessary condition for the a local optimum, can be written as:

$$s(\xi) \left\{ \int_{-\infty}^{\infty} d\xi' \, \left[k(\xi - \xi') s(\xi')\right] - \lambda \right\} = 0, \tag{2.131}$$

where $s(\xi)$ must be nonnegative, and λ is a constant Lagrange multiplier whose value is chosen so that both this stationarity condition and the energy constraint are satisfied, and can be written as:

$$\lambda = \frac{\int_{-\infty}^{\infty} d\xi \int_{-\infty}^{\infty} d\xi'\, k(\xi - \xi') s(\xi) s(\xi')}{\int_{-\infty}^{\infty} d\xi\, s(\xi)} = \frac{u}{U}. \tag{2.132}$$

The stationarity condition implies that almost everywhere, either

$$s(\xi) = 0 \tag{2.133}$$

or

$$\int_{-\infty}^{\infty} d\xi' \, k(\xi - \xi') s(\xi') = \lambda. \tag{2.134}$$

Using the parity of the kernel and a change of variables we can write the latter requirement as:

$$\int_{-\infty}^{\infty} d\xi' k(\xi' - \xi) s(\xi') = \int_{-\infty}^{\infty} d\xi' k(\xi') s(\xi' + \xi) = \lambda. \tag{2.135}$$

Now suppose this condition holds identically for any source whatsoever over some interval (possibly infinite) in ξ. Then by overall translation invariance, it will also hold for some source $s(\xi)$ within an interval $(-\xi_0\,,\,\xi_0)$ containing the origin, where $0 < \xi_0 \leq \infty$. The condition (2.135) can be expressed equivalently in Fourier space as

$$\int_{-\infty}^{\infty} d\omega \, K(\omega) S(\omega) e^{i\omega\xi} = \lambda \quad \forall\, \xi \in (-\xi_0\,,\,\xi_0). \tag{2.136}$$

Differentiating both sides with respect to ξ and evaluating at $\xi = 0$, this requires

$$\int_{-\infty}^{\infty} d\omega \, \omega^n \, [\omega K(\omega) S(\omega)] = 0 \quad \text{for } n = 0, 1, 2, \ldots, \tag{2.137}$$

which together with linearity further implies

$$\int_{-\infty}^{\infty} d\omega \, h(\omega) \, [\omega K(\omega) S(\omega)] = 0 \tag{2.138}$$

for arbitrary analytic functions $h(\omega)$, that can be written as a power series in ω, i.e., for all convergent sums $h(\omega) = \sum_{j=0}^{\infty} \alpha_j \, \omega^j$ for appropriate complex coefficients α_j, $j = 0, 1, 2, \ldots$.

Since we may use analytic functions to approximate any bump function with arbitrary accuracy, from this we can infer that in fact $[\omega K(\omega) S(\omega)] = 0$ almost everywhere, which in turn implies that $K(\omega)S(\omega) \propto \delta(\omega)$, and substituting back into the integral relation we find $K(\omega)S(\omega) = \lambda\delta(\omega)$. From the asymptotically-exponential decay in the impulse response $G(\xi)$, we know that $K(\omega)$ will be finite everywhere, including at $\omega = 0$, so it must be $S(\omega)$ that instead contains some singular generalized function(s) at $\omega = 0$, i.e., $S(\omega) = \sum_{j=0}^{\infty} \beta_j \, \delta^{(2j)}(\omega) + \ldots$ for some constant complex expansion coefficients β_j. The ellipses indicate that $S(\omega)$ might also include certain nonsingular (bounded) terms, and we have omitted any odd derivatives of delta-functions as they would lead to odd contributions in

$s(\xi)$, violating the nonnegativity requirement. But this distributional form for $S(\omega)$ implies that $S(\omega = 0) = \infty$, while the energy constraint requires that $S(\omega = 0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} d\xi\, s(\xi) = \frac{1}{\sqrt{2\pi}} U < \infty$, leading to a contradiction.

So in fact the second branch of the stationarity condition (2.131) cannot be satisfied together with the nonnegativity and energy constraints over any non-zero interval (at least in the nontrivial case where $\lambda > 0$), and we can conclude that in fact $s(\xi) = 0$ almost everywhere, i.e., all the energy in the source $s(\xi)$ must be concentrated in impulsive terms, so that the source can be written as $s(\xi) = \sum_j c_j\, \delta(\xi - t_j)$, where the c_j are some number of positive constants, chosen such that the energy constraint $\sum_j c_j = U$ is satisfied. Substituting into the above stationarity condition we find:

$$\sum_m c_m \Big[\sum_n c_n k(t_m - t_n) - \lambda \Big] \delta(\xi - t_m) = 0, \tag{2.139}$$

which can only be satisfied if

$$\sum_n c_n\, k(t_m - t_n) = \lambda \quad \forall\, m, \tag{2.140}$$

where $\lambda = \frac{u}{U} = \frac{1}{U} \sum_m \sum_n c_m c_n\, k(t_m - t_n)$.

The most obvious solution is just a single delta-function, i.e., $s(\xi) = U\delta(\xi - t_1)$, for which $u = U^2\, k(0)$. For generic $k(\Delta)$, driving sources involving two or more impulses can also be found. However, the latter are only local optima, and single-impulse solution is in fact the *unique global* maximum. Since the unique absolute maximum of the kernel $k(\Delta)$ occurs at $\Delta = 0$, a source with two or more impulses, such that $t_m \neq t_n$ for some m, n, will have:

$$u = U\,\lambda = \sum_m \sum_n c_m c_n k(t_m - t_n) \;<\; \sum_m \sum_n c_m c_n\, k(0) = \Big(\sum_m c_m \Big)^2 k(0) = U\, k(0). \tag{2.141}$$

Because this impulsive source independently maximizes each filament energy contribution u, it will also maximize *any* transversely-weighted wake energy $\mathscr{U}_w$. Thus, as in the homogeneous plasma case, a single delta-function source leads to the largest wake response for fixed source energy. After trudging through all this mathematics, this can also be seen from the filament energy expression (2.128), which can be thought of as a kind of weighted average of the energy response kernel $k(\Delta)$ with the weight function given by the source autocorrelation $C(\delta)$, apart from overall normalization by $\int\int d\Delta d\xi\, s(\xi) s(\xi - \Delta) = \big(\int d\xi'\, s(\xi') \big)^2 = U^2$, which is fixed at a constant value by the energy constraint. Since $k(\Delta)$ peaks at $\Delta = 0$, the integral obviously can get no larger than when $C(\Delta)$ is concentrated at $\Delta = 0$, i.e., when

$C(\Delta) \propto \delta(\Delta)$. This can in principle seemingly occur in any number of ways: if, for example, $s(\xi)$ is a white noise, which, however, cannot simultaneously satisfy the nonnegativity and fixed energy constraints, or if $s(\xi)$ is itself a delta-function, which is the solution found above.

In fact, the more careful analysis above demonstrated that a delta-function is the unique (up to overall translations) globally optimal shape which is of specified energy and is everywhere nonnegative. The situation is perhaps even more illuminating in Fourier space, where the energy response is given by (2.130). The kernel $K(\omega)$ must be finite, real, and even, and generically $K(|\omega|)$ will look something like a Lorentzian, peaking near but not exactly at the natural plasma frequency $\omega \approx 1$ (because of the effective channel damping, the exact resonance occurs at complex frequency), and then falling off algebraically (See Fig. 2.5.) Ideally, it would seem that we should concentrate $|S(\omega)|$ at the maximum of $K(|\omega|)$ in order to optimize the wake response, and indeed this would be the case, but for the constraints. The finite-energy constraint implies that $s(\xi)$ cannot be sinusoidal, so we cannot concentrate $|S(\omega)|$ entirely as impulses at the (positive and negative-frequency) maxima of $K(|\omega|)$, but at most only in some finite peaks centered near the maxima. However, such a solution cannot satisfy the positivity constraint, which implies that the absolute maximum of $|S(\omega)|$ must occur at $\omega = 0$, while $|S(\omega = 0)| = S(0) > 0$ is proportional to the envelope energy U.

In real space, $s(\xi)$ must be everywhere nonnegative, implying the average value of $s(\xi)$ must be finite and positive and just large enough to ensure that $s(\xi)$ never goes negative, yet does vanish sufficiently rapidly as $|\xi| \to \infty$. Therefore the value of $|S(\omega)|$ near the peaks in $K(\omega)$ can be no larger than $|S(\omega = 0)| = S(0)$, which is fixed by the energy constraint, and hence we can evidently do no better than demanding $|S(\omega)|$ is in fact constant, with phases chosen to reproduce a real, positive $s(\xi)$. Any choice of phases such that $S(-\omega) = S(\omega)^*$ will produce a real source $s(\xi)$, but it turns out that only linear phases, i.e., for which $\arg[S(\omega)] = \omega t_0 + \theta_0$ for some real constants t_0 and θ_0 will also yield a source which is everywhere nonnegative, again leading to the delta-function solution. This is indicated schematically in Fig. 2.5

This spectral representation reveals two somewhat surprising features of the solution. First, as we have already noted, the energy and positivity constraints are in effect *maximally* operative, since the unconstrained optimum would evidently just be a single sinusoid with zero bandwidth and infinite duration, while the constrained solution is a delta-function with infinite bandwidth and zero duration. Thus the positivity of the source, itself a consequence of the ponderomotive nature of the laser-driven forcing, is highly relevant to issues of wake

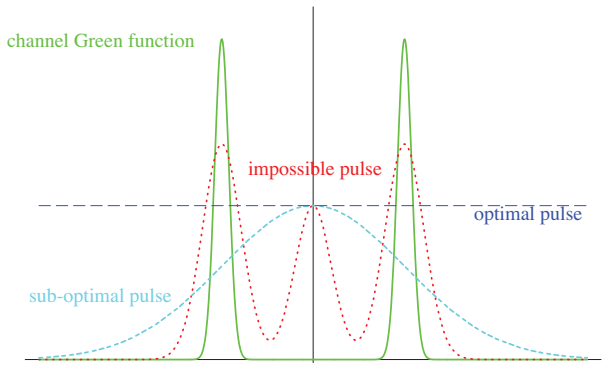

Figure 2.5. Schematic of the source spectra and plasma response. Plotted are the absolute value of the channel Green function along the real ω axis, along with power spectra of: a proposed but inconsistent pulse which does not satisfy the nonnegativity constraint, a possible but sub-optimal pulse, and the optimal pulse corresponding to a delta function.

generation in the LWFA. Second, we see that because of the constraints, most of the energy in the source is far from the resonance (or more precisely, the quasi-resonance) in the wake response. Thus the energy delivered to the wake continues to increase towards its maximum as the effective width of the driving pulse shrinks, even after it narrows beyond the "resonant" width of $O\left(\frac{1}{2}\mathcal{T}_p\right) = O\left(\frac{\pi}{\omega_p}\right)$ widely though to be optimal. As discussed above, this contrasts with the situation in which peak amplitude, rather than envelope energy, is held fixed as the pulse width is narrowed, for which the energy in the drive must decrease toward zero, leading to a peak in wake response at some intermediate width with balances the effects of diminishing absolute driver energy and increasing relative spectral content near resonance. Of course, as the pulse is narrowed past the scale of the plasma period, the wake will response will continue to increase (disregarding differential dispersive effects), but with diminishing marginal returns as the plasma is increasingly unable to resolve features that become much shorter in duration that the plasma period.

2.7.4 Bandwidth Constraints

Very short impulse-like drives, or square pulses with abrupt cutoffs or other sharp features are *not* very realistic, because of the excessive bandwidth required. Here we take some preliminary steps into incorporating more realistic spectral limitations on the actual shaping-process and resulting pulse-shape.

Consider a linear wake response in a homogeneous plasma or channel, excited by some driving pulse $a(\xi)$ produced in an idealized CPA system, with Fourier transform (in ξ):

$$a(\omega) = e^{L_g|G(\omega)|}\,|T(\omega;L_g)|\,e^{i\theta(\omega;L_g)}D(\omega), \tag{2.142}$$

where $L_g \geq 0$ is an effective gain length and/or strength, $|G(\omega)|$ is the *relative* gain curve, $\theta(\omega; L_g)$ represents both unavoidable dispersion or phase shifts in the amplifier and any phase-masking intentionally introduced, and $|T(\omega; L_g)|$ represents losses in the amplifier as well as the action of a possible amplitude mask, while $D(\omega)$ is the spectrum of the seed pulse prior to amplification, typically reasonably well described by a Gaussian or Lorentzian of bandwidth $\Delta\omega_{\mathrm{D}}$ exceeding the characteristic gain bandwidth $\Delta\omega_G$ associated with $|G(\omega)|$.

Experimentally, one can image in principle controlling L_g through the adjustments to the amplifier geometry and extent of pumping, and parts of $|T(\omega; L_g)|$ and $\theta(\omega; L_g)$ through filters, masks, gratings, electro-optic modulators, or other linear or nonlinear optical elements, while having relatively little control over the remaining terms for a given gain medium and seed laser system.

In the linear regime, it turns out that wake response will increase monotonically with increasing L_g, and optimal masking occurs for:

$$|T(\omega; L_g)| = 1, \tag{2.143}$$

and

$$\theta(\omega; L_g) = -\arg[D(\omega)] + \omega\xi_0 + \theta_0; \tag{2.144}$$

for some constants $\xi_0 \in \mathbb{R}$ and $\theta_0 \in \mathbb{R}$ associated with an overall translation in time and some arbitrary global phase shift. That is, the optimal shape is not really shaped at all! As nearly as possible, the drive pulse is made to have linear phases and as a result look as look as narrow as possible in ξ, given its initial power spectrum and the response of the amplifier.

Since power cannot be easily moved from one Fourier component to another, but only added by the amplifier or absorbed by a mask, in the linear regime it is apparently never efficient to remove energy from the driver pulse in any frequency interval once it has been "paid for," provided those Fourier components can be properly phased to minimize the pulse duration.

Whether this remains true in the nonlinear regime is unknown. In the self-modulated case (SM-LWFA), it has been found both theoretically and experimentally[23, 24, 25] that moderate shaping that actually tends to increase the overall pulse length can improve performance. By tweaking the CPA optics before the final compression, the pulse can be skewed such that the leading edge of the pulse is steepened and the remainder is stretched, resulting in measurable enhancement of electron acceleration. The idea is that a sharper

leading edge is more effective at initiating the instabilities that modulate the remainder of the pulse envelope, which then in turn resonantly excite the wake.

However, this regime is far different from the standard LWFA, which naturally operates at shorter intrinsic pulse lengths (for given plasma density) and for which Raman and other instabilities tend to be deleterious. In the case of short-pulse LWFA operating in the strongly nonlinear regime, we conjecture, but have not proven, that optimal masking will still not remove energy already supplied by the amplifier from any frequency band, but will arrange the phases so as to decrease the pulse-length until some compromise with dispersion is achieved.

Details of the Bandwidth-Limited Linear Case

Given that the wake response, in the absence of dispersive or depletion effects, is maximized for fixed laser energy by a driving pulse which is essentially as short and peaked as possible, we ought to take account of limitations on the spectral content of the driving laser envelope, which have so far been neglected, leading to unphysically narrow delta-function pulses as optimal shapes for wake excitation. However, just imposing an absolute bandwidth cutoff (i.e., rectangular spectral window) on the spectrum of the laser pulse along with an upper bound on energy is neither very convenient mathematically nor very realistic physically.

Here we continue to treat only the linear wake response in either a homogeneous plasma or channel, excited by some driving pulse $a(\xi)$ produced in an idealized CPA system. Instead of assuming an intense drive pulse of given energy and duration, let us consider an idealized laser amplification process, in which a large-bandwidth, low-energy seed oscillation is linearly amplified, possibly in multiple stages, and perhaps optically shaped via filters or phase or amplitude masks or other optical elements, in order to produce the post-amplified source pulse actually used to ponderomotively excite the wake.

We can express the net overall pulse amplification and shaping process in Fourier space as

$$A(\omega) = e^{L_g|G(\omega)|} e^{i\theta_0(\omega;L_g)} e^{i\theta_1(\omega;L_g)} \, |T_0(\omega;L_g)| \, |T_1(\omega;L_g)| \, D(\omega). \tag{2.145}$$

Here for convenience $A(\omega)$ is taken to be the Fourier transform of the driving (post-amplifier) laser pulse envelope $\tilde{a}(\xi)$, i.e., with the fast carrier dependence removed, so ω can be regarded as the shifted and scaled co-moving wavenumber conjugate to ξ, but it will usually just be referred to as a frequency, because the amplifier itself is approximately time-invariant

(the time to make one pass is typically large compared to the carrier frequency), but not spatially translation-invariant, so inside the amplifier it is the frequencies that provide the naturally invariant labels for the spectral components. All other Fourier transforms or transfer functions discussed shortly are to be similarly shifted, so as to be centered near $\omega \approx 0$, rather than near the scaled carrier frequency $\frac{\omega_0}{\omega_p}$.

$D(\omega)$ is the spectrum of the initial, or "raw" laser seed oscillation envelope, assumed to have a small but non-zero amplitude and a characteristic bandwidth $\Delta\omega_D$ in scaled units. Often the seed envelope spectrum $D(\omega)$ can be approximated by a Gaussian with standard deviation $\sigma_\omega \sim \Delta\omega_D$, which then can be written as $D(\omega) \propto e^{i\omega\xi_0} e^{-\frac{\omega^2}{2\Delta\omega_D^2}}$ for some constant ξ_0 determining the overall timing of the initial pulse, or else by a Lorentzian with FWHM $\sim 2\Delta\omega_D$.

$L_g > 0$ is an effective gain length and/or strength over the entire amplification process, which is determined by details of the amplifier design, such as the structure and geometry of the gain medium (such as as titanium-sapphire crystal) and auxiliary optics, etc, as well as the overall amount of external pumping. The unimodular function $e^{i\theta_0(\omega;L_g)}$ for some $\theta_0(\omega; L_g) \in \mathbb{R}$ represents dispersion or other unavoidable phase shifts in the amplifier optics, while the corresponding term involving $\theta_1(\omega; L_g) \in \mathbb{R}$ represents the effects of any additional phase-masking *intentionally* introduced to shape the laser envelope at any stage. The function $|T_0(\omega; L_g)|$ satisfies $0 \leq T(\omega) \leq 1$, and represents the intrinsic net transmission coefficient, including any unavoidable sources of absorption or loss in the amplification process or elsewhere in the optics. The $|T_1(\omega; L_g)|$ term also satisfies $0 \leq T(\omega) \leq 1$, but is intended to represent the possible effects of some additional (passive) amplitude-masking deliberately introduced to shape the laser before, during, or after the actual amplification.

With absorption included in $T_0(\omega; L_g)$ and phase shifts accounted for in $\theta_0(\omega; L_g)$, the remaining function $e^{L_g|G(\omega)|}$ represents pure gain, where $|G(\omega)| \geq 0$ is a normalized (and shifted) gain curve representing relative amplification for frequencies around the laser carrier frequency, and can be taken to independent of L_g for a linear amplifier. Presumably, $|G(\omega)|$ peaks at some large relative gain value somewhere near $\omega \approx 0$, and then falls off toward zero as $|\omega| \to \infty$ with some characteristic gain-bandwidth scale $\Delta\omega_G$ which is typically smaller than the initial seed bandwidth $\Delta\omega_D$.

With this flexible model, we take a step or two closer to more realistic limitations, yet are still able to largely ignore many of the complicated details of the shaping and amplification, typically involving intricate and high precision optics, several stages of compression

and dilation with gratings , amplification in Ti:sapphire, Neodymium-Yttrium-Aluminum-Garnet (Nd-YAG) or other crystals, possibly phase or amplitude masking using filters, holographic gratings, nonlinear crystals or electro-optic modulators, etc., and just consider the net process in terms of an overall linear transfer function.

As natural constraints on the longitudinal pulse-shape optimization, one should fix the initial (i.e., unamplified) seed envelope energy, approximately proportional to

$$\tfrac{1}{2}\int d\omega \; |D(\omega)|^2 = U_0 \tag{2.146}$$

to be some finite, and in fact realistically small, value, and specify its initial bandwidth $\Delta\omega_D$ to be finite but realistically large; otherwise an unrealistically narrow or intense near-impulsive post-amplifier pulse could be produced despite our taking the trouble to incorporate the finite gain and bandwidth of the amplifier.

One must of course ensure that only a finite amount of gain can be achieved. One could bound the total amount of final (i.e., post-amplified) envelope energy, proportional to $U = \frac{1}{2}\int d\omega \; |A(\omega)|^2$, but because absorption might occur during intermediate stages of amplification, the difference between this final pulse energy U and the initial energy U_0 does not necessarily represent the total energy that is delivered to the pulse by the amplifier during the amplification and shaping process. Unfortunately, the latter quantity depends on the precise details of the amplification and the actual time-history of the pulse, i.e., exactly when and where in the amplification process the absorption occurs. Physically, however, the most important constraint seems not to be the final pulse energy *per se*, nor the net energy increase, nor even the total time-integrated energy expended in the amplification process, but rather the extent of amplification possible in the particular apparatus used, which is embodied in the gain curve and effective gain length. Therefore, it seems most natural not to fix U directly, but rather assume $|G(\omega)|$, $\theta_0(\omega; L_g)$, $T_0(\omega; L_g)$, L_g, and $D(\omega)$ are specified, as they are fixed by the actual experimental considerations and all the attendant technological limitations, and then determine the optimal choice of the net absorption and phase-mask functions $|T_1(\omega; L_g)|$ and $\theta_1(\omega; L_g)$, and then finally examine the dependence of the optimized pulse on the specified amplifier and seed parameters.

The total beam energy will then be limited not by an explicit constraint but by the initial driver energy and by the maximum gain achievable. From the spectral representation of the wake response, it is evident that the energy (and along with it, the peak amplitude) in the induced wake will increase if $|S(\omega)|$ can be made larger anywhere in spectral space, where $S(\omega)$ is the Fourier transform of then envelope $s(\xi) = \frac{1}{2}\tilde{a}(\xi)^2$. If any trade-off must be faced

in maximizing different frequency components of $|S(\omega)|$ for fixed driver energy and amplifier gain, then obviously it would be preferable to make $|S(\omega)|$ larger near the resonance, where the wake response is largest. However, as we have seen, the nonnegativity constraint on $s(\xi)$ greatly limits the extent to which $|S(\omega)|$ can be increased for any $\omega \neq 0$ without also increasing $|S(\omega = 0)| \propto U$, which is indirectly limited by the seed energy introduced and total amplification that can be achieved.

In fact, we can see that $|S(\omega)|$ can be made as large as possible *everywhere*, without having to worry about any such compromises affecting *where* it should be made larger. Because $s(\xi) = \frac{1}{2}|a(\xi)|^2$, the spectrum $S(\omega)$ can be written in terms of the autocorrelation of the Fourier transform $A(\omega)$ of $a(\xi)$:

$$S(\omega) = \tfrac{1}{2\sqrt{2\pi}} \int d\omega' \, A(\omega') A^*(\omega' - \omega), \tag{2.147}$$

where specifically we take $A(\omega) = \frac{1}{\sqrt{2\pi}} \int d\omega \, \tilde{a}(\xi) e^{i\omega\xi}$. Using the Cauchy-Schwarz inequality, we can infer that

$$|S(\omega)| \leq \tfrac{1}{2\sqrt{2\pi}} \int d\omega' \, \left| A(\omega') A^*(\omega' - \omega) \right|, \tag{2.148}$$

with equality if and only if $A(\omega)$ satisfies

$$A(\omega') A^*(\omega - \omega') = \left| A(\omega') A^*(\omega' - \omega) \right| e^{i\theta_2(\omega)} \tag{2.149}$$

for all ω and ω' and some real function $\theta_2(\omega)$. This is in turn equivalent to

$$A(\omega) A^*(\omega' - \omega) = \left| A(\omega') \right| \, \left| A(\omega' - \omega) \right| e^{i\theta_2(\omega)}, \tag{2.150}$$

or

$$\arg[A(\omega')] - \arg[A(\omega' - \omega)] = \theta_2(\omega), \tag{2.151}$$

for which the only class of solutions valid for arbitrary ω and ω' corresponds to linear phases for $A(\omega)$, i.e.,

$$\arg[A(\omega)] = \omega \xi_0 + \psi_0 \tag{2.152}$$

for some arbitrary real constants ξ_0 and ψ_0.

Defining $\theta(\omega; L_g) = \theta_0(\omega; L_g) + \theta_1(\omega; L_g)$ and $T(\omega; L_g) = T_0(\omega; L_g) T_1(\omega; L_g)$, the transfer function form (2.145) can be written somewhat more compactly as

$$A(\omega) = e^{L_g |G(\omega)|} e^{i\theta(\omega; L_g)} \, |T(\omega; L_g)| \, D(\omega). \tag{2.153}$$

Substituting the form for $A(\omega)$ into the inequality (2.148) we have

$$\begin{aligned}|S(\omega)| &\leq \tfrac{1}{2\sqrt{2\pi}} \int d\omega' \left|A(\omega')\right| \left|A(\omega'-\omega)\right| \\ &\leq \tfrac{1}{2\sqrt{2\pi}} \int d\omega' e^{(|G(\omega')|+|G(\omega'-\omega)|)L_g} \left|T(\omega';L_g)T(\omega'-\omega;L_g)\right| \left|D(\omega')\right| \left|D(\omega'-\omega)\right|.\end{aligned} \tag{2.154}$$

Noting that $|T(\omega';L_g)T(\omega'-\omega;L_g)| = |T(\omega';L_g)|\,|T(\omega'-\omega;L_g)| \leq 1$, with equality everywhere if and only if $|T(\omega;L_g)| = 1$ for almost all ω, we infer

$$\begin{aligned}|S(\omega)| &\leq \tfrac{1}{2\sqrt{2\pi}} \int d\omega' \left|A(\omega')\right| \left|A(\omega'-\omega)\right| \\ &\leq \tfrac{1}{2\sqrt{2\pi}} \int d\omega' e^{(|G(\omega')|+|G(\omega'-\omega)|)L_g} \left|D(\omega')\right| \left|D(\omega'-\omega)\right|.\end{aligned} \tag{2.155}$$

where this upper bound can be achieved if and only if $T(\omega) = 1$ everywhere $D(\omega) \neq 0$, and $\theta(\omega) = -\arg[D(\omega))] + \omega\xi_0 + \psi_0$ for some constant ω_0 and ξ_0. In the linear regime (with no group velocity dispersion in the subsequent propagation), it is never effective to impose any amplitude masking. Because $|G(\omega)| \geq 0$, and $L_g \geq 0$, $\exp\left[\left(|G(\omega')| + |G(\omega'-\omega)|\right)L_g\right] \geq 1$, so clearly $|S(\omega)|$ increases monotonically with L_g as well, implying that the benefits of greater source energy seem to outweigh any potentially deleterious effects of gain broadening.

If the initial driving spectrum $D(\omega)$ has linearly-varying phases, then $|S(\omega)|$ is made as large as possible at all ω simply by amplifying $D(\omega)$ as much as possible, without any additional adjustments to the phasing, and with any absorption made as small as possible at all frequencies. Otherwise, if $D(\omega)$ has some non-trivial (i.e., nonlinear) phase dependence, then $\phi(\omega)$ should be chosen if possible to cancel the nonlinear part of the phase advance, while $|T(\omega)|$ should still be kept as large as possible at all frequencies. Using calculus of variations, it can be shown that this linear re-phasing is equivalent to minimizing the RMS duration of the pulse envelope (with weighting proportional to $a(\xi)^2$), without changing the power spectrum.

In short, under these conditions, where we constrain the amplification process rather than the amplified driver energy, the optimal pulse shape is not really "shaped" at all, but made as spectrally broad, or temporally narrow, as possible, with as much power as possible at all frequencies, consistent with the implied constraints.

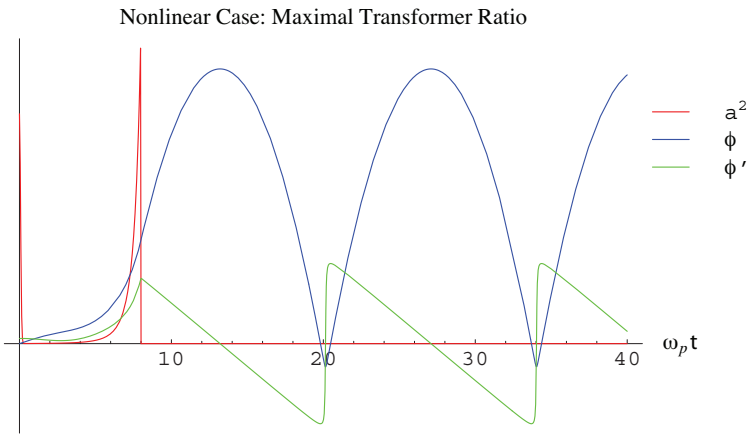

Figure 2.6. A laser envelope consisting of an impulse followed by a nonlinear ramp that is purported to maximize the transformer ration R_3 within the nonlinear quasi-static regime, along with the resulting (normalized) potential and electric field.

2.7.5 Nonlinear Regime: Maximizing the Transformer Ratio

Now we turn to pulse-shaping within the 1D nonlinear QSA dynamical description, with diffraction, depletion, and dispersion still ignored so that the laser envelope continues to be regarded as a prescribed function of ξ.

In this regime, Spitkovsky, Chen, et al. claim[8, 9, 10, 11] that the optimal family of pulse shapes maximizing what we have called the transformer ratio R_3 may be expressed as as an impulse followed by a nonlinear (concave) ramp:

$$a^2(\xi) = 2\kappa\delta(\xi) + \Theta_0(\xi)\,\Theta_1(\xi_f - \xi)\left[\frac{4\kappa^2}{(1-\kappa\xi)^5} + \frac{1}{(1-\kappa\xi)^2} - 1\right], \tag{2.156}$$

parameterized by the nonnegative constants $\kappa \geq 0$ and $\xi_f \geq 0$, where for simplicity the pulse has been assumed to begin at $\xi_0 = 0$, and in order that the envelope energy actually remain finite, the turn-off time ξ_f must satisfy $0 \leq \xi_f < \frac{1}{\kappa}$. An example of such a pulse shape and the resulting wake is shown in Fig. 2.6.

Unlike the linear case, this pulse shape must necessarily be of limited duration for fixed κ, but can still accommodate any finite envelope energy by abruptly shutting off ever closer to the pole at $\xi = \kappa^{-1}$. If instead both $\kappa \ll 1$ and $\kappa\xi_f \ll 1$, the analytic part of the source profile may be Taylor expanded and the result resembles that obtained in the linear case.

Because of the presence of the impulse and the sharp trailing edge, the solution is not consistent with either finite intensity constraints or finite bandwidth constraints. In practice the sharp features (impulse and cusp) will be smoothed in optical systems supporting only a finite bandwidth. Spitkovsky and Chen have performed a preliminary analysis of how

such smoothing might be expected to decrease the peak wake field and decrease the average value while increasing the variation in $\kappa(\xi)$.

As in the linear case, no absolute maximum of R_3 exists, and one must impose suitable additional constraints, for example on the value of the total envelope energy and some length-scale (for example, either on the gross pulse duration or the predicted depletion length), or on both the depletion and pulse lengths, before the optimization is even well-posed. Actually, we will see that for either of these choices, one additional consistency condition (an inequality) must be satisfied by the constrained parameters to ensure the existence of a solution.

Recall that our improved estimate for the predicted depletion length $L_{\rm pd}$ (2.45) is inversely proportional to the maximum rate of photon red-shift, which itself is proportional to the term $\max\big[\kappa(\xi)\big]$ constituting the denominator of R_3. If we can somehow neglect variations in $\sqrt{\tilde{\gamma}_\perp}$ over the range of possible pulse-shapes considered, then under constraints of fixed depletion length $L_{\rm pd}$ and either fixed envelope energy U or fixed pulse duration ξ_f, maximization of the ratio R_3 in the form (2.70) is therefore equivalent to maximizing the peak accelerating wakefield under the same constraints.

The proposed solution (2.156) to this constrained maximization problem then leads to a value

$$R_3 = \sqrt{\tilde{\gamma}_\perp}\frac{2\omega_0}{\omega_p}\frac{\sqrt{1+\xi_f^2\left(1-\kappa\xi_f\right)^3}}{\left(1-\kappa\xi_f\right)^2} \tag{2.157}$$

for the transformer ratio, and peak scaled wakefield

$$\max\big[|\mathcal{E}_+|\big] = \tfrac{\omega_p}{2\sqrt{\tilde{\gamma}_\perp}\omega_0}\sqrt{\tilde{\gamma}_\perp}\tfrac{2\omega_0}{\omega_p}R_3, \tag{2.158}$$

with the photon deceleration function

$$\kappa(\xi) = \Theta_{\frac{1}{2}}(\xi)\Theta_1(\xi_f-\xi)\kappa \tag{2.159}$$

rising instantaneously in the impulsive precursor but then remaining constant everywhere within the main part of the pulse.

The depletion length $L_{\rm pd}$ and constant photon deceleration factor κ are related by (2.43), while κ, the pulse duration, and the envelope energy U, and are related by

$$U = \kappa + \tfrac{1}{2}\xi_f\frac{\kappa^2+(1-\kappa\xi_f)^3}{(1-\kappa\xi_f)^4}, \tag{2.160}$$

where the two terms represent the envelope energy in the delta-function precursor and nonlinear ramp, respectively.

The claimed advantages of such pulses are traced back to the prediction of constant photon deceleration throughout the bulk of the pulse. Unshaped pulses deplete preferentially in regions where the photon deceleration rate $\kappa(\xi)$ is largest, whereas in (2.156) all the photons in the ramp decelerate uniformly and lose all of their energy in a single depletion length as defined, enhancing energy deposition in the wake.

We have already seen that the pulse duration must be shorter than κ^{-1} for given κ and any finite U. Since both terms on the RHS of (2.160) are nonnegative, for fixed energy U, the value of κ can reach at most U, which is achieved if and only if $\xi_f \to 0$, and the pulse then consists of a single delta function without any subsequent ramp.

The delta function shape therefore excites the largest possible wake amplitude but has the smallest depletion length (largest κ) and smallest value of R_3 amongst all pulses of a fixed envelope energy within the general family (2.156). Spitkovsky and Chen claim this distinction for the delta function in comparison to all pulse-shapes of fixed envelope energy, but this is *not* true in general, but only amongst contiguous pulses or sufficiently short competitors. As in the linear case, we see that for fixed energy, the peak wakefield and transformer ratio are not simultaneously maximized, but rather are inversely related.

This impulsive-precursor-plus-ramp envelope does *not* lead to the largest possible wakefield subject only to a constraint on the envelope energy, or even in general when upper bounds are placed on both energy and pulse duration, except when the latter is less than $\mathcal{T}_{\text{NL}}$. Actually, in general the parameterized family of solutions (2.156) is *not* even quite optimal as claimed when R_3 is accepted as the figure-of-merit for generic prescribed values of envelope energy U and duration ξ_f.

Each of these negative results arises for a similar reason, namely failure to exploit what we have called the synergy inherent in the nonlinear wake excitation process. For example, If $\xi_f > \mathcal{T}_{\text{NL}}$, then by using two or more impulses in the precursor it becomes possible to excite an initial wake with scaled photon deceleration rate at the same value κ as the proposed solution but with less overall precursor energy than the single impulsive precursor, leaving more energy available to deposit in the bulk of the ramp pulse, which will increase the peak value of $\phi'(\xi)$ achievable behind the pulse while maintaining the same constant rate of deceleration for drive photons in the ramp, and thus increasing the value of R_3 achieved. Such solutions may be deduced using the Pontryagin formalism.

To assure that a pulse of the form (2.156) is truly optimal, one must either impose the additional requirement that the envelope be contiguous (i.e., of connected support), or

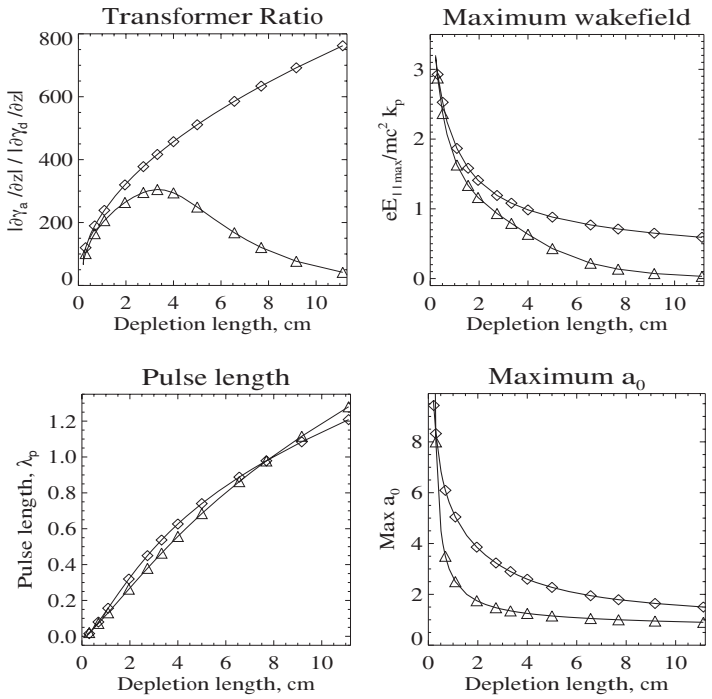

Figure 2.7. A comparison, reproduced from [10], of the transformer ratio R_3, maximum wakefield, pulse length, and peak normalized vector potential (excluding delta functions) as a function of the predicted depletion length for impulse-plus-ramp (diamonds) and Gaussian (triangles) pulses, both of energy $\mathcal{H}_{\rm EM} = 0.5$ J, optical wavelength $\lambda_0 = 1$ μm, assumed effective cylindrical cross section of $r_\perp = 10$ μm, in a cold plasma of density $n_0 = 10^{19}$ cm^{-3}.

ensure that the duration ξ_f is sufficiently short so that multiple sub-pulses in the precursor cannot be properly timed with respect to the plasma wave phase to improve performance.

Nor do shapes of the form (2.156) necessarily optimally suppresses modulational instabilities, and/or even maximize uniformity of photon deceleration, as has been suggested. It is true that almost everywhere drive photons are present, i.e., with respect to the measure $\Theta\big(a(\xi)^2\big)d\xi$, photons are predicted (albeit in a non-self-consistent model) to experience the same rate of red-shift, but it is not the case that almost all photons in the drive pulse experience the same rate of deceleration, since with respect to the more natural measure $a^2(\xi)d\xi$ (which is proportional to the photon density within the eikonal approximation), the impulsive precursor contains some finite fraction of the total number of drive photons, which increases as the pulse duration decreases. Inside the impulse, photons will tend to suffer significant linear dispersion (initial velocity spread) and rapidly changing $\kappa(\xi)$ (deceleration spread), suggesting that more instability-resistant or more uniformly-decelerating shapes can exist when these figures-of-merit are averaged over all photons in the drive pulse, including any precursor.

In Fig. 2 of [10], reproduced here for reference as Fig. 2.7, a revealing numerical comparison is made between an impulse-and-ramp shaped pulse and a Gaussian pulse of the same envelope energy U, as well as identical carrier frequencies ω_0 and both propagating in underdense plasma with the same $\omega_p \ll \omega_0$, where the pulse lengths (total duration for the ramped, FWHM for the Gaussian) are varied to also make the predicted depletion lengths equal.

For both pulses, the pulse lengths are comparable and grow monotonically with the depletion length. Over a good portion of the range of depletion lengths shown, it is somewhat puzzling as to how pulses with such similar pulse lengths but rather disparate values for the peak a can nevertheless have the same envelope energy, but it is presumably explained by their very different structure, in particular the sharply rising cusp in the ram and the long tails in the Gaussian, as well as the fact that the maximum a^2 excludes the leading impulse which contains some fraction of the pulse energy. Longer depletion lengths are associated with weaker laser drives, as expected.

As the pulse length decreases, the peak wakefield increases monotonically in either case, but slower than the depletion length decreases, so the value of the transformer ratio R_3 decreases. As the length of the Gaussian pulse increases, its peak wakefield drops well below that produced by the ramped pulse, due to plasma oscillations inside the support of the laser envelope which can cause a trailing portion of the laser in regions where $\phi'(\xi) < 0$ to reabsorb some of part of the energy previously delivered by an earlier slice. So for the Gaussian R_3 decreases at large pulse lengths as well, peaking at a near-resonant width where the FWHM is equal to one-half of a plasma period.

In contrast, by deliberate construction the ramped solution of any length suppresses wake oscillations inside the pulse, i.e., $\phi'(\xi) > 0$ everywhere within the rising ramp, so once delivered no energy is subsequently removed from the growing potential, and as the pulse length increases, the peak wake amplitude decreases but more slowly than in the Gaussian case. With the shaped ramp, the transformer ratio can be made arbitrarily large even as the peak wakefield falls by increasing the depletion length.

For us the most important lessons are perhaps not what the authors originally intended. A large peak accelerating wakefield is not positively correlated with a large transformer ratio, and in practice may be uncorrelated or even negatively correlated. Increasing the wakefield amplitude at fixed energy is accomplished by decreasing the pulse length, even below the "resonant" width. Over most of the range shown, the very large values for $L_{\rm pd}$ only underscore our contention that the depletion length-scale is simply not terribly relevant

to the LWFA in typical operating regimes. For the parameters shown, the de-tuning length for electrons in the plasma wave are expected to be $L_{\rm det} \sim O(1\text{ mm})$, and even the total length of the plasma might be only $L_0 \sim O(5\text{ mm})$ for a supersonic gas jet or $O(1\text{ cm})$ for a capillary. Where $L_{\rm pd}$ might shrink to some scale comparable to L_a and become marginally relevant, observable differences between the responses to the different types of pulse rapidly vanish, because as the duration decreases appreciably below the plasma period, all ultra-short pulses of given energy begin to look asymptotically like delta functions to the plasma, despite any internal structure.

After our various disputations, we should emphasize that while we remain in fundamental if good-natured partial disagreement with Spitkovsky, Chen, and co-workers over the import of maximizing R_3, we commend the value of their important contribution: it stimulated greater interest in issues of pulse-shaping in the LWFA, it encouraged us and others in the community to clarify our own ideas about the depletion length and its relevance (or not) in the case of the LWFA, and it led directly to the development of an estimate for the depletion length that, while not entirely self-consistent, is expected to be more accurate than the previous rough scaling-law, provided the wake potential $\phi(\xi)$ can be determined via numerical integration of the quasi-static plasma response equation. Also, as pointed out in [10], many of the laser-plasma instabilities are initially seeded by dynamic perturbations to the effective index of refraction associated with the wake field, but the authors idea of shaping so as to keep $\kappa(\xi)$ constant can be generalized to control other aspects of the resulting nonlinear refractive index, suggesting the possibility for determining pulse shapes that are predicted to optimally delay the onset of deleterious instabilities in the LWFA or perhaps optimally hasten the onset of desirable modulational instabilities in the case of the SM-LWFA.

2.7.6 Nonlinear Regime: Maximizing Wakefield Amplitude

In the nonlinear case, we have instead advocated for maximizing the *peak wake amplitude* $\mathcal{E}_+ = \max\bigl[|E_+(\xi)|\bigr]$ directly, subject to possible constraints on:

1. *peak normalized intensity* of the drive: $a^2(\xi) \leq a_0^2$;
2. *maximum envelope pulse energy*: $\frac{1}{2}\int d\xi\, a^2(\xi) \leq U$;
3. total *gross duration* of the envelope: $\bigl\{\xi \mid a(\xi)^2 > 0\bigr\} \subseteq [\xi_0\,,\, \xi_0 + T]\,;$
4. total *net duration* of the drive pulse: $\int d\xi\, \Theta_0\bigl(a(\xi)^2\bigr) < T$;

5. total number N_a of *contiguous* sub-pulses.

These resulting optimization problems are amenable to solution via the Pontryagin procedure, using the first integral to deduce the peak amplitude of the wakefield behind the drive pulse. The proof in the case of the intensity and duration constraints or intensity and energy constraints are not difficult but somewhat long and with mathematical details outweighing physical insight, so is omitted. The nature of the solution for other possible constraint sets can be deduced from this fundamental one.

Constraints on Peak Normalized Intensity

If the case where the normalized laser strength is rigidly bounded, i.e., $a^2(\xi) \leq a_0^2$, it has long been conjectured, even assumed[26, 27, 28, 29, 3], that the optimal shape consists of a series of appropriately timed square pulses at the peak value a_0^2. While certainly plausible, and in fact true, this had never actually been definitively demonstrated.

As it turns out, one can rigorously prove that in the nonlinear QSA model with prescribed drive pulses, constraining the peak laser strength still leads via the Pontryagin formalism to a solution of the "bang-bang" form, i.e., the optimal shape will be a series of square pulses, with either $a^2(\xi) = a_0^2$ or $a^2(\xi) = 0$ almost everywhere (with respect to the usual Lebesgue measure).

The catch is that because of the nonlinear synergy in the wake excitation, the precise locations of the switch-points where the optimal drive should abruptly turn on or off must be determined through a complicated additionally multivariate optimization problem, depending in intricate ways on the precise values of envelope energy or gross pulse duration that are allowed. In the nonlinear regime, excitation is most effective where $\frac{\phi'(\xi)}{1+\phi(\xi)}$ is large and positive, but unlike the linear case it is not always globally optimal to apply the drive in all regions where this quantity is positive.

Suppose that in addition to the intensity the gross duration is limited to T. The obvious pulse-train to consider consists of a series of square pulses, which are turned on and then off for half-periods $\frac{1}{2}T_{\text{NP}}$ of the local nonlinear plasma period $\mathcal{T}_{\text{NL}} = \mathcal{T}_{\text{NL}}\big[\phi'(\xi)\big]$, which is itself a functional of the wakefield amplitude. Such a solution can be compactly defined as

$$a^2(\xi) = \begin{cases} a_0^2 & \text{if } \xi_0 \leq \xi \leq \xi_0 + T \text{ and } \phi'(\xi^-) \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.161}$$

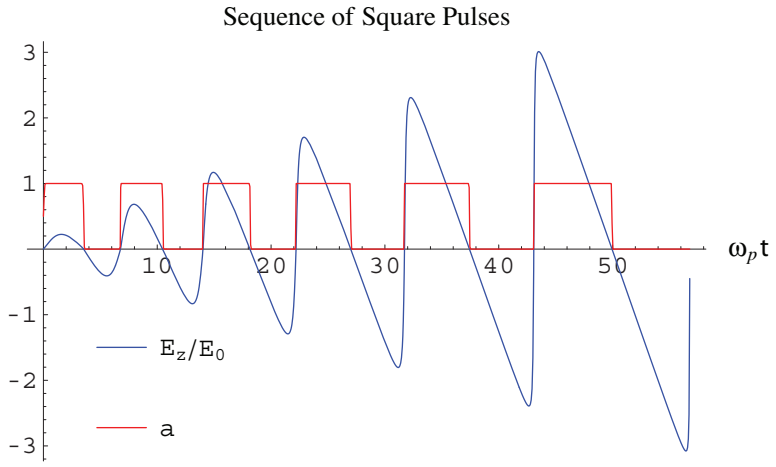

Figure 2.8. A sequence of square pulses of normalized vector potential $a = 1$, timed to coincide with intervals favorable for forcing the nonlinear Langmuir wave, along with the resulting electric field. The results were obtained by numerically integrating the quasi-static plasma response, although exact solutions can be written in terms of Jacobi Elliptic functions.

for some scaled start time ξ_0, for which a typical example is shown in Fig. 2.8. The lengthening of the nonlinear plasma period with amplitude, and the corresponding stretching of the matched square sub-pulses are clearly seen.

This solution has been widely assumed to be optimal[26, 27, 28, 29, 3]. However, while this simple solution is certainly superior to some generic timing of the sub-pulses, it may not be a global or even local optimum under the given constraints, except possibly for certain carefully chosen values of T. Due to the particular form of the quasi-static relativistic nonlinearity, and the fact that driving the nonlinear plasma wave to higher amplitude increases the effective period, and therefore the time before the next favorable opportunity to force arises, for certain sufficiently long values $T > \mathcal{T}_p$ and sufficiently strong laser strengths $a_0 \gtrsim 1$, after an initial plasma wave is first excited, it may be beneficial early on to actually abstain from driving during some ostensibly favorable opportunities to force (i.e., anywhere $\phi'(\xi)$ remains positive) or possibly even to apply the drive in regions which are locally unfavorable where $\phi'(\xi) < 0$, in order to hasten the arrival of subsequent opportunities to force in regions where $\frac{\phi'(\xi)}{1+\phi(\xi)}$ is larger and the efficacy of the forcing is enhanced.

The exact trade-offs involved can lead to complicated and sensitive dependence on the values of T and a_0. For moderate laser strengths $a \lesssim O(1)$ and pulse duration $T \lesssim O(5\ \mathcal{T}_{\rm NP})$, we expect that the standard timing in (2.161) should not be too far from optimal. It is also truly optimal if we just constrain the total number N_a of sub-pulses, rather than the total

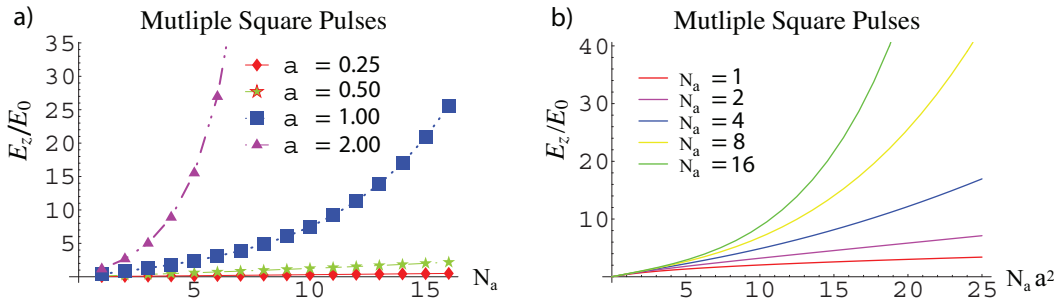

Figure 2.9. Benefits of applying multiple properly-timed square pulses. In a), the peak wakefield achieved after the last sub-pulse, as predicted by nonlinear theory, is plotted as a function of the number N_a of sub-pulses in the pulse-train, for various values of the normalized laser strength a. In b) the peak wakefield amplitude behind the last sub-pulse is plotted as a function of the quantity $N_a a^2$, for various values of N_a.

duration or total energy, which may be a more natural constraint in this context, and is certainly simpler.

In any case, what is known to be true is that if the drive is to be applied at all, it should be applied at the maximum possible strength, and the gains from multiple pulses increase (eventually nonlinearly) with increasing $a = a_0$ and with increasing N_a, as shown in Fig. 2.9 for a train of square pulses of given intensity, with timings of the form (2.161). From Fig. 2.9(a), we see that the peak wakefield achieved after the final sub-pulse always increases with the number N_a of sub-pulses, and while these gains start out small and grow approximately linearly in N_a for $a < 1$, they become dramatic and highly nonlinear for $a \gtrsim 1$. This should not be surprising: in the nonlinear regime each successive sub-pulse of duration $\frac{1}{2}\mathcal{T}_{\rm NL}$ is longer than the preceding in an absolute sense, so contains more envelope energy at a fixed value of a; and additionally each successive pulse is applied in the presence of a larger wake, and can benefit from the nonlinear synergy of the laser-wake coupling. These benefits of multiple pulses in a highly nonlinear regime are also evident in Fig. 2.9(b), where the peak wakefield amplitude behind the pulse-train is plotted versus the quantity $N_a a^2$, for trains with various number of sub-pulses. In the linear regime, we expect that $\max\big[|E_+(\xi)|\big] \propto N_a a^2 E_0$, but in the nonlinear regime we see an increasing concavity in the curves, indicating that adding more square pulses of a fixed intensity to the pulse train is more effective than increasing the peak normalized intensity a^2 of the existing square pulses.

Of course, in order to take advantage of multiple sub-pulses, their individual widths and spacings must be properly controlled, seeming to require precise knowledge (which we may not have) of the laser intensity (which may be subject to large shot-to-shot fluctuations) and background plasma density (also prone to fluctuations and inhomogeneities), prediction

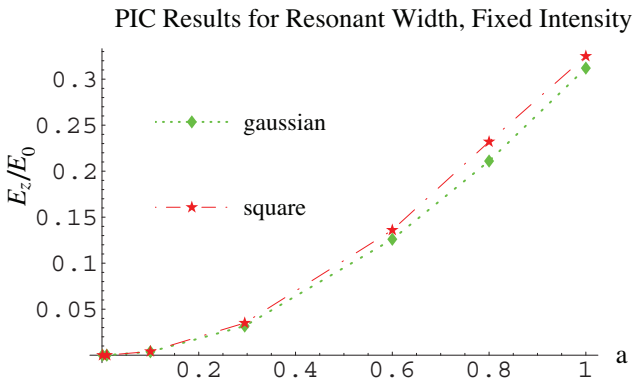

Figure 2.10. Comparison of a single square(-ish) pulse to a Gaussian pulse of the same peak value of a, with durations of $O\left(\frac{1}{2}\mathcal{T}_{\rm NL}\right)$ optimally chosen in each case. For numerical purposes, the square pulse actually had exponentially-decaying leading and trailing edges, with characteristic width comparable to the carrier wavelength λ_0.

of the switching times by numerical integration of the nonlinear wake response equation or some other model (which may not be very accurate), and the capability to produce several closely-spaced ultra-short laser pulses with steep fronts and small pedestals, all timed precisely according to the predicted switch-points (any of which may tax existing laser technology). It remains to be seen if these difficulties can be overcome in practice, but obviously it will be important to address numerically and theoretically the sensitivity of the wake response to various errors, uncertainties, and imperfections. We take just a few preliminary steps here.

Fortunately, because each of the sub-pulses in an optimized pulse train will be shorter than the nonlinear plasma period, the response is not expected to be especially sensitive to inevitable smoothing of the square pulses due to the finite bandwidth of actual lasers. The reassuring results for a simple comparison for single pulses of fixed intensity using PIC simulations are shown in Fig. 2.10. At fixed intensity, a square pulse of duration $\frac{1}{2}\mathcal{T}_{\rm NL}$ is the optimal unimodal ($N_a = 1$) pulse shape, but only marginally outperforms a Gaussian pulse of the same peak intensity but whose duration (FWHM) has been optimally chosen according to the predictions of nonlinear theory.

As a simple first look at the sensitivity to the precise timing between sub-pulses, we performed PIC simulations using two (approximately) square pulses of fixed intensity, and each of optimal "resonant" duration as predicted by the nonlinear quasi-static model assuming perfect timing, but with a variable gap between their respective times of onset.

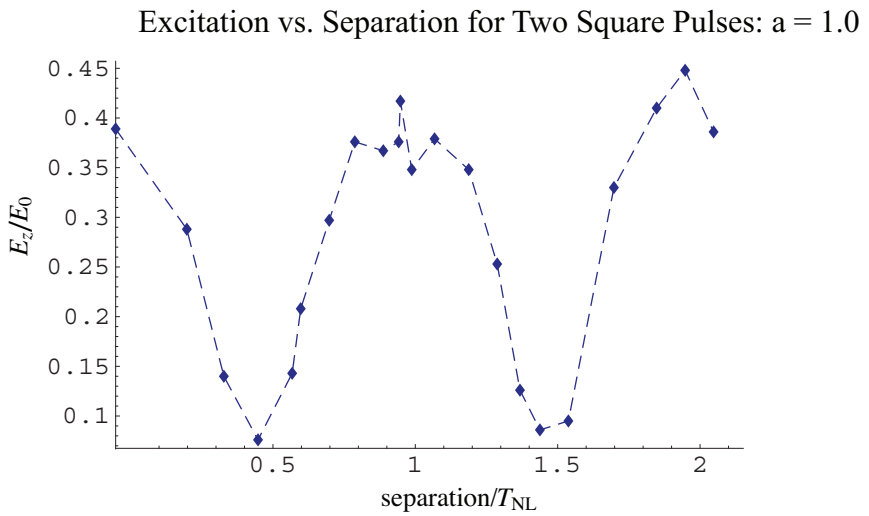

Figure 2.11. Sensitivity of the peak wakefield achieved to the delay between two square pulses. Sub-pulse durations are fixed at the optimal values predicted by nonlinear theory assuming perfect delay, and the actual separation between pulses is measured from their leading edges, so zero separation corresponds to total overlap, creating a single pulse of larger a^2, and a separation of $0.5\mathcal{T}_{NL}$ corresponds to end-to-end placement, creating a single long pulse.

The peak wakefield achieved after the second pulse is plotted as a function of the delay, expressed as a fraction of the nonlinear plasma period as estimated by nonlinear theory for the plasma wave after the first pulse. The expected resonant dependence on separation is clearly seen. Higher-frequency modulations may results from some nonlinear harmonic effects, but are more likely numerical artifacts. Some reasonable fraction of the maximum possible amplitude, of $O(10\%)$, can be achieved even if the nonlinear period is imperfectly known, or pulses are imperfectly timed, with errors or uncertainties of about $O(30\%)$. As a^2 increases, the sensitivity required to take full advantage of the nonlinearity is likely to increase, because the regions where $\frac{\phi'(\xi)}{1+\phi(\xi)}$ is especially large become quite narrow compared to the overall plasma period.

Still supposing $a(\xi) \leq a_0$, instead of specifying the gross pulse duration or the number of contiguous sub-pulses, we might also bound the envelope pulse energy or, equivalently, the net pulse duration. Local optima must still be of the bang-bang (square-pulse) form, but in these cases there is no global maximum in principle. With an ever larger number of ever narrower pulses, the peak wakefield amplitude asymptotically approaches that which can be achieved with only the envelope energy bound but no intensity bound imposed, which we analyze below. Of course, eventually we will run into bandwidth limitations which determine how narrow pulses can actually become, and timing challenges because of nonlinear wave-

front steepening, such that the intervals over which the nonlinear ponderomotive forcing is most effective narrow as the wave amplitude increases, as well as non-collisional damping processes, instabilities, or other mechanisms which will degrade the energy or coherence of the plasma wave and limit the time over which multiple sub-pulses can act synergistically.

Constraints on Envelope Energy

The simplest way to determine optimal solutions maximizing peak wakefield amplitude in this case of an envelope energy rather than peak intensity constraint is to consider the previous case of fixed U and a_0 and possibly gross duration T, where the solutions are known to be square pulses of normalized intensity a_0^2, and then relax the constraint on a_0. If $0 < U < \infty$, then as $a_0 \to \infty$, the intensity constraint becomes slack, and we see that the optimal family of solutions must converge to a sum of Dirac delta functions,

$$a^2(\xi) = \sum_{j=1}^{N_a} \alpha_j^2 \, \delta\left(\xi - \xi_j\right), \tag{2.162}$$

where N_a is the number of impulses in the pulse train, such that $1 \leq N_A \leq \infty$, the $\alpha_j^2 > 0$ are positive weights satisfying

$$\tfrac{1}{2} \sum_{j=1}^{N} \alpha_j^2 \leq U, \tag{2.163}$$

and the constants $\xi_j \in \mathbb{R}$ determine the spacing of the impulses.

All this is reminiscent of the linear case, and subject to the same caveats. Impulses require infinite bandwidth, and will suffer maximal dispersion and depletion. But again we may conclude that if envelope energy remains fixed, wake response is expected to improve asymptotically as that pulse energy is concentrated in a sequence of ever shorter pulses if their spacing can be properly timed, at least until pulse lengths shrink to the point where group velocity dispersion becomes important.

As in the linear case, it is easy to see that all of the energy allowed by the constraint should in fact be utilized to achieve the maximum possible wake amplitude, whatever constraints on pulse duration may have been imposed. First consider the case of an optimal train for finite N_a, but suppose $\frac{1}{2} \sum_{j=1}^{N} \alpha_j^2 < U$. From the quasi-static first integral (2.24) we see that the change in the QSA Hamiltonian induced by the final impulse is just

$$\Delta\mathcal{H}\left(\xi_{N_a}^+\right) = \tfrac{1}{2}\alpha_{N_a}^2 \frac{\phi'\left(\xi_{N_a}^-\right)}{\left[1+\phi(\xi_{N_a})\right]^2} + \tfrac{1}{4}\alpha_{N_a}^4 \frac{1}{\left[1+\phi(\xi_{N_a})\right]^4}, \tag{2.164}$$

and of course the maximum wakefield behind the pulse is just proportional to $\sqrt{\mathcal{H}\left(\xi_{N_a}^{+}\right)}$. Since this pulse-train is assumed optimal, it must be the case that $\Delta\mathcal{H}\left(\xi_{N_a}^{+}\right) \geq 0$, or else this last impulse would not have been imposed where it was, or imposed at all. This in turn implies

$$\phi'\left(\xi_{N_a}^{-}\right) \geq -\tfrac{1}{2}\frac{\alpha_{N_a}^2}{\left[1+\phi\left(\xi_{N_a}\right)\right]^2}, \tag{2.165}$$

and since $\alpha_{N_a}^2 > 0$ this implies

$$\phi'\left(\xi_{N_a}^{-}\right) > -\frac{\alpha_{N_a}^2}{\left[1+\phi\left(\xi_{N_a}\right)\right]^2}, \tag{2.166}$$

strictly, from which we can deduce

$$\frac{\partial\Delta\mathcal{H}\left(\xi_{N_a}^{+}\right)}{\partial\alpha_{N_a}^2} > 0, \tag{2.167}$$

so we can always increase the final wake amplitude by just dumping the remainder of the budgeted envelope energy into the last impulse without making any changes to its relative position. Since this strict inequality holds for arbitrary N_a it will also hold in the limit as $N_a \to \infty$ for any convergent sum of impulse strengths.

But unlike the linear case, in the absence of an explicit constraint on the number of impulses, or on the gross or RMS duration of the drive pulse, due to the synergistic nature of the nonlinearity it can be shown that in principle $N_a + 1$ pulses of energy U can always strictly out-perform N_a pulses if the strengths and delays are suitably chosen. This can be shown by some lengthy but straightforward algebraic manipulations of quantities like (2.164), but the detailed proof will be omitted. The basic idea is to shave a little energy off of the last (N_ath) impulse of a supposedly optimized pulse-train and add an additional impulse at the next optimal opportunity. For a given intensity of a carefully-timed laser pulse, the effectiveness of the drive increases with the pre-existing plasma wave amplitude up to the nonlinear wave-breaking limit.

Despite previous claims in the literature [10, 11], without enforcing $N_a = 1$ explicitly, or demanding that $T < \mathcal{T}_{\mathrm{NL}}$, or perhaps incorporating damping on the plasma wave, the single delta function does not excite the largest wakefield for given envelope energy.

Of course, a drive pulse of a given finite energy U should not be expected to excite a wakefield of arbitrarily large amplitude, so eventually the improvements associated with adding another impulse will suffer from diminishing marginal returns. For a given U, the actual asymptotic limit is not known analytically, but presumably monotonically approaches

$E_{\rm WB}$ as U increases. We have not pursued these question further, considering them not terrifically relevant within a model without any self-consistent depletion effects on the drive.

The addition of a constraint on gross or RMS duration lead to the solutions within the same family consisting of a train of impulses, but complicate the secondary maximization problem involved in apportioning the envelope energy between the α_j^2 and choosing the locations ξ_j. For finite (and in practice, reasonably small) values of N_a, the impulse strengths and locations can be found by numerically optimizing $\sum_{j=1}^{N_a} \Delta\mathcal{H}\left(\xi_j^+\right)$ subject to the energy constraint. In general the impulses will be spaced at intervals close to *but not exactly at* multiples of the effective nonlinear plasma period $\mathcal{T}_{\rm NL}$. For any value of α_j^2, the (locally) optimal location ξ_j for the jth impulse will involve some trade-off between making $1+\phi(\xi_j)$ small and making $\phi'(\xi_j)$ large so as to enhance the nonlinear laser-wake coupling. As in the case of square pulses, with the total duration also limited, some potential effectiveness of the jth impulse may be sacrificed to hasten or improve favorable opportunities at which to apply the $(j+1)$th.

Comparison of Pulse-Shaping at Fixed Energy or Intensity

For single pulses, results in the nonlinear regime are qualitatively similar to those analyzed previously in the linear case, apart from the higher fields achieved. Typical results for Gaussian pulses are shown in Fig. 2.12. For given peak intensity, the plasma response is expected to peak for pulse widths somewhere near the nonlinear resonant width $\mathcal{T}_{NL}$ and fall off towards zero for longer or shorter pulses. The maximum wake field $\mathcal{E} \sim 1.2$ occurs at about $\omega_p\sigma \approx 1.3$, but the response is much less sensitive to the Gaussian width than was linear case, and pulses with $0.5 \lesssim \omega_p\sigma \lesssim 2.5$ are all nearly equally effective (within about 10%) in wakefield generation.

Comparing nonlinear theory at fixed envelope energy U to PIC simulations performed at fixed vacuum EM energy $\mathcal{H}_d$, we see that the simulations and QSA both predict that the plasma responds continues to increase monotonically until $\sigma \sim O(2\lambda_0)$, where the eikonal approximation for the pulse energy breaks down, and the actual value of a_0^2 at fixed EM energy actually vanishes rather than diverges in the limit $\omega_p\sigma \to 0$.

Of course in either case, by splitting the same energy between two or more pulses of just the right widths and separation, performance can be further enhanced.

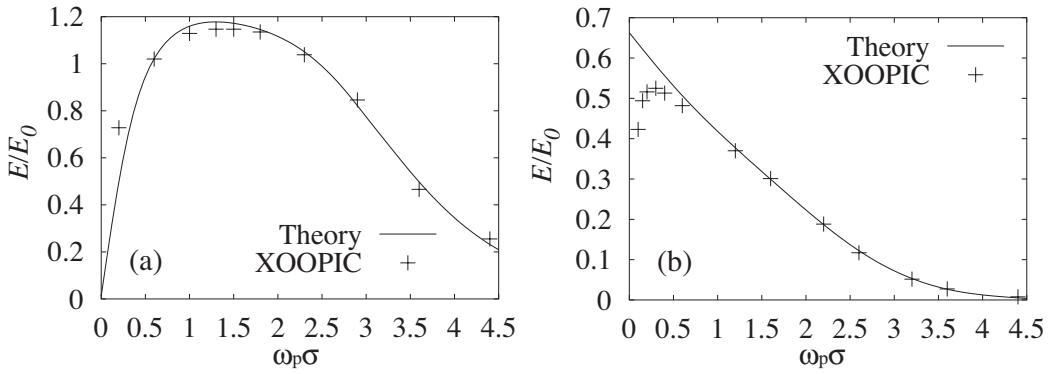

Figure 2.12. Nonlinear plasma response E/E_0 for Gaussian pulses of fixed (a) peak intensity and (b) vacuum EM energy. Theory solves QSA response numerically for $\omega_0/\omega_p = 10$ assuming fixed envelope energy U and (a) $a = 3.0$, (b) $a = 1.12$ for $\omega_p\sigma = 1.2$. For short pulses, $\sigma \sim \lambda$, the envelope approximation breaks down as seen in (b).

2.8 A Caveat: the Importance of Physical Electric Fields in PIC codes

When evaluating the role of pulse-shaping in optimizing LWFA performance, it is important to ensure that the field profiles considered are physically realizable, at least in principle, or else spuriously large wakefields can result.

Originally, out of convenience or naiveté, we first performed certain LWFA and other laser-plasma simulations assuming the transverse electric field of the Gaussian laser pulse might be described by only the eikonal term, i.e.

$$\mathcal{E}_x(\xi) = a_0 \tfrac{\omega_0}{\omega_p} e^{-\frac{\xi^2}{2\tilde{\sigma}^2}} \cos\left(\tfrac{\omega_0}{\omega_p}(\xi - \xi_0)\right), \tag{2.168}$$

our thinking being that for long pulses the remaining term arising from the derivative of the envelope will be small, while for short pulses the exact form of the oscillation within a narrow impulse-like Gaussian envelope will not matter much anyway. Our thinking was flawed.

Fig. 2.13 plots typical results for a moderate value of $a \sim 0.3$. Results for the case of fixed intensity appear fine, but at fixed energy the behavior for small $\sigma\omega_p$ looks bizarre, with the peak wakefield amplitude spiking well above the predictions of QSA theory near the leading edge of the plasma where the pulse enters, but then falling off in more interior points.

Subsequent simulations showed strong dependence on the precise phase of the carrier oscillation, as determined by ξ_0, and independent 1D fluid simulations convinced us that

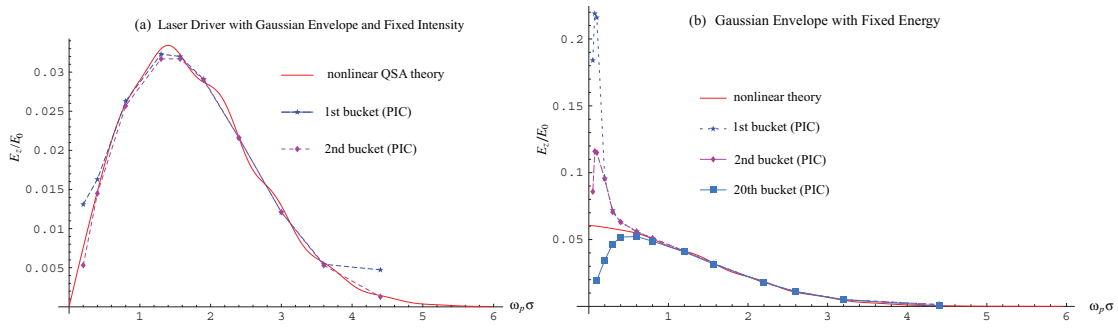

Figure 2.13. Comparison of plasma responses E/E_0 for optimized Gaussian pulses of fixed (a) normalized intensity corresponding to $a = 0.295$ and (b) energy corresponding to $a = 0.295$ at $\omega_p\sigma = 1.57$. Wakefield buckets are counted in the temporal order in which they were excited, i.e., in from the leading edge of the plasma. The theoretical (nonlinear QSA) prediction is based on constraining the envelope energy as usual, while the PIC simulations assumed an eikonal form for the electric field. The spurious spiking in the wake immediately behind the pulse will be explained below.

this spiking was not due to dispersive effects. Further simulations showed that spurious, persistent transverse EM fields could also be excited in the plasma behind the laser by such sources, and simple calculations of growth rates indicated that (inverse) two-phonon decay or other instabilities were not to blame.

Finally, we realized these spurious fields were an artifact of the chosen form for the electric field, where what we thought was a minor simplification led to a laser pulse which, we shall argue, cannot be physically realizable, and which is therefore in turn capable of all sorts of unphysical effects.

Typically in PIC codes, radiation fields are launched as time-dependent boundary conditions on the transverse electric field at one or more planar surfaces outside the plasma, often in the form of a sinusoidal carrier modulated by a prescribed envelope function. While these pulse shapes will then automatically satisfy (apart from ordinary numerical errors) the Maxwell equations in the initial vacuum region ahead of the plasma, one must also take care that their corresponding "sources" be physically realizable.

In particular, we will see that under reasonable assumptions, the radiative field components must average to zero at any fixed observation point sufficiently far from the actual source of the radiation. Such limitations were discussed as far back as 1849 by Stokes[30], and have more recently been invoked in the debate surrounding vacuum laser acceleration[31]. Radiation fields launched in the vacuum with a non-trivial DC component can excite unphysically large longitudinal fields or erroneous transverse fields after passing through the plasma, particularly in short pulses of duration Δt about $O(10)$ laser wave-

lengths or less, or in longer pulses with features (such as step functions) which rise or fall too steeply.

In three dimensions, the usual argument against unipolarity proceeds by noting that physically realizable sources for classical radiation fields will ultimately consist of a set of charges which, although temporarily accelerated, remain bounded in space for all relevant time; that is, at any time t, all accelerating charges contributing to the radiation lie within some finite distance R from, say, their centroid $\boldsymbol{x}_0$ at $t = 0$.

If $\boldsymbol{E}(\boldsymbol{x}, t) = -\frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{A}(\boldsymbol{x}, t)$ includes the radiative component of the electric field, such that $\boldsymbol{A}(\boldsymbol{x}, t)$ is the vector potential in the Coulomb gauge ($\nabla \cdot \boldsymbol{A}(\boldsymbol{x}, t) = 0$), then for any observation point $\boldsymbol{r}$ sufficiently distant from all sources, i.e., $|\boldsymbol{r} - \boldsymbol{x}_0| \gg R$, we must have

$$\int_{-\infty}^{\infty} dt\, \boldsymbol{E}(\boldsymbol{r}, t) = \tfrac{1}{c}\boldsymbol{A}(\boldsymbol{r}, -\infty) - \tfrac{1}{c}\boldsymbol{A}(\boldsymbol{r}, +\infty) = \boldsymbol{0}. \tag{2.169}$$

For supposing otherwise, in the frequency-domain representation of the field, namely

$$\boldsymbol{E}(\boldsymbol{r}, \omega) = \tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dt\, e^{i\omega t} \boldsymbol{E}(\boldsymbol{r}, t), \tag{2.170}$$

a non-vanishing DC component will be present, i.e., $\boldsymbol{E}(\boldsymbol{r}, \omega = 0) \neq \boldsymbol{0}$. But any DC component of the electric field will necessarily correspond to a *static* solution to Maxwell's equations, so that far from the sources it must fall off asymptotically as $|\boldsymbol{E}(\boldsymbol{r}, \omega = 0)| \sim |\boldsymbol{r} - \boldsymbol{x}_0|^{-2}$ (as in Coulomb's law), or even faster if a monopole moment is not present. It can easily be shown from energy conservation arguments, however, that true radiative fields must fall off with the slower $|\boldsymbol{E}(\boldsymbol{r}, \omega)| \sim |\boldsymbol{r} - \boldsymbol{x}_0|^{-1}$ scaling.

However, in truly one-dimensional geometries, what were considered point sources in $3D$ are interpreted as transverse charge sheets of uniform surface charge density, which are not bounded in three-dimensional space, and whose fields do not fall off with distance from the sheet, so the above argument fails.

Yet, in one-dimensional situations, it is particularly critical to obtaining physically realistic results that the radiative electromagnetic fields average to zero, because in this geometry transverse canonical momentum is exactly conserved. From (2.169), a non-vanishing DC component of the radiative electric field implies

$$\boldsymbol{A}(\boldsymbol{r}, t = \infty) \neq \boldsymbol{A}(\boldsymbol{r}, t = -\infty). \tag{2.171}$$

In order that total transverse canonical momentum remain conserved in the presence of this unphysical field, a plasma electron oscillating around the position $\boldsymbol{r}$ under the influence of

the field will be left with a compensatory spurious component of kinetic momentum, which through ponderomotive effects in the tail of the pulse can generate unphysical longitudinal wakefields, and also can act as a source for spurious transverse electromagnetic components.

In a PIC simulations of short pulses, this can have disastrous effects, as shown above in Fig. 2.13, or again in Fig. 2.14 at a small value of $a = 0.005$ to avoid any nonlinear effects. For poor choices of the carrier phase, the error involved becomes quite pronounced for pulses whose envelopes change on the order of the fast carrier time-scale $\sim \omega_0^{-1}$. It turns out that this error is solely due to the inappropriately launched electric field, that leaves electrons with unphysical kinetic momentum. Electrons "miss" the components dropped from the electric field, and act out in strange ways.[13]

However, the errors are not simply the usual ones of an eikonal approximation pushed too far. If in (2.168) we use $\xi_0 = \frac{\pi}{2}\frac{\omega_p}{\omega_0}$ instead of ξ_0 the error incurred from the perspective of an asymptotic eikonal expansion is of exactly the same order in $\frac{1}{\omega_0 \sigma}$, but the electric field is now of odd parity and its integral will automatically vanish, along with the spurious electric fields left behind in the plasma.

In [32], we presented what we thought was a cogent argument as to why fields must still integrate to zero in one-dimension. This proof, summarized here, now appears incorrect, but we will muster new arguments in favor of the same conclusion.

Obviously a function of the form $E_x(z,t) = \Theta(z - ct)$ solves the homogeneous wave equation, but does not integrate to zero, so any proof will have to limit the kinds of fields that can be achieved by actual sources. Our argument was as follows: in the case of a 1D system, the solenoidal current density is just the geometrically-transverse component of the full current density, and the Fourier transform of the Coulomb-gauge vector potential satisfies the Helmholtz equation

$$\left(\frac{\partial^2}{\partial z^2} + \frac{\omega^2}{c^2}\right) \boldsymbol{A}(z,\omega) = -\frac{4\pi}{c}\left[\boldsymbol{J}(z,\omega) - (\hat{\boldsymbol{z}} \cdot \boldsymbol{J}(z,\omega))\hat{\boldsymbol{z}}\right]. \tag{2.172}$$

For a source consisting of a possibly large but finite number of charge sheets, each of finite surface charge density and moving with finite transverse velocity, and each assumed to remain bounded in *longitudinal* position z for all time, the Fourier transform $\boldsymbol{J}(z,\omega)$ of the source current density, regarded as a function of ω at any fixed position z, should possess no poles at $\omega = 0$.

We then concluded (apparently wrongly) that the only possible singular contribution to the DC component of the source might be a term proportional to a Dirac delta function

[13] Like cats left home alone for too long, or teenagers pretty much all the time.

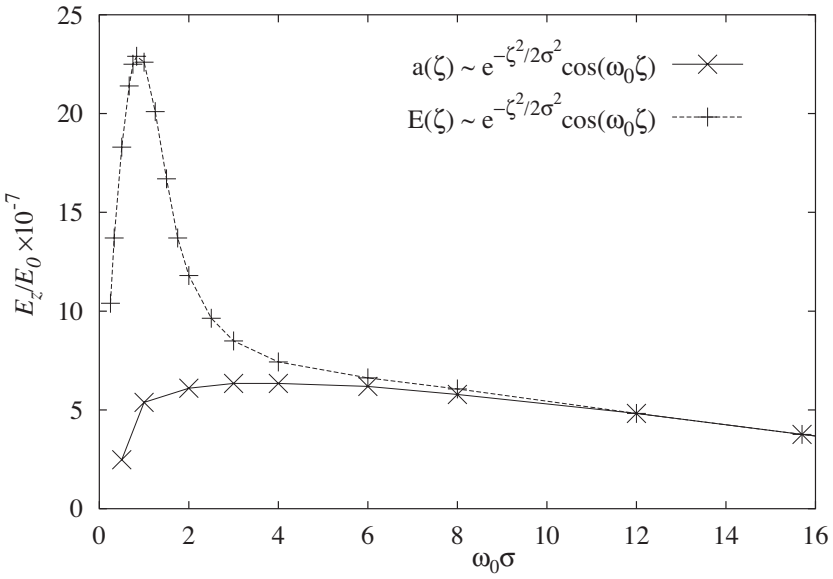

Figure 2.14. The folly of not enforcing $\int E_\perp \, dt = 0$. The solid line based on a physically-consistent electric field for a laser of $a_0 = 0.005$ correctly predicts E_z, whereas the dotted line is based on an inconsistent approximation to is in error by a factor of $O(5)$, far greater than the absolute or relative error between the two forms of the drive field. Note that the inaccurate pulse is simply a Gaussian that does not average to zero because of the poor choice in the carrier phase; there is otherwise nothing "pathological" in the profile. For notational convenience we here work in the coordinate $\zeta = \xi/\omega_p$.

$\delta(\omega)$ corresponding to some constant current density at one or more longitudinal positions. Higher-order derivatives of delta functions cannot appear: a term proportional to $\frac{d^n}{d\omega^n}\delta(\omega)$ in the frequency domain would correspond in the time domain to a contribution in the transverse current density growing as $\sim t^n$ indefinitely, which is impossible because only a finite amount of surface charge density can accumulate at any longitudinal position, and any charge must move with a subluminal velocity.

Supposing all of this is true, the conclusion quickly follows. By convention, one can take $\boldsymbol{A}(z = -\infty, \omega = 0) = \boldsymbol{0}$ as a constant of integration, and by integrating the Helmholtz equation up to z for $\omega = 0$, one finds that for any longitudinal position z, as $\omega \to 0$, $\boldsymbol{A}(z, \omega)$ remains either finite or, possibly, contains a term proportional to $\delta(\omega)$. In either case it follows that

$$\lim_{\omega \to 0} \boldsymbol{E}(z, \omega) = \tfrac{i}{c} \lim_{\omega \to 0} \omega \, \boldsymbol{A}(z, \omega) = \boldsymbol{0}, \tag{2.173}$$

implying the infinite-time average of $\boldsymbol{E}(z, t)$ vanishes.

The problem with this argument seems to be that, while we somehow allowed for the possibility that $\boldsymbol{J}(z, \omega)$ might contain a Dirac delta function $\delta(\omega)$, we missed another gen-

eralized function, the principle value $\mathcal{P}\frac{1}{\omega}$. The action of both of these as kernels can be thought of in terms of integrating ω^{-1} along some different contour in the complex frequency plane, namely a small circle for the delta function, and the real line with a small segment around $\omega = 0$ excluded for the principle value. The distribution $\mathcal{P}\frac{1}{\omega}$ is not the same thing as $\frac{1}{\omega}$ so our prohibition against actual poles is evaded, while $\omega\mathcal{P}\frac{1}{\omega} = 1 \neq 0$. In the time domain this will correspond to step function terms in $\boldsymbol{J}(z,t)$ and $\boldsymbol{E}(z,t)$.

However, even if our mathematical argument was flawed, we still argue on physical grounds that 1D fields should be non-unipolar. The world is three-dimensional, and the sources that we can arrange and manipulate are finite, so what we call 1D radiation fields are ultimately idealizations, introduced for our mathematical convenience, of 3D fields where the characteristic diffraction angle $\theta_d \sim \frac{\lambda}{\sigma_\perp}$ and perhaps other ratios of length scales are suitably small. Here λ is the characteristic wavelength for the radiation (or one band of it), and $\sigma_\perp$ is a characteristic scale-length for the transverse spot size of the radiation. Where this idealization is not convenient, even if ostensibly valid, we should avoid or postpone it.

In many cases we can usually get away with taking the appropriate limit of the 3D fields

$$\boldsymbol{E}(z,t) = \lim_{\sigma_\perp \to \infty} \boldsymbol{E}(x, y = 0, z = 0; \sigma_\perp) \tag{2.174}$$

first, i.e., at the start of some problem, and then proceed in truly one-dimensional geometry. However, this procedure may occasionally lead to difficulties if this infinite limit does not commute with other limits, which we contend is the case here.

If we define the (principle-value) integral of a 1D radiation electric field at some position z remote from the source as a double-limit of the form

$$\int_{-\infty}^{\infty} dt\, \boldsymbol{E}(z,t) = \lim_{T_0 \to \infty} \int_{-T_0}^{T_0} dt \lim_{\sigma_\perp \to \infty} \boldsymbol{E}(x = 0, y = 0, z; \sigma_\perp), \tag{2.175}$$

there there appears to be no unassailable mathematical or physical reason of which we are aware that definitively rules out a non-zero answer. However, if we reverse the order of limits, the quantity

$$\begin{aligned} \int_{-\infty}^{\infty} dt\, \boldsymbol{E}(z,t) &= \lim_{\sigma_\perp \to \infty} \lim_{T_0 \to \infty} \int_{-T_0}^{T_0} dt\, \boldsymbol{E}(x = 0, y = 0, z; \sigma_\perp) \\ &= \lim_{\sigma_\perp \to \infty} \int_{-\infty}^{\infty} dt\, \boldsymbol{E}(x = 0, y = 0, z; \sigma_\perp) = \lim_{\sigma_\perp \to \infty} \big[0\big] = 0, \end{aligned} \tag{2.176}$$

because for any physical 3D field of any specified characteristic spot size $\sigma_\perp$ satisfying $\lambda_0 < \sigma_\perp < \infty$, we know the radiation electric field must average to zero at any position remote from the source.

We contend that the latter choice makes more physical sense. Operationally, for any given experimental configuration, we can easily imagine approaching the $T_0 \to \infty$ limit for any given value of $\sigma_\perp$ simply by collecting data longer. Granted, no experiment (or experimentalist) can last forever, but in a typical case where the pulse eventually turns off or decays rapidly (perhaps exponentially or better), any truncation error due to a finite observation time will eventually decrease at the same rate, and rapidly fall below other sources of measurement error. In contrast, in order to approach the $\sigma_\perp \to \infty$ limit, while keeping the observation time T_0 and certain other physically relevant parameters such as the carrier wavelength and on-axis intensity fixed, every increase in $\sigma_\perp$ envisioned would require a new experimental setup, with proportionally larger optics and sources, more power, etc., if even possible at all.

Said another way, the integral $\int_{-\infty}^{\infty} dt\, \boldsymbol{E}(x = 0, y = 0, z; \sigma_\perp)$ will vanish for any finite value of the diffraction angle θ_d, so in any sort of asymptotic expansion of this integral for small θ_d (or large $\sigma_\perp$), there is no sense in which any non-zero term (of finite order, or even of exponential accuracy) in θ_d can be meaningfully said to be small compared to zero.

In any case, avoiding such spurious fields in PIC codes or other numerical simulations is not difficult. At the boundary in question where the radiation field is launched , one should choose a prescribed analytic form for the transverse vector potential $\boldsymbol{A}(t)$ that vanishes as $t \to \pm\infty$, rather than the electric field $\boldsymbol{E}(t)$, and then determine the latter via $\boldsymbol{E}(t) = -\frac{1}{c}\frac{d}{dt}\boldsymbol{A}(t)$, retaining all terms, not only the usual eikonal contribution consisting of the derivative of the carrier phase multiplied by the pulse envelope, but also the carrier phase multiplied by the derivative of the envelope. This is shown for the simple Gaussian example in Fig. 2.14, where correct behavior is found when the electric field is properly derived from the vector potential.

2.9 Colliding Beam Accelerator

Finally, we consider the generalization to the case of a so-called Raman-pumped LWFA, otherwise known as the Colliding Beam Accelerator (CBA).[14] The basic idea is simple: a

[14]Ideas in this section were developed in particularly close collaboration with R. R. Lindberg.

long counter-propagating (with respect to the electron beam) laser pump beam of peak strength a_1 can supply most of the energy for wake excitation via a Raman-type instability, while a short co-propagating seed pulse of peak strength a_0 can provide the coupling, resonant or otherwise. With a total energy budget for seed and pump combined, one might anticipate some optimal shape for the seed and optimal apportionment of energy between seed and pump. In actual fact, for fixed energy this approach can never exceed the efficiency (in energy transfer) or efficacy (*viz a viz* peak fields produced) of a single intense short pulse or a carefully-spaced train of co-propagating pulses, but it might enjoy certain practical advantages if available technology limits the energy which can be supplied in the short pulse or the accuracy in timing multiple pulses.

Because of the greater difficulties in analyzing the CBA, the results here are somewhat less rigorous, and hence less definitive, than some of those above, but we believe our general arguments for the greater energetic efficiency of the LWFA are valid.

2.9.1 Wake Generation in the CBA

The proposed colliding beam accelerator is realized using two counter-propagating lasers: a short seed pulse whose duration is of order $O\left(\omega_p^{-1}\right)$, and a long, essentially CW pump beam, upshifted in carrier frequency from the seed pulse by $\delta\omega \sim \omega_p$. Assuming as usual a cold-but-collisionless underdense electron plasma, the dynamical plasma response was developed in [6] by considering the equation of motion for a single electron. The phase $\psi_j \equiv 2k_0\, z_j - \delta\omega\, t_j$ of the jth electron in the beat wave then satisfies

$$\frac{\partial^2}{\partial\xi^2}\psi_j + \frac{\omega_{\rm B}^2}{\omega_p^2}\sin\psi_j = -\sum_{\ell=1}^{\infty} n_\ell\, e^{i\ell\psi_j} - 2\frac{\omega_0}{\omega_p}\frac{E_{\rm fast}}{E_0} + c.c., \tag{2.177}$$

where $\omega_{\rm B} \equiv 2\omega_0\sqrt{a_0 a_1}$ is the nonlinear bounce frequency in the ponderomotive bucket.

From (2.177) we can identify two distinct plasma waves generated by the lasers. The first wave arises from the sum over the Fourier harmonics n_ℓ of the density, predominantly involving photon deceleration (i.e., red-shift) associated with the ponderomotive beating of the two lasers. Because of its slow phase velocity ($\frac{v_{\rm slow}}{c} = \frac{\Delta\omega}{2\omega_0} \ll 1$), this plasma wave is unsuitable for electron acceleration (but could prove useful for accelerating ions.)

The second plasma wave $E_{\rm fast}$ has a non-zero average over the electron phase; its importance was previously noted in [33] for the proper analysis of Raman backscatter. This second wave is dominated by photon exchange between the two lasers which, in order to conserve linear momentum, must deposit net momentum in the plasma. Because the mo-

mentum exchange occurs on the time-scale governed by the short pulse duration $\sim O\left(\omega_p^{-1}\right)$, significant momentum can be delivered to the plasma electrons. The associated electron current, in turn, generates a plasma wave via Faraday's law: $\frac{\partial}{\partial t}E_{\text{fast}} = -4\pi\langle J_e\rangle$. Because this wave has a phase velocity equal to the seed pulse's group velocity ($v_{\text{fast}} \approx v_g \lesssim c$), it can accelerate relativistic electrons.

The CBA as envisioned by [6] has two significant attractive characteristics: its potential applicability as part of a multi-staged plasma LINAC; and its potential capability to create enhanced acceleration gradients for a given laser intensity. The former would rely on manipulation of the plasma wave phase through the use of multiple pumps, each appropriately timed and detuned from the seed pulse. While this does present the interesting possibility of a staged accelerator unlimited by the de-phasing length, its analysis is really beyond the scope of the present chapter, and we will not consider it further here. Instead, we focus on the second feature, and compare the plasma response of the single-stage CBA and LWFA under what we hope are at least moderately realistic constraints.

2.9.2 The Linear Regime of the CBA

If the slow plasma wave remains linear, the harmonics n_ℓ for $\ell > 1$ in (2.177) can all be dropped, and one can in principle find a closed form expression for the accelerating

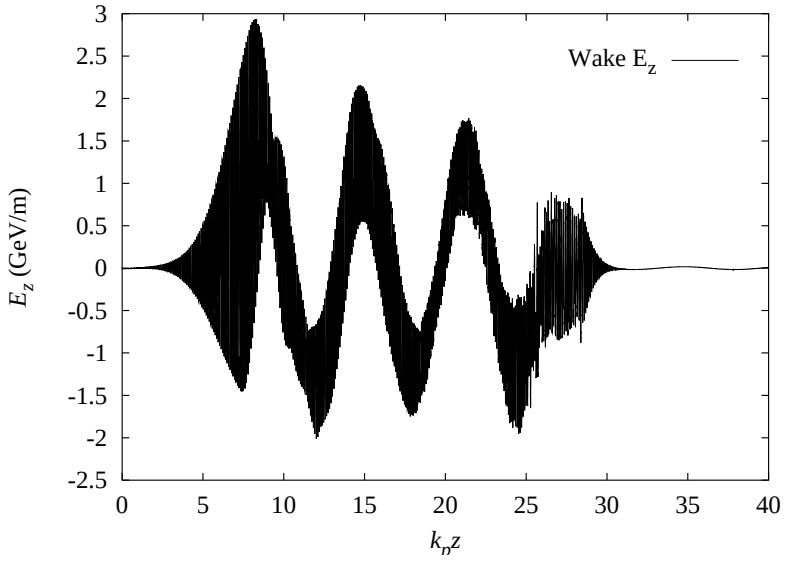

Figure 2.15. CBA plasma response when the slow wave remains linear. The fast, accelerating wake of amplitude ~ 1 GeV/m is clearly superimposed on the high frequency slow wake. Simulated with 1D-XOOPIC, using parameters $a_0 = a_1 = 0.02$, $\delta\omega = 1.1\omega_p$, $\sigma = 2\omega^{-1}$.

wakefield. For simplicity, consider as a canonical example a Gaussian seed pulse of RMS width $\sigma = 2\omega_p^{-1}$, and pump detuning $\delta\omega = 1.1\,\omega_p$. In this case, the peak electric field can be shown to reach a maximum value

$$E_{\text{fast}} \approx 0.6\,E_0 \frac{16 a_1^2 \omega_0^3}{\omega_p^3} a_0^2. \tag{2.178}$$

In the linear regime of the LWFA, for a single pulse with a Gaussian profile $|a|^2 = a_0^2 e^{-\xi^2/\sigma^2}$ the optimal width has $\sigma\omega_p = 2$, for which the peak wakefield amplitude is

$$E_z \approx E_0 \sqrt{\tfrac{\pi}{4e}} a_0^2 \approx 0.6\,E_0 a_0^2, \tag{2.179}$$

so in this case, the addition of the pump multiplies the longitudinal field (2.179) achieved by the optimized, single-Gaussian-pulse LWFA by an additional factor of $16\,a_1^2\,\frac{\omega_0^3}{\omega_p^3}$, and therefore the CBA outperforms the single- pulse LWFA whenever $4a_1 > \left(\frac{\omega_p}{\omega_0}\right)^{3/2}$. For this reason, it is apparent that the CBA might effectively use low intensity seed and pump lasers ($a_0, a_1 \sim 0.004$ – 0.02) in order to achieve plasma wakefields $E_z \sim O\,(1$ GeV/m) in the linear regime.

One should also note that the accelerating wakefield will be accompanied by a large amplitude, high frequency, slow plasma wave. Because this slow plasma wave averages to zero over the phase of the electrons, it should not significantly affect particle acceleration, although it will inevitably detract from efficiency of desirable energy transfer.

To illustrate a pump-enhanced wake, we have simulated the CBA with a density $n_0 = 10^{18}$ cm^3 and $a_0 = a_1 = 0.02$ using the 1D-XOOPIC Particle-In-Cell code. We plot the longitudinal field in Figure 2.15. The seed pulse propagates to the left, colliding with the pump at $k_p z \approx 30$, so that for $8 \lesssim k_p z \lesssim 25$, one can see both the accelerating wake (of wavelength $\sim \lambda_p$) and the superimposed slow-velocity wake (of wavelength about $\frac{1}{2}\lambda_0$). Figure 2.15 exhibits an enhanced accelerating wake $E_{\text{fast}} \approx 1$ GeV/m, which is about ~ 43 times greater than that for the single-pulse LWFA. To achieve the same plasma response, an short-pulse LWFA system would require a laser with intensity $4.5{\cdot}10^{16}$ W/cm^2, as compared to the CBA case with seed and pump intensity at 10^{15} W/cm^2. The enhancement observed in the simulations is, however, considerably less than the factor of ~ 235 increase predicted by (2.178). Presently, it is unclear if a more complete analysis of (2.177) will give greater agreement; based on simulation results over a wide range of system parameters, we shall consider the analytic enhancement formula (2.178) as an upper bound for the linear CBA wake.

2.9.3 Particle-Trapping Regime for the CBA

As either laser driver intensity is increased, larger wakes are possible, but the previous linear analysis may no longer hold. For sufficiently intense lasers, one has $\omega_{\rm B}^2 = 4\omega_0^2 a_0 a_1 > \omega_p^2$, so that the RHS of (2.177) becomes relatively unimportant, and may be neglected. In this regime, electrons can become trapped in the ponderomotive bucket, where they can acquire an average momentum provided the seed pulse duration and detuning are appropriately chosen. In the case where $\omega_{\rm B}^2 \gg \omega_p^2$, one can estimate the average momentum imparted, which produces an accelerating wake amplitude

$$E_{\rm fast} \approx 2E_0 \left(a_0 a_1\right)^{1/2} . \tag{2.180}$$

Because of this scaling, the CBA wakefield can be considerably larger than that of the conventional LWFA for a_0^2, $a_1^2 \ll 1$. As an example, we show the 1D-XOOPIC results for some typical model parameters in Figure 2.16. In this case, the two pulses collide at $k_p z \approx 80$, where one can see an approximate factor of ~ 25 increase in the wake from $k_p z > 80$ (single pulse regime) to $15 < k_p z < 80$ (combined pulse and pump). We find that, in order to obtain the same wake using the LWFA, the laser power must be increased by more than an order-of-magnitude (from $2.7 \cdot 10^{16}$ W/cm^2 to 10^{18} W/cm^2). This increase is also less than the theoretically-predicted factor of ~ 125 from (2.180).

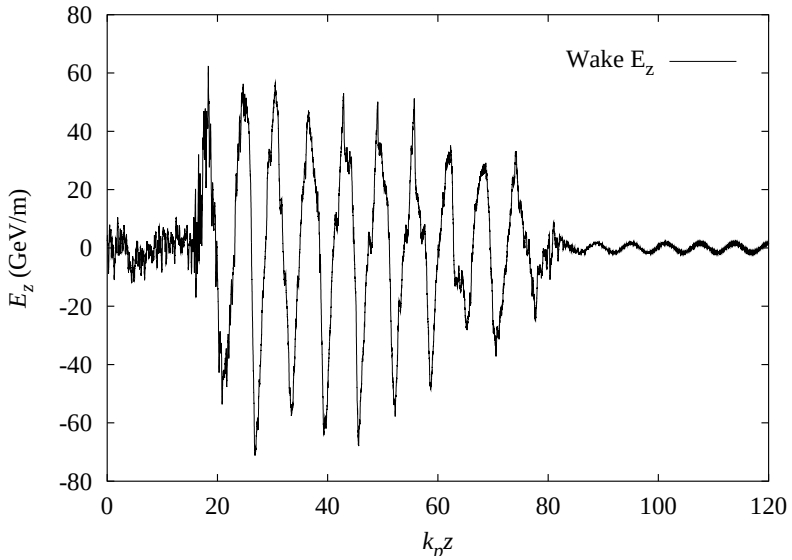

Figure 2.16. The particle trapping regime of the CBA ($4\omega_0^2 a_0 a_1 \gg \omega_p^2$), for which the enhanced wake $E_{fast} \approx 40$ GeV $\approx 25 E_{LWFA}$. Simulated using 1D-XOOPIC, with parameters $a_0 = a_1 = 0.1$, $\omega_0/\omega_p = 10$, $\Delta\omega = 2\omega_p$, $\sigma = \omega_p^{-1}$.

2.9.4 Comparison of CBA to Conventional LWFA

As we have seen, for fixed laser *intensities*, one can use the counter-propagating pump of the colliding beam accelerator to significantly enhance (by one or two orders-of-magnitude) the peak longitudinal field achievable by a single Gaussian pulse. From the theoretical formulae given by (2.178) and (2.180), we predict that for sufficiently small laser intensities, i.e., for which $a_0 a_1 < \frac{\omega_p}{\omega_0}$, a pump of amplitude a_1 increases the LWFA wake (i.e., beyond that which would be produced by the seed alone) by a factor of $16\, a_1^2\, \frac{\omega_0^3}{\omega_p^3}$.

As the pump and/or seed pulse become more intense, the wake amplification factor scales as $\frac{1}{\sqrt{a_0 a_1}}$. Thus, the CBA might be an attractive alternative to the LWFA if limited to very low laser powers, although this has not been the main direction of contemporary plasma-based accelerator research, which has continually moved toward ever higher peak powers.

Because the CBA still requires a short, near-resonant seed pulse of nontrivial amplitude (and hence a CPA system in practice), it is natural to ask whether it would be simpler or more effective to implement the counter-propagating, detuned pump of the CBA with the seed remaining at lower power, or rather to improve the amplifier and optics for high-power single-pulse LWFA experiments. This question depends in part on technological requirements and capabilities that are constantly evolving. Thus, we will simplify the comparison by presuming an existing CPA system producing a Gaussian pulse of minimal width σ and a certain total energy budget U, yet keep in mind that any findings must inevitably be weighed by more detailed experimental considerations.

In order to accelerate electrons over the length L_a, the CBA requires a pump at least of length $2L_a$. For the present case, we will consider an acceleration length to be equal to the de-phasing length $L_a = L_{\rm det} \approx \lambda_p \frac{\omega_0^2}{\omega_p^2}$. If one instead assumes the acceleration length to be comparable the depletion length $L_{\rm pd}$, our conclusions will still basically hold, although one must also consider the relevant technological issues involved for each system – the CBA would requires multiple, differently de-tuned pumps, while for the LWFA, tapered plasma channels[15] have been proposed[18, 19].

For the purposes of energy bookkeeping, we therefore assume a CBA pump length of

[15] It is easy to see that the longitudinal phase and group velocities of an EM wave in a channel are related by $v_\phi v_g = U^2$ for some constant velocity scale U if and only if the dispersion relation is of the usual form arising from a separable solution of Helmholtz's equation, i.e., $\omega(k)^2 = U^2 k^2 + \omega_c^2$ for some constant cutoff frequency ω_c. However, for slow longitudinal variations in the transverse profile, this reciprocal relationship should approximately hold locally.

$L_{\rm pump} = 2L_{\rm det} = 2\lambda_p \left(\frac{\omega_0}{\omega_p}\right)^2$, and the total energy requirement becomes

$$U_{\rm CBA} \propto 2\sqrt{\pi}a_0^2 + 2\pi \left(\frac{\omega_0}{\omega_p}\right)^2 a_1^2. \tag{2.181}$$

The resulting CBA wakefield has the limiting formula given by (2.178) in the linear case and (2.180) for the particle-trapping regime. On the other hand, the energy requirement for a single "resonant" LWFA pulse of amplitude a_s is given by

$$U_{\rm LWFA} \propto 2\sqrt{\pi}a_s^2. \tag{2.182}$$

First, we consider the case of where CBA wake is linear, and (2.178) holds. Equating the CBA and LWFA energies (2.181) and (2.182), and applying the conditions for wake enhancement $[4a_1 > \left(\frac{\omega_p}{\omega_0}\right)^{3/2}]$ and linearity $[\omega_{\rm B}^2 = 4a_0a_1\omega_0^2 < \omega_p^2]$, one can actually show that the single-pulse LWFA generates higher gradients than a CBA with two lasers of the same total energy if $\frac{\omega_p}{\omega_0} \gtrsim 10^{-3}$. For the linear analysis to be valid, the tenuous plasmas for which CBA wins require very low-intensity lasers, so that $E_{\rm fast} < 10^{-4}\,\frac{\omega_0}{\omega_p}E_0$. As an example, for $\lambda_0 = 1\ \ \mu$m light in the linear regime, this requirement implies that if the CBA is to produce a larger plasma response than the LWFA of identical total energy, the wakefield can be at most $E_z \sim 1$ MeV/m, and cannot compete with conventional devices. For CBA parameters identical to those of the simulation in Figure 2.15, a single short pulse of identical energy would have $a_s \approx 0.8$, and a nonlinear wake approximately ~ 20 times that of the CBA.

In the particle-trapping regime, setting (2.181) equal to (2.182) and using the trapping condition $\omega_{\rm B}^2 > \omega_p^2$, one can show that for a given laser energy, the wakefields produced by the LWFA and CBA satisfy

$$E_{\rm LWFA} > \left[\frac{1}{2\sqrt{e}}\left(\sqrt{\pi}a_0 + \frac{\pi}{a_0}\right)\right] E_{\rm CBA}, \tag{2.183}$$

where here e is the base of the natural logarithm, not the magnitude of the electron charge. Since the term in square brackets in (2.183) is always greater than unity, we arrive at an even stronger conclusion than in the linear case: for a fixed total energy and acceleration length $L_a \approx L_{\rm det}$, the single-pulse LWFA always theoretically outperforms the CBA. As an example, using a single pulse with the same energy as available to the CBA in the example shown in Figure 2.16, a Gaussian pulse with peak $a_s = 1.33$ is predicted to give rise to a nonlinear wake about ~ 2.5 times that of the CBA.

To complete our discussion, we recall that given the additional constraint bounding peak *intensity* (together with total pulse duration and total energy), a single-pulse LWFA is no

longer the optimal solution. Instead, one would need to use the available laser energy to inject a train of appropriately-timed short pulses. In the linear regime ($a_s^2 \ll 1$), a sequence of N_a such pulses, separated by integer multiples of the linear plasma period, and all with the same intensity as the CBA seed, will excite the wake described in (2.179) using a total energy given by (2.182), just with a_s^2 replaced by $N_a\, a_s^2$. Because of this simple substitution, our previous comparison and conclusions in favor of the LWFA still hold essentially unchanged, but the additional technical difficulties apparent in such a carefully-timed pulse train might limit its usefulness in practice.

2.10 Conclusions and Future Directions

We have proven a number of results within simplified analytic frameworks, mainly consisting of a cold, collisionless, quasi-static 1D model with prescribed laser drive traveling at fixed group velocity, although certain results in a 2D paraxial channel were also established in the linear regime. We have corrected or clarified certain ambiguous, erroneous, or misleading claims in the literature, and provided the first definitive proofs of certain long-standing conjectures.

Although we have come to much better mutual understanding, we remain in disagreement with Spitkovsky, Chen, *et al.*[8, 9, 10, 11] over the usefulness of the transformer ratio concept in the LWFA. At bottom, the fundamental disagreement is over their claim[10] that of all the parameters that can effect issues of performance or the choice of pulse shape in the LWFA, "the only two meaningful parameters that describe laser shape from the stand point of wake excitation are the total pulse energy and the depletion length." We remain skeptical that the depletion length-scale can be of primary interest, or perhaps even of any relevance, to the LWFA in regimes where it is expected to significantly exceed other length-scales that will first limit particle acceleration, or possibly even exceed the size of the plasma itself. Accelerating electrons feel the wakefield directly, but cannot know about the depletion length under such circumstances. Despite certain the claims that optimizing the transformer ratio R_3 also results in large wake gradients, we have seen that these goals are in general not equivalent, and in actual fact complementary. Because the transformer ratio R_3 is related to the ideal energy gain only if particles can be accelerated at something close to the peak field over the entire depletion length, its relevance in other cases too is suspect. If acceleration is limited by diffraction or de-phasing or other effects to some shorter length-scale, then the best-case energy gain achievable is instead usually a monotonic function of

the peak wakefield amplitude, which should be optimized instead. Short of modeling the actual trajectories of injected particles, this would seem the best figure-of-merit to use.

2.10.1 Summary for 1D Results

Which definition of transformer ratio is used matters greatly, but the pulse shape maximizing R_3 is an impulse-plus-ramp in the linear case, but in the nonlinear case the impulse followed by concave ramp is not quite optimal, because of the nonlinear synergy in the wake excitation. In general, for fixed envelope energy, the transformer ratio can be increased only by going to longer pulse lengths, while the peak wake amplitude is increased for shorter pulses. Optimizing for wakefield amplitude and optimizing for efficiency or transformer ratio are distinct and often conflicting goals. If particle acceleration is limited by diffraction, de-tuning, or perhaps fast-growing instabilities to some distance less than the pump depletion length, which is usually the case in the LWFA, then the relevance of the transformer ratio is unclear, and overall efficiency will be limited whatever shape is chosen, so trying to achieve a large transformer ratio or high efficiency is arguably less important than simply exciting large peak gradients.

For fixed peak intensity, the pulses optimizing the peak wakefield behind the pulse consist of a train of square pulses, as predicted. The optimal pulse widths and separations are resonant in the linear case, but depend in complicated ways on how much total time is allotted for the pulse in the nonlinear case. The "greedy" algorithm involving driving whenever the laser can add energy to the wake works in the linear case but is in general sub-optimal in the nonlinear regime, but should outperform other simple-minded recipes. As pulses narrow below the plasma wavelength, they begin to look the same to the plasma, so at fixed peak intensity the performance tends not to be too seriously degraded by using smoother or less-squarish pulses of appropriate width. Performance of multiple pulses is of course sensitive to timing, but in some cases is somewhat more robust than might have been feared. Somewhat surprisingly, for single unimodal pulses in the nonlinear regime, performance seems somewhat less sensitive to precise choice of the pulse-width than in the linear regime.

For fixed envelope energy, either single impulses or sums of resonantly spaced-impulses are optimal in the linear regime within homogeneous plasma, in the sense of maximizing the peak accelerating wakefield, at least as long as dispersion may be ignored. For linear channels, single impulses are best, because of leakage. With frequency bandwidth restrictions

on the laser amplifier, in the homogeneous or channeled linear regimes, any capabilities to shape the pulse though masking or other optical manipulations should be exploited so as to minimize the removal of any energy from the pulse spectrum, but just arrange the phases to make the pulse as short as possible in time. In the nonlinear regime, sums of impulses are optimal, and for more realistic pulses of finite width, performance improves monotonically as the pulse narrows well below the resonant pule width, until dispersion becomes important. Because of the nonlinearity, two pulses are better than one, and performance in fact improves monotonically with the number of pulses (but eventually with diminishing marginal returns), as long as the pulses can be properly timed. At fixed EM energy rather than eikonal envelope energy, behavior is very similar until the pulse length becomes comparable to a few optical wavelengths, when the eikonal approximation breaks down, but also where dispersion is expected to become severe and likely limit the usefulness of the pulses anyway.

In the CBA, if the intensity of the short pulse is limited, then the addition of counter-propagating pump can definitely enhance the amplitude of the wakefield, especially in low-a regimes. However, it seems to be the case that performance is never as good as that which could be achieved in a conventional LWFA if the pump energy were instead somehow squeezed into the short pulse.

PIC codes have generally corroborated the analytic results, expect that very short intense pulses can suffer significant dispersion or distortion, and ultimately a breakdown in the applicability of the envelope approximation, for pulse lengths corresponding to a few laser carrier wavelengths or less.

2.10.2 Limitations and Extensions

Eventually, optimization frameworks balancing the goals of large wake gradients, useful wake shape, and efficient use of the driver energy, leading to particle beam of sufficiently high current and low emittance should probably be employed and evaluated, but maximization of the transformer ratio by itself does not seem to supply such a metric. Perhaps the related but distinct principle of minimizing differential deceleration will prove useful. All these question seem beyond simple analytic theory, but are well suited for numerical study.

In a 1D theory with a prescribed laser envelope obviously much physics is excluded, and no amount of analysis or or speculation based on these simplifying assumptions can be reliably applied when the ignored effects become important. For a sufficiently intense pulse

or sufficiently effective interaction, these approximations almost by definition must break down.

The use of intense short pulses implies that both linear group velocity dispersion and changes to the nonlinear refractive index due to ponderomotive blow-out or other effects may become significant. To definitively address questions about optimal excitation, efficiency, or energy exchange between laser and plasma, or between plasma and the accelerated beam, one must ultimately incorporate self-consistent feedback of the plasma on the laser pulse, and perhaps self-consistent beam loading of the beam in the plasma wave as well.

We have begun to verify and supplement the analytic results using 1D PIC simulations that in principle include certain dispersion, depletion, and distortion effects, as well as certain 1D instabilities like RBS, but obviously not diffraction or $2D$ instabilities.

To incorporate important transverse dynamics, including the effects of channel guiding or nonlinear self-guiding, other aspects of nonlinear or non-paraxial propagation, including filamentation, hosing, or other instabilities that do not appear in 1D geometry, numerical simulations should be extended to at least 2D or even 3D.

If Landau damping and particle self-trapping effects can be neglected, relativistic fluid simulations specialized to the case ultra-short laser propagation in underdense plasmas promise to be far more efficient than general PIC-based simulations. However, at realistic electron temperatures, Landau damping may not be negligible, especially if multi-pulse trains are of interest, because the Langmuir wave can decay or decohere between sub-pulses, so at least some kinetic effects may need to be modeled. Vlasov simulations are possible in 1D but not yet computationally practical for fully $2D$ geometries (or $4D$ phase spaces).

In such a setting, further optimization over laser spot size or transverse profile more generally, plasma density or channel profile, and other parameters can be envisioned. it is not clear how nonlinear and transverse aspects of the laser-plasma interaction and nontrivial back-reaction on and evolution of the laser may affect our conclusions If short pulses in a channel can beat linear dispersion and hold together long enough (due to nonlinear effects) to deliver most of their energy to the wake, then they may continue to out-perform other pulse-shapes in the strongly nonlinear regimes. Otherwise interest may continue to shift to the SM-LWFA regime, where certain instabilities are harnessed rather than avoided. While longitudinal pulse-shaping in the case of the nonlinear SM-LWFA has been introduced, questions of optimality remain largely unanswered, or even unasked.

Increasingly, questions of optimizing not just wakefield amplitude but aspects of the electron injection process, or of the actual energy gain and energy spread, are becoming of experimental concern and theoretical interest. In all these cases, the demands of multi-dimensional numerical simulation suggest that, rather than exploring large swaths of parameter space in search of better solutions, we should focus simulations more directly on optimizing specific schemes being tried in practice, or on a few particularly promising directions guided by insights from new theory.

Acknowledgements

This research was performed in collaboration with R. R. Lindberg, and benefited greatly from many interactions with B. A. Shadwick while he worked as a post-doctoral fellow for Professor Wurtele at LBNL. We also thank A. Spitkovsky, formerly at U.C. Berkeley, for many useful conversations, insightful challenges, and forceful defenses of his own points-of-view, and E. Esarey of L'OASIS group at LBNL, for many informative discussions.

Chapter 3

Robust Autoresonant Excitation in the Plasma Beat-wave Accelerator

...no resonance goes completely unheard....

John Gardner

The greatest stories are those that resonate our beginnings and intuit our endings, ... and dissolve them both into one.

Ben Okri

3.1 Introduction and Overview

The Plasma Beat-Wave Accelerator (PBWA) was first proposed by Tajima and Dawson (TD) [34] as an alternative to the short-pulse Laser Wake-Field Accelerator (LWFA), based on earlier analysis of beat-wave excitation as a plasma probe [35] or as a mechanism for plasma heating [35, 36, 37]. Subsequently the PBWA concept has been studied extensively theoretically, numerically, and experimentally [38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51]. (For a review, see [52].) In the original scheme, two laser fields co-propagating in an underdense plasma are detuned from each other by a frequency shift close to the electron plasma frequency, so that the modulated envelope resulting from the beating between the two laser fields can act ponderomotively on the plasma electrons to resonantly excite a large-amplitude, high-phase-velocity plasma wave suitable for particle acceleration.

For a fixed beat frequency, performance of the PBWA is constrained by what is now

known as the Rosenbluth-Liu (RL) limit, after the pioneering study in [37]. As the plasma wave amplitude grows, relativistic detuning effects eventually prevent further growth of the peak longitudinal electric field E_z beyond a maximum value $E_{\rm RL}$, which, for any realistic laser and plasma parameters, lies below the cold, non-relativistic, one-dimensional (1D), wave-breaking limit E_0 [53, 54]:

$$|E_z| \leq E_{\rm RL} \equiv E_0 \left(\tfrac{16}{3} \tfrac{\omega_p^2}{\omega_1 \omega_2} \tfrac{|E_1||E_2|}{E_0^2} \right)^{1/3} < E_0 \equiv \frac{mc\omega_p}{e}, \tag{3.1}$$

where c is the speed of light, m is the mass of the electron, e is the magnitude of its electric charge; ω_p is the cold, linear electron plasma frequency, defined in Gaussian units as

$$\omega_p = \left(\tfrac{4\pi n_0 e^2}{m} \right)^{1/2}, \tag{3.2}$$

where n_0 is the ambient electron number density; E_1 and E_2 are the peak electric fields of the beating drive lasers; and ω_1 and ω_2 are their respective carrier frequencies. For the plasma waves of interest here, with high but sub-luminal phase velocities $v_p \lesssim c$, the cold non-relativistic limit itself is smaller than the cold, relativistic (or nonlinear) wave-breaking limit [53, 52]:

$$E_0 < E_{\rm WB} = \left[2(\gamma_p - 1) \right]^{1/2} E_0, \tag{3.3}$$

which sets an upper bound on the amplitude of coherent plasma oscillations useful for particle acceleration in a cold plasma. Here

$$\gamma_p = \left[1 - \tfrac{v_p^2}{c^2} \right]^{-1/2} \tag{3.4}$$

is the relativistic kinematic factor associated with the phase-velocity v_p of the excited plasma wave, approximately equal to the characteristic group velocity $\bar{v}_g$ (to be precisely defined below) associated with the envelope of the beating lasers, both of which will be nearly equal to the speed of light c for sufficiently underdense plasmas. (Large thermal spreads increase particle trapping and lower this threshold for wave-breaking, as discussed in [55, 56, 57].)

It is not difficult to see the origin of this detuning effect. The effective nonlinear plasma frequency, in the presence of the laser drive and a plasma wave, is approximately given by

$$\omega_{p_{\rm NL}} \approx \frac{\left(1 + \overline{\delta n}/n_0\right)}{\sqrt{\gamma_{\rm RMS}}} \omega_p, \tag{3.5}$$

where $\overline{\delta n}$ is an average density perturbation associated with the plasma wave, and $\gamma_{\rm RMS}$ is the root-mean-square relativistic factor of the electrons, including effects of both the transverse quiver in laser fields and the longitudinal velocity imparted by the Langmuir wave itself. Except for highly nonlinear situations where ponderomotive blow-out becomes important,

the average density perturbation is typically small, and will be neglected here. The RMS transverse quiver velocity is roughly constant for the fixed-amplitude driving lasers typical of PBWA, but as the longitudinal motion of the perturbed electrons associated with the excited Langmuir wave becomes even weakly relativistic, $\gamma_{\rm RMS}$ will begin to increase, causing $\omega_{p_{\rm NL}}$ to decrease. Because of the sensitivity of resonance phenomena, $\gamma_{\rm RMS}$ need not grow much before this resonance shift severely limits the efficacy of the driving beat-wave. As this dynamical process is effectively reversible, if the detuning occurs before wave-breaking or other phase-mixing or dissipative effects, then the ponderomotively excited Langmuir wave is not actually saturated *per se* at some fixed value given by the RL limit, but exhibits a slow (compared to the plasma frequency) nonlinear modulation or beating of the wave amplitude, periodically peaking near the maximum value predicted by RL and then decreasing, as energy is exchanged between the Langmuir wave and the laser fields.

Tang, Sprangle, and Sudan (TSS) [40, 41] pointed out that one can, in principle, achieve somewhat higher amplitudes than that predicted by the RL limit by simply detuning the beat frequency to a value somewhere below the linear plasma frequency, matching the drive frequency to the nonlinear plasma frequency for some non-trivial plasma wave amplitude. However, McKinstrie and Forslund [58] noted fundamental practical difficulties with this approach. For downward detunings at or beyond some small critical frequency shift, multiple solutions to the envelope equations appear, and the high-amplitude branch cannot be accessed reliably from the initially quiescent plasma state, especially in the presence of any small amount of damping. For sufficiently small detunings, moderate improvement over the RL limit is possible, but typically with peak amplitudes still lying below even the cold linear wave-breaking limit. Furthermore, as the detuning grows in magnitude toward this critical value, the excitation becomes increasingly non-monotonic, and the time needed to reach the peak amplitude increases and becomes comparable to the time-scales for ion modulational or decay instabilities that degrade the coherence of the plasma wave together with its utility for accelerator applications.

Matte *et al.* (MMEBP) [59] suggested the use of a plasma with a time-varying density, carrying out the beat-wave excitation during the actual ionization process, when the free electron density is growing. The increase in plasma density can then compensate for the increase in $\gamma_{\rm RMS}$, and thus the effective plasma resonance may be maintained with a fixed driving beat frequency for a longer time. However, in order to yield appreciable improvement, such an approach would require impractically precise timing of lasers and some means to control the ionization rate, matching it, at least roughly, to the expected growth rate of

the plasma wave. The usual method involving laser-induced field ionization leads to cross sections which are exponentially sensitive to the laser intensity and is therefore unsuitable for such an approach.

Better still, Deutsch, Meerson, and Golub (DMG) [60] suggested incorporating compensatory time-dependence into the drive lasers rather than in the plasma density. They proposed to partially overcome the relativistic detuning effect by using a chirped beat-wave excitation scheme, where the frequencies of one or more of the laser pulses are chirped downward starting from the linear resonance so as to compensate for the change in the nonlinear frequency of the growing plasma wave, ideally allowing amplitudes approaching the nonlinear wave-breaking limit. The nonlinear modulations are thereby reduced but not eliminated, consisting of a ringing about some non-zero saturation value rather than a full beating as in the original (RL/TD) scheme.

Chirping too fast will cause the plasma oscillations to detune before they have a chance to grow much, while chirping too slowly appears to result in a slow, highly non-monotonic (ringing) excitation, where the long interaction times required to build up sufficient amplitude (on the order of a few ion plasma periods) can allow instabilities of the ionic background to grow and disrupt the coherence of the plasma wave [61, 46]. Intuitively, one expects for the DMG approach that there exists an optimal intermediate chirp rate, or in fact even an entire optimal chirp profile, in which the driving beat frequency starts at the linear plasma resonance, where the initial coupling is strongest, and then self-consistently matches the instantaneous frequency to the detuning caused by the expected growth of the plasma wave.

However, solving such a complicated nonlinear control problem so as to exactly match the nonlinear plasma frequency with the laser chirp, with the help of either numerical simulation or experimental calibration, would be both computationally daunting and experimentally impractical. The linear plasma frequency will be imperfectly known in practice, and even if it were obtained, the instantaneous value of the nonlinear plasma frequency actually depends on the initial absolute phase of the plasma wave, which is determined by the exact initial conditions of the plasma and the initial beat phase of the driving lasers, which are not normally controlled. However, DMG argue that such *extrinsic* frequency matching is not actually necessary, and that entrainment can be achieved without recourse to carefully-tailored pulses or to external feedback. They invoke the nonlinear dynamical phase-locking phenomenon known as autoresonance [62, 63], arguing that as long as the chirp is sufficiently slow (compared to the time-scales for the nonlinear growth and modulation described by

RL) and in the right direction (namely, downward in frequency), then the externally forced, nonlinear dynamical system can self-adjust and automatically maintain approximately the desired phase entrainment between the driven wave and the ponderomotive drive, resulting in an average increase in the oscillation amplitude.

However, proposed as it was before the understanding of autoresonance had matured, especially with regard to the threshold behavior for establishing and maintaining entrainment and the importance of the initial detuning, the DMG scheme sometimes fails to produce appreciable phase-locking. This is largely due to its prescription of beginning the chirp on resonance, and becomes particularly problematic if the plasma density is imperfectly known or subject to shot-to-shot jitter, when it can lead to nonuniform growth or early saturation. In fact, there appears to be some uncertainty in the original description of the scheme [60] as to exactly how autoresonant the proposed excitation mechanism is, as to whether or to what extent "the precise form of the time-dependence of the frequency is inessential" or else there "exists an optimal chirp rate that provides the highest excitation rate."

In fact the DMG chirped scheme can often achieve higher longitudinal electric fields than the original (RL/TD) approach, but does not do so universally or robustly, and cannot always produce fields approaching even the linear wave-breaking limit for experimentally accessible parameter values. The peak amplitude and spatio-temporal uniformity of the final plasma wave may be sensitive to the exact chirp rate, to the initial detuning, and most especially to the precise value of the plasma frequency, which, once again, often is imperfectly known or subject to significant variation, due to limited measurement precision, to shot-to-shot jitter, or to single-shot density fluctuations.

Informed by more recent results, we propose a novel variant of the chirped plasma PBWA concept which also exploits autoresonance, but enhanced via Adiabatic Passage Through Resonance (APTR) [64, 65, 66]. A brief summary of our scheme can be found in [67]. Rather than chirping downward from the linear resonance, we actually start, counter to naive intuition, with a frequency shift well above the resonance, and then slowly sweep the beat frequency through and below resonance. With this approach, when starting from a quiescent plasma, the final state of the plasma wave is insensitive to the exact chirp history, and the excitation is more robust with respect to imprecise characterization of the plasma density, or actual shot-to-shot fluctuation of the density or single-shot variation in it, provided only that the spatio-temporal scales of variability are long compared to the frequency and wavenumber of the plasma wave, and the relative range of fluctuation or uncertainty is not too large. The chirp rate need not be matched *a priori* to any anticipated

rate of growth in the plasma wave or resulting rate of relativistic detuning, but must only be chosen to be sufficiently slow so as to satisfy a certain adiabatic trapping condition and thereby ensure phase-locking. Because of the nonlinear nature of the ponderomotive forcing in the relativistic regime, this adiabaticity condition becomes increasingly difficult to maintain as the wave amplitude grows, and for a constant chirp rate eventually the plasma wave would fall out of phase-locking with the drive. This leads to a true saturation in amplitude, with significantly less of the undesirable ringing seen in the original RL/TD scheme. Before the onset of saturation, the frequency of the excited plasma wave can be closely entrained to the instantaneous beat frequency of the drivers, and the overall phase can remain reasonably entrained as well, while the plasma wave amplitude monotonically grows to automatically and self-consistently adjust itself to the monotonically decreasing beat frequency.

Not only is our excitation scheme more robust, so too is our method of analysis. The Lagrangian fluid formalism developed by RL and then used by DMG employs a power series expansion which is valid only for weakly relativistic motion in both the transverse and longitudinal directions, thereby limiting its domain of applicability to laser fields of sufficiently low intensity (implying non-relativistic quiver motion), and to excited plasma waves of sufficiently low amplitude (with electric fields well below wave-breaking). Therefore their equations of motion become increasingly untrustworthy precisely in the regime of interest, where the plasma wave amplitude grows to some appreciable fraction of the wave-breaking value E_0. Here, we instead employ a fully nonlinear, Eulerian fluid model that allows for arbitrarily relativistic electron motion below the nonlinear wave-breaking limit E_{WB}, similar to that discussed in [68, 52] for general laser-plasma interactions and in [44] for PBWA investigations, and also to a pioneering treatment in [53]. In the weakly relativistic limit, this Eulerian approach can be shown to agree exactly with the Lagrangian treatment of RL.

This analytical fluid model, and the various physical assumptions which go into it, are discussed in Section 3.2. In order to treat the autoresonant nature of the problem more easily and transparently, we re-formulate the dynamical equations in a fully Hamiltonian form in Section 3.3. In Section 3.4, we use this canonical formalism to analyze the autoresonant aspects of the beat-wave excitation, from the initial linear phase-locking regime through the weakly and strongly nonlinear trapped phases to non-adiabatic saturation. In Section 3.5, we discuss features of some realistic examples relevant for possible experimental implementation and investigation. Section 3.6 summarizes our preliminary assessment of the

merits and limits of our autoresonant PBWA, especially in comparison to other beat-wave approaches. We then offer brief conclusions from our initial investigation and prospects for future study[1] in Section 3.7.

3.2 Fundamental Equations

Our study of wake excitation is based on an approximate, but convenient and widely-used model. The plasma is treated as a cold, collisionless, fully relativistic electron fluid moving in a stationary, neutralizing, ionic background, coupled to electromagnetic fields governed by Maxwell's equations. We restrict our analysis to one-dimensional (1D) geometry, where all dynamical quantities depend only on the longitudinal position z and the time t, and we assume a completely homogeneous and initially quiescent background state of the plasma. The plasma is assumed to be highly underdense, i.e., $\omega_p \ll \omega_1, \omega_2$. In addition, we assume prescribed laser fields, neglecting throughout the entire interaction any changes to the laser envelopes due to linear effects such as diffraction or group-velocity dispersion, or any nonlinear back-action of the plasma on the lasers such as depletion, self-focusing, photon acceleration due to ionization or density variation, as well as Raman scattering, self-modulation, and other instabilities [69, 70]. This model, although simplified, nevertheless reveals the essential features of autoresonance and its potential advantages for the PBWA. Possible extensions to more realistic models, as well as some arguments for the validity of our general results beyond the strict applicability of this model, especially in the light of density fluctuations, are discussed in Sec. 3.6 and Sec. 3.7.

The cold, collisionless fluid treatment assumes that the electron temperature is sufficiently small so that: the thermal energy $k_{\mathrm{B}}T_e$ is negligible compared to the typical kinetic energy associated with the transverse quiver motion in the driving laser fields; thermal corrections to the Langmuir dispersion relation are small, which, for waves with relativistic phase velocities simply requires that $k_{\mathrm{B}}T_e$ is much smaller than the electron rest energy mc^2; and collisional damping of the laser fields is small over the interaction time, as are Landau and collisional damping of the excited plasma wave. Neglect of ion dynamics is strictly valid provided the time-scale for ion motion, typically of the order of a few times

[1] Because of the highly collaborative nature of the work, some of this research is expected to be reviewed in the thesis of R.R. Lindberg. In addition, we expect that he will report subsequent investigations where he compared the results of the simplified analytical model developed here to numerical simulations from one-dimensional PIC and nonlinear fluid codes.

ω_i^{-1}, where

$$\omega_i = \left(\frac{4\pi n_0 Z_i^2 e^2}{M_i}\right)^{1/2} \tag{3.6}$$

is the ion plasma frequency for ions of mass M_i and charge $+Z_i e$, remains longer than the duration T of the lasers: i.e., $\omega_i T \lesssim 2\pi$.

Within this model, the continuity, momentum, and Poisson equations, respectively, are, in Gaussian units:

$$\frac{\partial n_e}{\partial t} + \boldsymbol{\nabla} \cdot (n_e \boldsymbol{v}) = \quad 0, \tag{3.7a}$$

$$\frac{\partial \boldsymbol{p}}{\partial t} + (\boldsymbol{v} \cdot \boldsymbol{\nabla})\boldsymbol{p} = \quad e\left(\boldsymbol{\nabla}\Phi - \frac{\boldsymbol{v}}{c} \times (\boldsymbol{\nabla} \times \boldsymbol{A})\right), \tag{3.7b}$$

$$\nabla^2 \Phi = -4\pi e(n_0 - n_e), \tag{3.7c}$$

where $\boldsymbol{A}$ is the vector potential and Φ the scalar potential in the Coulomb gauge; n_e is the electron number density and n_0 the background ion number density, assumed to be equal in the absence of perturbations; $\boldsymbol{v}$ is the electron velocity; and $\boldsymbol{p} = m\gamma\boldsymbol{v}$ is the kinetic momentum, where

$$\gamma = \left[1 - \frac{|\boldsymbol{v}|^2}{c^2}\right]^{-1/2} \tag{3.8}$$

is the relativistic factor associated with electron motion.

Neglect of diffraction or other transverse effects in the laser fields requires that characteristic laser spot size w_0 sufficiently exceeds both the characteristic laser wavelength λ and the transverse length-scale $\Delta r_\perp$ for variations in the plasma, while $\Delta r_\perp$ itself should greatly exceed the Langmuir wavelength $\lambda_p \sim c/\omega_p$ to ensure the validity of our assumption of a 1D, homogeneous medium. Neglect of the nonlinear evolution of the lasers effectively imposes certain constraints on the duration and intensity of the pulses, as summarized in [52]. In the Coulomb gauge, the vector potential $\boldsymbol{A}$ is solenoidal (i.e., divergenceless), which in 1D also implies that it is geometrically transverse, so $\boldsymbol{A} = \boldsymbol{A}_\perp(z,t)$ will be polarized perpendicular to the propagation direction $+\hat{\boldsymbol{z}}$. For simplicity each drive laser will be assumed to possess right-handed circular polarization, and to have a flat-topped profile of total duration T, where the leading edge of each laser enters the left edge of the plasma at $z = 0$ at the initial time $t = 0$.

Defining a scaled, or dimensionless, time $\tau \equiv \omega_p t$, co-moving position $\xi \equiv \omega_p(t - z/\bar{v}_g)$, longitudinal velocity $\beta \equiv v_z/c$, Langmuir phase velocity $\beta_p \equiv v_p/c = \bar{v}_g/c$, vector potential $\boldsymbol{a} = \boldsymbol{a}_\perp \equiv \frac{e}{mc^2}\boldsymbol{A}$, electrostatic potential $\phi \equiv \frac{e}{mc^2}\Phi$, and number density perturbation $\rho \equiv (n_e - n_0)/n_0$, the continuity equation (3.7a) can be written in scaled variables as

$$\frac{\partial}{\partial\tau}\rho + \frac{\partial}{\partial\xi}\left[(1 - \beta/\beta_p)(1 + \rho)\right] = 0, \tag{3.9}$$

and Poisson's equation (3.7c) becomes

$$\frac{\partial^2}{\partial \xi^2}\phi = \beta_p^2\,\rho. \tag{3.10}$$

Assuming an initially quiescent plasma in the initial absence of laser fields (i.e., η, ϕ, β, $\boldsymbol{a} \to 0$ as $\tau \to -\infty$ for fixed ξ or as $\xi \to -\infty$ for fixed τ), the conservation of transverse canonical momentum implies that

$$\boldsymbol{p}_\perp = mc\,\boldsymbol{a}_\perp = mc\,\boldsymbol{a}. \tag{3.11}$$

The relativistic factor γ then factors into transverse and longitudinal contributions as

$$\gamma \equiv \gamma_\perp \gamma_\parallel = \sqrt{1+|\boldsymbol{a}|^2}\,\frac{1}{\sqrt{1-\beta^2}}. \tag{3.12}$$

Using these relations, we eliminate any explicit appearance of $\boldsymbol{v}_\perp$ or $\boldsymbol{p}_\perp$ in the longitudinal component of the momentum equation (3.7b), thereby obtaining

$$\frac{\partial}{\partial \tau}(\gamma\beta) + (1-\beta/\beta_p)\,\frac{\partial}{\partial \xi}(\gamma\beta) = -\frac{1}{\beta_p}\frac{\partial}{\partial \xi}\phi + \frac{1}{2}\frac{1}{\beta_p \gamma}\frac{\partial}{\partial \xi}\,|\boldsymbol{a}|^2\,. \tag{3.13}$$

Now, we further make the Quasi-Static Approximation (QSA) (see, e.g., [44] and [71]), wherein we assume $\rho = \rho(\xi)$ and $\beta = \beta(\xi)$, i.e., that the plasma response is independent of time τ in the co-moving frame, implying that the plasma wave itself moves without dispersion at the fixed average group velocity β_p of the driving lasers. The QSA also specifically requires that any distortion or change in the shape of the laser envelope remains negligible during the typical interaction time with any one transverse slice of plasma, which will be $O(T)$. This constraint is satisfied automatically in our model, since we have already assumed that laser envelope evolution is negligible throughout the entire interaction, which is of duration $O(T + L/c)$, where L is the total plasma length. In the QSA, all τ derivatives in equations describing the plasma response can then be neglected, and the continuity equation (3.9) can be immediately integrated, yielding, for our quiescent initial conditions,

$$(1-\beta/\beta_p)\,(1+\rho) = 1. \tag{3.14}$$

Expressing $|\boldsymbol{a}|^2$ in terms of γ and β, the longitudinal momentum equation (3.13) can also be integrated to obtain:

$$\gamma\,(1-\beta_p\beta) - \phi = 1. \tag{3.15}$$

After some algebra, the Poisson equation (3.10) then becomes

$$\frac{\partial^2}{\partial \xi^2}\phi = \beta_p^2\gamma_p^2\left[\beta_p\left(1 - \frac{1+|\boldsymbol{a}|^2}{\gamma_p^2(1+\phi)^2}\right)^{-\frac{1}{2}} - 1\right]. \tag{3.16}$$

This single, second-order equation describing the nonlinear plasma response is valid, within the QSA, for arbitrary relativistic electron velocities provided the Langmuir wave amplitude remains below the nonlinear wave-breaking limit determined by:

$$\phi > -1 + \frac{1}{\gamma_p}\sqrt{1+|\boldsymbol{a}|^2} = -1 + \frac{\gamma_\perp}{\gamma_p}. \tag{3.17}$$

In the high phase-velocity limit appropriate to highly underdense plasmas, where $0 < 1-\beta_p \ll 1$ and $\gamma_p \gg 1$, the dynamical equation (3.16) may be Taylor expanded and further simplified to

$$\frac{\partial^2}{\partial\xi^2}\phi = \frac{1}{2}\left[\frac{1+a^2}{(1+\phi)^2} - 1\right]. \tag{3.18}$$

Because we have effectively taken $\gamma_p \to \infty$, this equation remains mathematically well-defined for any $\phi > -1$ and any $|\boldsymbol{a}|^2 \geq 0$, but the actual bound (3.17) remains a more physically trustworthy limit for wave-breaking. After solving for $\phi(\xi)$, we can determine the electric field using

$$\frac{\partial}{\partial\xi}\phi = \beta_p \varepsilon_\parallel, \tag{3.19}$$

where $\varepsilon_\parallel = \varepsilon_\parallel(\xi) \equiv E_z/E_0$ is the scaled longitudinal electric field within the QSA. The density perturbation $\rho(\xi)$ may then be determined using the scaled Poisson's equation (3.10), the velocity $\beta(\xi)$ using the first integral (3.14), and finally $\gamma(\xi)$ by using (3.12).

Under our assumptions, the normalized vector potential

$$\boldsymbol{a} = \boldsymbol{a}(\xi,\tau) = \tfrac{e}{mc^2}\boldsymbol{A}\big(z = (\tau-\xi)\bar{v}_g/\omega_p, t = \tau/\omega_p\big) \tag{3.20}$$

is taken to be a prescribed function of time and position, describing the unperturbed laser fields of duration T throughout the plasma of length L, with a modulated envelope which is slowly chirped but travels along at a fixed group velocity $\bar{v}_g$. For otherwise arbitrary laser fields, $|\boldsymbol{a}|^2$ will still have both fast (carrier frequency) and slow (beat frequency) dependence on both τ and ξ, although considerable simplification is possible for the assumed PBWA form, namely: two weakly detuned, slowly-chirped, co-propagating, nearly-plane-wave, flat-topped lasers with positive helicity. The normalized vector potential $\boldsymbol{a}$, representing the beating drive lasers may then be explicitly written as

$$\boldsymbol{a} = \boldsymbol{a}_1 + \boldsymbol{a}_2 = \tfrac{1}{2}\left[\hat{\boldsymbol{e}}_+ a_1 e^{i\psi_1} + c.c.\right] + \tfrac{1}{2}\left[\hat{\boldsymbol{e}}_+ a_2 e^{i\psi_2} + c.c.\right], \tag{3.21}$$

where $\hat{\boldsymbol{e}}_+ = \frac{1}{\sqrt{2}}\left(\hat{\boldsymbol{x}} + i\hat{\boldsymbol{y}}\right)$.

At the leading edge ($z = 0$) of the plasma, the laser phases ψ_j ($j = 1, 2$) are given, for $t \geq 0$, by

$$\psi_j(z=0,t) = \psi_{j_0} - \int_0^t dt'\,\omega_j(t')\,, \tag{3.22}$$

where the ψ_{j_0} are real constants depending on initial laser conditions, and the instantaneous carrier frequencies $\omega_j(t) \equiv \omega_j(z=0,t) \equiv -\frac{d}{dt}\psi_j(z=0,t)$ allow for slow chirping of one or both of the lasers, such that

$$|\omega_j^{-1}\tfrac{d}{dt}\omega_j| \ll \omega_p \ll \omega_j \tag{3.23}$$

and

$$|\omega_1 - \bar{\omega}_1| \lesssim |\omega_2 - \bar{\omega}_2| \lesssim |\omega_1 - \omega_2| \sim |\bar{\omega}_1 - \bar{\omega}_2| \sim \omega_p, \tag{3.24}$$

where we define the average carrier frequencies

$$\bar{\omega}_j = \frac{1}{T}\int_0^T dt'\,\omega_j(t') = \big[\psi_j(z=0,t=0) - \psi_j(z=0,t=T)\big]/T, \tag{3.25}$$

and the overall average carrier frequency is then defined as

$$\bar{\omega} \equiv \tfrac{1}{2}\left(\bar{\omega}_1 + \bar{\omega}_2\right). \tag{3.26}$$

Although taken as piecewise constant here, the amplitudes a_j, $j = 1, 2$ could more generally also include suitably slow (i.e., slower than the plasma frequency, but possibly comparable to the chirp rate) time and position dependence as well, to model more realistic ramping of the drive fields and some 2D effects.

Within the plasma ($0 \le z \le L$), we assume that the local laser frequencies $\omega_j(z,t) = -\frac{\partial}{\partial t}\psi_j$ and local wavenumbers $k_j(x,t) = \frac{\partial}{\partial z}\psi_j > 0$ each satisfy the 1D electromagnetic dispersion relation

$$\omega^2 = \frac{\omega_p^2}{\gamma_0} + c^2k^2, \tag{3.27}$$

where the constant $\gamma_0 \ge 1$ parameterizes an average global nonlinear shift in the effective plasma frequency due to transverse electron quiver. For electromagnetic waves satisfying this dispersion relation, the group velocity v_g is given by

$$v_g \equiv v_g(\omega) \equiv \tfrac{d}{dk}\omega(k) = c^2\tfrac{k(\omega)}{\omega} = c\left[1 - \tfrac{\omega_p^2}{\gamma_0\omega^2}\right]^{1/2}, \tag{3.28}$$

which in our underdense case may be approximated as

$$v_g(\omega) \approx c\left[1 - \tfrac{1}{2}\frac{\omega_p^2}{\gamma_0\omega^2} + \ldots\right]. \tag{3.29}$$

We take as our reference group velocity $\bar{v}_g$, the expression (3.28) evaluated at the average carrier frequency $\bar{\omega}$:

$$\bar{v}_g \equiv c\beta_p \equiv v_g(\bar{\omega}) \approx c\left[1 - \tfrac{1}{2}\frac{\omega_p^2}{\gamma_0\bar{\omega}^2}\right] \approx \frac{\bar{\omega}_2 - \bar{\omega}_1}{k(\bar{\omega}_2) - k(\bar{\omega}_1)}, \tag{3.30}$$

which represents the characteristic velocity at which both laser envelope modulations and Langmuir phase-fronts travel.

Because the relevant time-scales satisfy $\omega_j^{-1} \ll \omega_p^{-1} \ll T$, the fast carrier oscillations in $|\boldsymbol{a}|^2$ at the harmonics $2\,\omega_1$ and $2\,\omega_2$ and the sum frequency $\omega_1 + \omega_2$ will average away, leaving only the slowly-varying contribution to $|\boldsymbol{a}|^2$:

$$a^2 = \langle |\boldsymbol{a}|^2 \rangle = \tfrac{1}{2}\left[|a_1|^2 + |a_2|^2 + a_1^* a_2 e^{i(\psi_2 - \psi_1)} + a_1 a_2^* e^{-i(\psi_2 - \psi_1)} \right]. \tag{3.31}$$

Since the group-velocity dispersion effects ($\frac{d^2}{dk^2}\omega$ and higher-order terms) remain small in the underdense regime, linear propagation into the plasma results in a beat phase given by

$$\psi_2(z,t) - \psi_1(z,t) = \Delta\psi_0 + \int_0^{t - z/\bar{v}_g} dt' \left[\omega_2(t') - \omega_1(t') \right] + O\left(\tfrac{1}{\gamma_0} \tfrac{\omega_p^3}{\bar{\omega}^3} \tfrac{\omega_p L}{c} \right), \tag{3.32}$$

where $\Delta\psi_0$ is just a constant, equal to the difference of the initial laser phases. The neglected terms limit the validity of the constant group velocity approximation to interactions lengths L less than the so-called dispersion length L_{disp} :

$$L \leq L_{\mathrm{disp}} \sim \frac{\bar{\omega}^3}{\omega_p^3} \frac{c}{\omega_p}. \tag{3.33}$$

For accelerator applications, the useful interaction length is already limited by the dephasing length L_{d}, beyond which accelerated electrons cannot gain energy from the electrostatic field:

$$L \leq L_{\mathrm{d}} \sim \frac{\bar{\omega}^2}{\omega_p^2} \frac{c}{\omega_p} \sim \frac{\omega_p}{\bar{\omega}} L_{\mathrm{disp}} \ll L_{\mathrm{disp}}, \tag{3.34}$$

so that our constant group velocity approximation imposes no further restriction on the interaction length.

By appropriate choice of the initial phases, we may take both a_1 and a_2 to be real and nonnegative. Defining the co-moving normalized beat frequency

$$\Delta\omega(\xi) = \left[\omega_2(t = \xi/\omega_p) - \omega_1(t = \xi/\omega_p) \right] / \omega_p, \tag{3.35}$$

the beat phase

$$\psi(\xi) = \psi(0) + \int_0^{\xi} d\xi' \Delta\omega(\xi'), \tag{3.36}$$

for $\psi(0) \equiv \Delta\psi_0$, the normalized beat amplitude

$$\epsilon = a_1 a_2, \tag{3.37}$$

the average normalized intensity per laser

$$\bar{a}^2 = \tfrac{1}{2}\left[a_1^2 + a_2^2\right], \tag{3.38}$$

and the electron quiver factor

$$\gamma_0 = \sqrt{1+\bar{a}^2}, \tag{3.39}$$

the reference group velocity $\bar{v}_g$ is fully determined, and the slow part of the normalized ponderomotive drive may be written as a function of ξ only:

$$a^2 = a^2(\xi) \approx \Theta(\xi)\,\Theta(\omega_p T - \xi)\left[\bar{a}^2 + \epsilon\cos\psi(\xi)\right], \tag{3.40}$$

where $\Theta(\xi)$ is the usual Heaviside step function. Using this form for the ponderomotive drive, the equation of motion (3.18) may be written as an ordinary differential equation in the co-moving coordinate ξ :

$$\tfrac{d^2}{d\xi^2}\phi = \phi''(\xi) = \frac{1}{2}\left[\frac{1+\bar{a}^2+\epsilon\cos\psi(\xi)}{(1+\phi)^2} - 1\right], \tag{3.41}$$

with the initial conditions $\phi(\xi=0) = \phi'(\xi=0) = 0$, valid for $\tau > 0$ and $\tau - \omega_p L/\bar{v}_g \le \xi \le \tau$, while otherwise $\phi(\xi,\tau) \equiv 0$. Equation (3.41) is used for all numerical simulations discussed subsequently, and is the starting point for our analysis of autoresonance.

3.3 Hamiltonian Formalism

To study autoresonance, we now develop the Hamiltonian formulation of (3.41). Our goal is an expression in terms of canonical action-angle variables, for which the phase-locking phenomenon is most readily analyzed. First, note that the dynamical equation for the electrostatic potential (3.41) can be derived from the Hamiltonian

$$\begin{aligned}\mathcal{H}(\phi,p;\xi) &= \frac{1}{2}p^2 + \frac{1}{2}\left[\frac{1}{1+\phi} + \phi - 1\right] + \frac{\bar{a}^2+\epsilon\cos\psi(\xi)}{2(1+\phi)} \\ &\equiv \mathcal{H}_0(\phi,p) + \frac{\bar{a}^2+\epsilon\cos\psi(\xi)}{2(1+\phi)},\end{aligned} \tag{3.42}$$

with the scalar potential ϕ regarded as the generalized coordinate, $p \equiv \phi' \equiv \frac{d}{d\xi}\phi = \beta_p\varepsilon_\parallel$ regarded as the canonical momentum conjugate to ϕ, and ξ taken as the time-like evolution variable. In this way, the plasma-wave dynamics are seen to be analogous to those of a one-dimensional forced nonlinear oscillator. The component $\mathcal{H}_0$ of (3.42) represents the Hamiltonian of the free oscillator, involving one term $\mathcal{T}(p) = \frac{1}{2}p^2$ analogous to the kinetic energy of the oscillator (which is in fact proportional to the electrostatic *potential* energy

density of the plasma wave), and another term corresponding to an anharmonic effective potential $\mathcal{V}(\phi) = \frac{1}{2}\left[1/(1+\phi) + \phi - 1\right]$. The remaining driving term $\Delta\mathcal{H} \equiv \mathcal{H}(\phi, p; \xi) - \mathcal{H}_0(\phi, p)$ corresponds to the time-dependent forcing of the oscillator. This forcing is external in that it depends on a prescribed function of ξ, but because of the nature of the relativistic nonlinearity, the effective strength of this forcing depends on both the intensity of the driving lasers and the instantaneous value of the dynamical variable ϕ associated with the excitation. Because of the ponderomotive nature of the coupling, the effective forcing is always positive, i.e., includes a constant as well as a purely oscillatory part. Both of these features differ somewhat from previously-studied autoresonant systems, and will have important implications for the dynamics.

In the absence of forcing (i.e., $\epsilon = \bar{a}^2 = 0$), the dynamics governed by $\mathcal{H}_0(\phi, p) = \mathcal{T}(p) + \mathcal{V}(\phi)$ are conservative, so the oscillator energy, i.e., the value H of the Hamiltonian $\mathcal{H}_0$ along any particular unperturbed orbit, remains constant. In the physically allowed region $\phi > -1$, the effective potential $\mathcal{V}(\phi)$ is nonnegative and possesses a single minimum $\mathcal{V} = 0$ at $\phi = 0$, while $\mathcal{V}(\phi) \to +\infty$ as $\phi \to -1^+$ or $\phi \to +\infty$. So for any value of energy H satisfying $0 \le H < \infty$, there exists a phase space trajectory $\big(\phi(\xi; H), \phi'(\xi; H)\big)$ which is a closed periodic orbit, mirror-symmetric about its turning points, for which $-1 < \phi(\xi) < \infty$ throughout, and which is unique up to its overall phase, or initial position at $\xi = 0$.

Making a canonical transformation to the action-angle variables of the free oscillator, $\phi = \phi(\mathcal{I}, \theta)$, $p = p(\mathcal{I}, \theta)$, we can express (3.42) as

$$\mathcal{H}(\mathcal{I}, \theta; \xi) = \mathcal{H}_0(\mathcal{I}) + \frac{\bar{a}^2 + \epsilon\cos\psi(\xi)}{2[1+\phi(\mathcal{I},\theta)]}, \tag{3.43}$$

where the action $\mathcal{I}$ is defined in terms the area in phase space contained within the unperturbed orbit $(\phi(\xi; H), \phi'(\xi; H))$ of energy H:

$$\mathcal{I} \equiv \frac{1}{2\pi}\oint p\, d\phi = \frac{1}{\pi}\int_{\phi_-}^{\phi_+} \phi'\, d\phi. \tag{3.44}$$

Here, ϕ_+ and ϕ_- are, respectively, the upper and lower turning points of the orbit, at which $\mathcal{V}(\phi_\pm) = H$:

$$\phi_\pm = H \pm \sqrt{H^2 + 2H}, \tag{3.45}$$

and in (3.44) we have used symmetry to reduce the integration path to the segment where $p \ge 0$. By making the change of variables $\phi = \phi_+ - (\phi_+ - \phi_-)\sin^2(u)$, the action (3.44) can be calculated with the help of a standard integral table, e.g., equation 2.584-13 on p. 163

of [72]:

$$\begin{aligned}\mathcal{I} &= \frac{2(\phi_+ - \phi_-)^2}{\pi\sqrt{1+\phi_+}} \int_0^{\pi/2} du\, \frac{\sin^2(u)\cos^2(u)}{\sqrt{1-\frac{\phi_+-\phi_-}{1+\phi_+}\sin^2(u)}} \\ &= \frac{4}{3\pi}\left[1+H-\sqrt{H^2+2H}\right]^{1/2}\left\{\frac{(1+H)E(\kappa)}{1+H-\sqrt{H^2+2H}} - K(\kappa)\right\}. \end{aligned} \tag{3.46}$$

Here, $K(\kappa)$, $E(\kappa)$ are complete elliptic integrals of the first and second kind, respectively, whose modulus satisfies

$$\kappa = +\left[2(1+H)\sqrt{H^2+2H} - 2H(2+H)\right]^{1/2}. \tag{3.47}$$

At this point, one could in principle use (3.46) to find $\mathcal{H}_0(\mathcal{I})$, but fortunately such a cumbersome inversion will not be necessary. The normalized (i.e., dimensionless) nonlinear frequency $\Omega(H)$ of the unforced oscillator is given by

$$\Omega(H) \equiv \tfrac{d}{d\xi}\theta = \tfrac{\partial \mathcal{H}_0}{\partial \mathcal{I}} = \left(\tfrac{\partial \mathcal{I}}{\partial H}\right)^{-1} = \frac{\pi}{2}\frac{\left[1+H-\sqrt{H^2+2H}\right]^{1/2}}{E(\kappa)}. \tag{3.48}$$

To put (3.42) in the desired form (3.43), the remaining ingredient needed is the canonical transformation $\phi = \phi(\mathcal{I},\theta)$. An explicit formulation of this will also not be needed, and we may proceed by formally assuming that we have made this substitution. Then, $\phi = \phi(\mathcal{I},\theta)$ is a periodic function of θ, and we can expand the driving term appearing in (3.43) in a Fourier series as

$$\frac{\bar{a}^2 + \epsilon\cos\psi(\xi)}{2\left[1+\phi(\mathcal{I},\theta)\right]} = \left[\bar{a}^2+\epsilon\cos\psi(\xi)\right]\sum_{n=-\infty}^{\infty} b_n(\mathcal{I})e^{in\theta}. \tag{3.49}$$

Because (3.49) is a real-valued function, the Fourier coefficients must satisfy $b_n = b_{-n}^*$, and are defined by

$$b_n(\mathcal{I}) = \tfrac{1}{2\pi}\int_0^{2\pi} d\theta\, \frac{e^{-in\theta}}{2[1+\phi(\mathcal{I},\theta)]}. \tag{3.50}$$

We put (3.50) into a form more amenable to calculation by changing variables using $\theta = \Omega(H)\xi$, and integrating over one period $0 \le \xi \le \Xi$ of the co-moving coordinate, where $\Xi \equiv \Xi(H) \equiv 2\pi/\Omega(H)$. Furthermore, if we choose the origin of the orbit associated with the energy H, or equivalently with the free action $\mathcal{I}(H)$, such that $\phi(\xi=0;H) = \phi_+$, the potential $\phi(\xi;H)$, and hence $[1+\phi(\xi;H)]^{-1}$, are symmetric about the point $\xi = \Xi/2$. Since the imaginary parts of (3.50) are obtained by integrating over the antisymmetric functions $\sin(n\Omega\xi)$, the b_n's are purely real and given by:

$$b_n(H) = b_n(H)^* = b_{-n}(H) = b_{-n}(H)^* = \tfrac{1}{\Xi(H)}\int_0^{\Xi(H)} d\xi\, \frac{\cos\left[n\Omega(H)\xi\right]}{2[1+\phi(\xi;H)]}. \tag{3.51}$$

Up to this point, no approximations have been made in the canonical transformations to the action-angle variables of the unforced oscillator. Now, we make the Single Resonance Approximation (SRA) [73] to (3.43), by assuming that the rapidly oscillating terms of the Hamiltonian average to zero and contribute negligibly to the dynamics. Anticipating the developments of Section 3.4, we realize that under certain frequency-locking conditions derived there, autoresonant excitation occurs, wherein the plasma wave amplitude, or equivalently free-oscillator energy $H(\xi) = \mathcal{T}\left(\phi'(\xi)\right) + \mathcal{V}\left(\phi(\xi)\right)$, grows secularly such that the dynamical frequency $\Omega(\xi)$ of the forced oscillator approximately matches that of the unforced oscillator at that same amplitude, $\Omega(\xi) \approx \Omega\big(H(\xi)\big)$. Simultaneously, this instantaneous oscillator frequency follows that of the driving beat frequency of the wave, $\Omega(\xi) \approx \Delta\omega(\xi)$. Under these assumptions, we can consistently neglect all terms in the sum (3.49) except the constant term or those terms with frequency dependence $\sim \pm\left[\Omega(\xi) - \Delta\omega(\xi)\right]$. The averaged Hamiltonian then becomes

$$\mathcal{H}(\mathcal{I}, \theta; \xi) = \mathcal{H}_0(\mathcal{I}) + \epsilon b_1(\mathcal{I}) \cos[\theta - \psi(\xi)] + \bar{a}^2 b_0(\mathcal{I}). \tag{3.52}$$

We now make an explicitly ξ-dependent canonical transformation to the rotating action-angle variables $(\hat{\mathcal{I}}, \Psi)$. Using the mixed-variable generating function

$$F_2(\hat{\mathcal{I}}, \theta; \xi) = \big[\theta - \psi(\xi)\big]\hat{\mathcal{I}}, \tag{3.53}$$

our old and new coordinates are related by

$$\Psi = \tfrac{\partial F_2}{\partial \hat{\mathcal{I}}} = \theta - \psi(\xi); \tag{3.54a}$$

$$\mathcal{I} = \tfrac{\partial F_2}{\partial \theta} = \hat{\mathcal{I}}. \tag{3.54b}$$

Dropping carets from the new action, the Hamiltonian in the rotating frame is

$$\mathcal{K}(\mathcal{I}, \Psi; \xi) = \mathcal{K}(\mathcal{I}, \theta, \xi) + \tfrac{\partial F_2}{\partial \xi} = \mathcal{H}_0(\mathcal{I}) - \Delta\omega(\xi)\mathcal{I} + \epsilon b_1(\mathcal{I}) \cos\Psi + \bar{a}^2 b_0(\mathcal{I}).$$

The resulting SRA canonical equations of motion are

$$\tfrac{d}{d\xi}\Psi = \quad \tfrac{\partial \mathcal{K}}{\partial \mathcal{I}} = \Omega(\mathcal{I}) - \Delta\omega(\xi) + \epsilon \frac{\partial b_1}{\partial \mathcal{I}} \cos\Psi + \bar{a}^2 \frac{\partial b_0}{\partial \mathcal{I}}, \tag{3.55a}$$

$$\tfrac{d}{d\xi}\mathcal{I} = -\tfrac{\partial \mathcal{K}}{\partial \Psi} = \epsilon b_1(\mathcal{I}) \sin\Psi, \tag{3.55b}$$

for which we next determine the required conditions for phase-locking.

3.4 Autoresonant Response

The essential ingredients for autoresonance, as exploited in the present scheme, are: a nonlinear, oscillatory degree of freedom (in our case, the plasma wave) evolving, in the

absence of any forcing, within an integrable region of phase space; a continuous functional relationship between the nonlinear frequency and energy of the oscillation, possessing a well-defined linear limit; an applied oscillatory driving force (in our case, the modulated ponderomotive envelope of the lasers) which is sufficiently small as to be considered perturbative, so that the notion of the nonlinear frequency of the unforced oscillator remains meaningful; and initial conditions and forcing profile consistent with Adiabatic Passage Through Resonance (APTR) – namely, an initially unexcited system (quiescent plasma), and an initial drive frequency sufficiently far from the linear resonance, with subsequent time-dependence that is sufficiently slowly-varying but otherwise arbitrary.

The key consequence of the autoresonant beat-wave generation is the robust entrainment between the three relevant (normalized) frequencies: the beat frequency $\Delta\omega(\xi)$ of the driving lasers, the instantaneous frequency $\Omega(\xi)$ of the driven plasma wave, and the nonlinear frequency $\Omega(H)$ of the unforced plasma wave. Assuming this phase-locking is achieved, amplitude control of the plasma wave can be simply understood via (3.48): because the frequency of the freely-evolving nonlinear plasma wave is a function of the energy, phase-locking of the driven wave to the envelope such that $\Delta\omega(\xi) \approx \Omega(\xi) \approx \Omega(H)$ implies that changing the drive frequency will correspondingly change the oscillator energy. In our case, (3.48) indicates that $\Omega(H)$ is a decreasing function of the energy, so that in order to increase the plasma wave amplitude one must decrease the beat frequency as a function of the co-moving position ξ (or, equivalently, in time t at the source of the driving lasers.)

While we have indicated how autoresonance can lead to large plasma waves, we have not yet shown under what conditions such phase-locking occurs. To answer this question, we first consider the linear and weakly nonlinear response, valid up to the point where the normalized sweeping drive frequency is of the order of the normalized linear resonance, $\Omega(H) \approx \Delta\omega(\xi) \gtrsim 1$. Next, we consider the fully nonlinear case, for which we derive adiabaticity requirements for autoresonance.

3.4.1 Small Amplitude Response and Phase-Locking

When the drive is first applied with its frequency above the linear resonance, the plasma wave amplitude (and, hence, H and $\mathcal{I}(H)$) are small and we can linearize equations (3.55a - 3.55b). In this limit, the oscillator is harmonic, with $\Omega(H) = 1$, $\phi(\xi) = \sqrt{2H}\cos(\xi)$, and

$H = \mathcal{I}$, so that $b_0 = -\frac{1}{2}H$, $b_1 = \sqrt{2H}/4$, and equations (3.55) become

$$\frac{d}{d\xi}\Psi = 1 - \tfrac{1}{2}\bar{a}^2 - \Delta\omega(\xi) + \tfrac{1}{4\sqrt{2}}\epsilon\tfrac{1}{\sqrt{\mathcal{I}}}\cos\Psi, \tag{3.56a}$$

$$\frac{d}{d\xi}\mathcal{I} = \tfrac{\epsilon}{2\sqrt{2}}\sqrt{\mathcal{I}}\sin\Psi. \tag{3.56b}$$

Note that the $(1-\frac{1}{2}\bar{a}^2)$ contribution to $\frac{d}{d\xi}\Psi$ corresponds, in normalized units, to the leading order expansion of the effective plasma frequency in the EM dispersion relation:

$$\omega_{p_{\text{eff}}} \equiv \omega_p/\gamma_0 = \omega_p/\sqrt{1+\bar{a}^2} \approx \omega_p\left(1 - \tfrac{1}{2}\bar{a}^2 + \ldots\right), \tag{3.57}$$

which is shifted from the bare value ω_p due to the transverse quiver motion of the electrons in the applied laser fields. While we never explicitly invoked any small a^2 approximation, only the leading-order correction appears because, in making the SRA, we ignored terms of the form $\bar{a}^2 e^{in\theta}$ for $|n| \geq 1$. But for sufficiently large intensity $\bar{a}^2$ such terms can appreciably effect the motion despite being off-resonance. One could, *a posteriori*, partially correct for this defect by replacing $(1 - \frac{1}{2}\bar{a}^2)$ with $\left(1+\bar{a}^2\right)^{-1/2}$ in the equation of motion (3.56a), but this will be unnecessary for the small values of $\bar{a}$ considered here. Physically, we should simply ensure that the drive frequency begins sufficiently far above, and then is slowly swept past, the effective (or renormalized) frequency $\omega_{p_{\text{eff}}}$, rather than the bare frequency ω_p, if the difference is not negligible. Precise knowledge of the exact value of the effective linear plasma resonance including the effects of transverse quiver is not needed.

For the driven plasma wave, we have $\Delta\omega(\xi) \sim \Omega(\xi) \sim 1$, while we seek solutions for which $|\Delta\omega(\xi) - \Omega(\xi)| \ll 1$ as a result of special initial and forcing conditions: an initially unperturbed plasma, $\phi(\xi = 0) = \phi'(\xi = 0) = 0$; an initial tuning of the beat frequency above resonance, i.e., $\Delta\omega(\xi = 0) > \omega_{p_{\text{eff}}}$, and subsequently a slow downward frequency chirp through resonance, where the chirp rate is be characterized by a parameter $\alpha \equiv \alpha(\xi) \equiv -\frac{d}{d\xi}\Delta\omega(\xi)$, with $0 \leq \alpha \ll 1$.

For a linear frequency chirp around the effective plasma frequency,

$$\Delta\omega(\xi) = 1 - \tfrac{1}{2}\bar{a}^2 - \alpha\xi, \tag{3.58}$$

the simple harmonic oscillator equations (3.56) have analytic solutions in terms of the Fresnel Sine and Cosine integrals [74]. We briefly summarize the extensive characterization of these solutions found in [64]. When the drive is first applied far from resonance, the oscillator response can be divided into two components: one ringing component precisely at $\omega_{p_{\text{eff}}}$, and the other at the driving frequency $\Delta\omega(0)$, both of small amplitude. The singular term $\sim \mathcal{I}^{-1/2}$ in (3.56a) allows for a large change in phase at small amplitude without

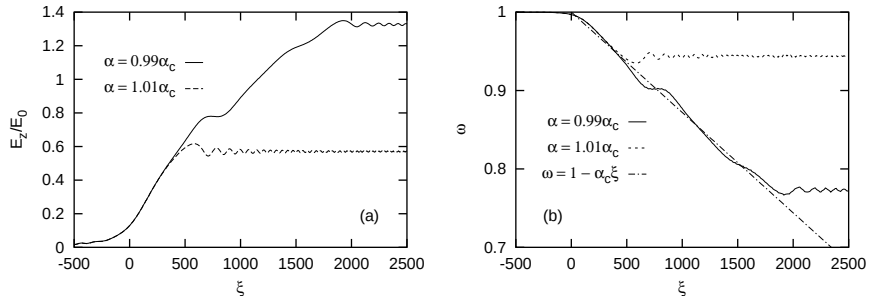

Figure 3.1. Demonstration of the critical autoresonant behavior for $\epsilon = 0.005$. In (a), chirp rates 1% below critical (solid line) yield final longitudinal fields above linear wavebreaking, whereas chirp rates 1% above saturate at $0.6E_0$. (b) compares the phase locking in autoresonance (solid line) to that of non-autoresonant behavior (dashed line). The critical drive (mixed line) is included for reference.

violating the requirement that $\phi(\xi)$ remain smooth, so that the response at the driven frequency can adjust itself to the drive, and phase-locking can occur. As the frequency is swept toward the resonance, these driven, phase-locked oscillations grow. Meanwhile, the response at the resonant frequency has no such phase relation with the drive, and remains small. In this way, we essentially have one growing, phase-locked plasma wave when the drive frequency reaches the resonant frequency.

As the resonance is approached, the amplitude of the plasma wave begins to become large and one must account for the growing nonlinearities. We therefore expand the expression for the free nonlinear frequency (3.48) to first order:

$$\Omega(H) = 1 - \tfrac{3}{8}H = 1 - \tfrac{3}{8}\mathcal{I}, \tag{3.59}$$

and continue to use the linearized frequency chirp (3.58). Making the change of variable $\mathcal{A} \equiv 4\sqrt{2\mathcal{I}}$, we have

$$\tfrac{d}{d\xi}\Psi = \alpha\xi - \tfrac{3}{256}\mathcal{A}^2 + \frac{\epsilon}{\mathcal{A}}\cos\Psi, \tag{3.60a}$$

$$\tfrac{d}{d\xi}\mathcal{A} = \epsilon\sin\Psi. \tag{3.60b}$$

This set of equations can be reduced to a single first-order ordinary differential equation by defining the complex dynamical variable $Z \equiv -\sqrt{256/3\alpha}\mathcal{A}e^{i\Psi}$, a re-scaled independent variable $\zeta \equiv \sqrt{\alpha}\xi$, and the dimensionless parameter $\mu \equiv \epsilon\sqrt{3/(256\alpha^{3/2})}$, to obtain

$$i\tfrac{d}{d\zeta}Z + \left(\zeta - |Z|^2\right)Z = \mu. \tag{3.61}$$

Thus, the weakly nonlinear problem is now described by a dynamical equation with a single parameter μ that combines the drive strength ϵ and the chirp rate α. It has been

found numerically [75] that the solution to (3.61) has a bifurcation at the critical value $\mu = \mu_c \approx 0.411$. For $\mu < \mu_c$, the plasma wave response quickly dephases from the drive, resulting in only small excitations. In contrast, for $\mu > \mu_c$, phase-locking occurs and the plasma wave can grow to large amplitude. This critical behavior with respect to μ translates into a critical drive strength ϵ and chirp rate α for the nonlinear oscillator to be autoresonantly excited. The condition is

$$\alpha \leq \left(\frac{3\epsilon^2}{256\mu_c^2}\right)^{2/3} \approx 0.169\,\epsilon^{4/3}. \tag{3.62}$$

Thus, for a given laser intensity one can readily find the maximum chirp rate that can be tolerated and still obtain high amplitude plasma waves. We demonstrate the sensitivity of this critical behavior for $\epsilon = 0.005$ in Fig. 3.1, which plots numerical solutions to the full quasi-static equation of motion (3.41). Plot (a) shows the excited longitudinal electric field for chirp rates just above and just below the critical rate $\alpha_c \approx 1.28 \times 10^{-4}$; Plot (b) demonstrates the dynamic frequency-locking that occurs in autoresonance, and compares this to the case where autoresonance fails to occur.

3.4.2 Fully Nonlinear Autoresonant Response

If phase-locking is maintained through the weakly nonlinear regime, the amplitude continues to grow and one must consider further nonlinearities beyond those included in (3.61). In this case, there arises more stringent restrictions on the chirp rate for adiabatic phase-locking to persist. We calculate this condition by first finding a second-order equation for the phase Ψ:

$$\frac{d^2}{d\xi^2}\Psi = \left\{\frac{d\Psi}{d\xi}, \mathcal{K}\right\} + \frac{\partial}{\partial\xi}\frac{d\Psi}{d\xi} = \frac{\partial^2\mathcal{K}}{\partial\mathcal{I}\partial\Psi}\frac{\partial\mathcal{K}}{\partial\mathcal{I}} - \frac{\partial^2\mathcal{K}}{\partial\mathcal{I}^2}\frac{\partial\mathcal{K}}{\partial\mathcal{I}} + \frac{\partial^2\mathcal{K}}{\partial\mathcal{I}\partial\xi} + \frac{\partial}{\partial\xi}\frac{d\Psi}{d\xi}, \tag{3.63}$$

where we have made use of the usual canonical Poisson bracket with respect to the rotating-frame action-angle variables:

$$\{F\,, G\} \equiv \frac{\partial F}{\partial\Psi}\frac{\partial G}{\partial\mathcal{I}} - \frac{\partial F}{\partial\mathcal{I}}\frac{\partial G}{\partial\Psi}. \tag{3.64}$$

Using this expression (3.63) as well as (3.55a), the phase Ψ is seen to obey the following second-order equation:

$$\tfrac{d^2}{d\xi^2}\Psi + \epsilon\frac{\partial b_1}{\partial\mathcal{I}}\sin\Psi\,\tfrac{d}{d\xi}\Psi + \left\{\tfrac{d}{d\xi}\Delta\omega - \epsilon b_1(\mathcal{I})\sin\Psi\left[\frac{\partial\Omega}{\partial\mathcal{I}} + \bar{a}^2\frac{\partial^2 b_0}{\partial\mathcal{I}^2} + \epsilon\frac{\partial^2 b_1}{\partial\mathcal{I}^2}\cos\Psi\right]\right\} = 0. \tag{3.65}$$

Now, we assume (see, e.g., [76]) that the free action can be written as

$$\mathcal{I} = \mathcal{I}_0 + \Delta\mathcal{I}, \tag{3.66}$$

where $\mathcal{I}_0 = \mathcal{I}_0(\xi)$ is the slowly-varying, secularly-evolving action about which there are small oscillations given by $\Delta\mathcal{I} = \Delta\mathcal{I}(\xi)$. These oscillations correspond to fluctuations in Ψ about its (slowly-varying) phase-locked value $\bar{\Psi} = \bar{\Psi}(\xi)$, an example of which can be seen in Fig. 3.1(b). In the autoresonant case (solid line), we see that as the plasma wave is excited, its frequency does indeed make small oscillations about the drive frequency. We further note that as one decreases the chirp rate α from its critical value α_c, we obtain more total oscillations in frequency over the longer excitation time, but with a slightly smaller magnitude of excursions from the drive frequency.

Using the form (3.66) for the action, the lowest-order equation for the phase is identical to (3.65), with $\mathcal{I}$ being replaced everywhere by $\mathcal{I}_0$. In this way, the phase itself is seen to obey a nonlinear oscillator equation, with an effective nonlinear damping (or anti-damping) term, and a conservative "forcing," described by the terms in the braces, which is derivable from an effective "potential" (not to be confused with any previously-mentioned potential) whose shape is dictated by the slowly-evolving action $\mathcal{I}_0$ and the drive parameters ϵ and α. Phase-locking then corresponds to trapping of Ψ in a basin of this effective potential. In order for the phase Ψ to be trapped about some value $\bar{\Psi}$, the effective potential must possess a local minimum there, and the non-conservative term must either provide damping or else remain sufficiently small if excitatory. In fact, the non-conservative term is expected to be small compared to the second term in the conservative force. Their ratio is given by $\left(\frac{1}{b_1}\frac{\partial b_1}{\partial \mathcal{I}_0}/\frac{\partial \Omega}{\partial \mathcal{I}_0}\right)\left(\frac{d\Psi}{d\xi}\right)$, where typically $\left(\frac{1}{b_1}\frac{\partial b_1}{\partial \mathcal{I}_0}/\frac{\partial \Omega}{\partial \mathcal{I}_0}\right) \sim \left(\frac{1}{b_1}\frac{\partial b_1}{\partial \Omega}\right) \sim \frac{1}{\Omega} \sim O(1)$, while $\frac{d\Psi}{d\xi} \ll O(1)$ because the nonlinear phase oscillations are slow compared to ω_p. Actually, over most of the typical range of parameter values, the last two terms in the conservative forcing are small compared to the first two terms, since they are higher order in the drive strength, and as a first approximation the phase oscillations are governed by the "biased" pendulum equation

$$\frac{d^2}{d\xi^2}\Psi \approx \alpha(\xi) + \epsilon b_1(\mathcal{I}_0)\frac{\partial \Omega}{\partial \mathcal{I}_0}\sin\Psi, \tag{3.67}$$

although we will continue to work with the full equation (3.65). Clearly, for a given $\mathcal{I}_0$, if the normalized chirp rate α remains sufficiently small compared to the normalized drive strength ϵ, then the effective potential will be of the tilted-washboard variety, with a series of periodically-spaced local minima in Ψ at intervals of 2π. As α increases, the depth of these wells decreases, until they finally disappear, as does any opportunity for phase-locking.

Thus, a necessary condition for trapping is that the effective force can actually vanish

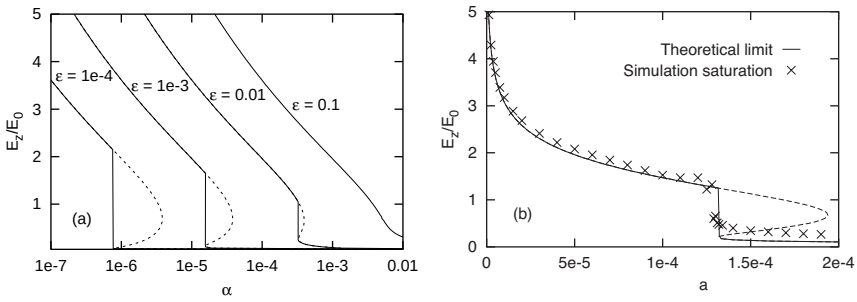

Figure 3.2. Maximum plasma wave amplitude obtainable before the "slowness" condition (3.69) is violated. In (a), we plot the maximum longitudinal field as a function of the chirp rate for four different driving strengths. Dotted lines indicate hysteresis at α_c given by (3.62). Part (b) compares the theory (solid line) with numerically determined saturation for $\epsilon = 0.005$

at some fixed point $\bar{\Psi}$:

$$\alpha(\xi) + \epsilon b_1(\mathcal{I}_0) \sin\bar{\Psi}\left[\frac{\partial\Omega}{\partial\mathcal{I}_0} + \bar{a}^2\frac{\partial^2 b_0}{\partial\mathcal{I}_0^2} + \epsilon\frac{\partial^2 b_1}{\partial\mathcal{I}_0^2}\cos\bar{\Psi}\right] = 0. \tag{3.68}$$

For given $\mathcal{I}_0$, ϵ, and α, this determines the average, slowly-varying phase $\bar{\Psi}$ about which trapping occurs. In the linear ($\mathcal{I}_0 \to 0$), weakly forced ($0 \le \epsilon \lll 1$) or weakly chirped ($\alpha \to 0^+$) regimes, this phase value is known [64] to be $\bar{\Psi} \approx \pi$, but for nonlinear plasma waves, stronger forcing, or faster chirping, this phase can shift appreciably. Since $|\sin\bar{\Psi}|$, $|\cos\bar{\Psi}| \le 1$, equation (3.68) also imposes an upper bound on the frequency chirp $\alpha(\xi)$ for which the phase can remain trapped in autoresonance, regardless of the actual value of the phase at which locking occurs. Setting $|\sin\bar{\Psi}| = |\cos\bar{\Psi}| = 1$ above, we obtain an upper bound on α beyond which any phase-locking is impossible:

$$0 \le \alpha(\xi) \le \epsilon|b_1(\mathcal{I}_0)|\left[\left|\frac{\partial\Omega}{\partial\mathcal{I}_0}\right| + \bar{a}^2\left|\frac{\partial^2 b_0}{\partial\mathcal{I}_0^2}\right| + \epsilon\left|\frac{\partial^2 b_1}{\partial\mathcal{I}_0^2}\right|\right]. \tag{3.69}$$

Again, for realistic parameters, the force balance typically resides predominately between the first two terms in (3.68), and hence this upper bound, although approximate, is expected to provide a reasonably tight cutoff for autoresonant phase-locking, which has been confirmed by numerical simulation. This inequality can also be thought of as giving the maximum achievable plasma wave amplitude (implicitly as a function of $\mathcal{I}_0$) for a given chirp rate and laser power. We show the dependence of the saturated longitudinal field on the chirp rate as a solid line in Fig. 3.2(a) for a number of different drive strengths. For a fixed drive strength ϵ, the maximum attainable electric field jumps discontinuously at the critical chirp rate α_c given by (3.62). The dotted lines correspond to solutions of (3.69) that cannot be accessed when starting from vanishing initial longitudinal field, and as such

constitutes a form of hysteresis in the excitation. The solid lines show the stable branches for the case of interest: excitation from a quiescent plasma via slow downward chirping from a drive initially above resonance. Fig. 3.2(b) shows a comparison of the theoretical maximum amplitude and that found by numerically integrating the quasi-static equation of motion (3.41).

Asymptotic expansions and numerical plots for $\Omega(\mathcal{I})$ and $b_n(\mathcal{I})$ reveal that the right-hand-side of (3.69) actually decreases with increasing $\mathcal{I}_0$ (at least beyond moderate values of $\mathcal{I}_0$), making the adiabatic nonlinear phase-locking condition (3.69) become increasingly stringent as the plasma wave grows. For fixed chirp rate and drive amplitude, the growing plasma wave will eventually fall out of frequency-locking. In the physical pendulum limit discussed above, note that when this cutoff is reached and autoresonance is lost, the average phase should be near $\pm\pi/2$. Mathematically, this saturation can be traced back to the nonlinear nature of the forcing in (3.41), where the effective strength of the forcing depends both on the external drive intensity and the plasma wave amplitude; physically, to the fact that as the plasma wave grows, the electrons become increasingly relativistic and therefore less susceptible to displacement via the ponderomotive forcing at a fixed intensity, so the effective strength of a given drive actually decreases. In principle, this may be counteracted by either slowly increasing the laser amplitude or, more practically, by slowly decreasing the chirp rate in time. In the latter case, the rate of growth of the plasma wave will fall off because it is directly tied to the decreasing drive frequency, but the wave can remain frequency-locked with the drive indefinitely (with some oscillating excursions.)

As a result of relativistic detuning, recall that in the original RL/TD scheme, the plasma wave amplitude exhibits slow (compared to ω_p^{-1}) nonlinear modulations which appear as beating, i.e., periodic amplitude oscillations up to the RL limit and back to a nearly unexcited state. As the wave is driven to a high amplitude its phase slowly slips until it is more than $\sim \pi/2$ out of phase with respect to the laser beat and then gives its energy back to the lasers, then continues to shift further out of phase, only to be re-excited when its amplitude approaches zero and it can re-establish phase matching with the drive. Both the frequency difference between plasma wave and drive and the phase-lag exhibit continuous oscillations in time.

In the DMG scheme, which relies on autoresonance but starts at linear resonance, the plasma wave amplitude not only can peak at a higher maximum than in the original scheme, but typically will sustain a higher average value at long times, exhibiting a nonlinear ringing about some non-zero saturated value rather than a full beating. We suspect that

this behavior results from two features of the chirping: when the autoresonant adiabatic condition fails, the wave phase is closer to its neutral value, with respect to energy exchange with the drive, and after autoresonance is lost, the frequency difference and phase lag each grow secularly in time rather than exhibiting oscillations. Depending on initial parameters, the initial growth up to the absolute maximum can either be essentially monotonic, or exhibit a "staircase" behavior, with intermittent plateaus or dips at local maxima between periods of resumed growth, before finally leveling off with some ringing. When starting from resonance, the originally published numerical examples [60] and our own simulations suggest that the observed ripples in the amplitude excitation are actually minimized at some intermediate value of the chirp rate, implying that the excitation may not be fully autoresonant in some sense.

In our autoresonant scheme based on APTR, the behavior more closely resembles that in DMG scheme, but exhibits more nearly monotonic growth, higher peak fields, and less ringing after saturation. Specifically, some slow oscillations in the amplitude may be present early on, but once nonlinear phase-locking has been achieved, the growth remains monotonic or virtually so until finally the amplitude levels off and appears to almost saturate, with only a very small amount of subsequent ringing.

The extent of the amplitude and phase excursions, which affect how regularly the excitation grows and how closely the phase is locked during autoresonant excitation, and also how much ringing persists after saturation, are determined by how deeply the phase is trapped in its effective potential well. This in turn depends both on how steep and how deep is the available well (determined by the plasma wave amplitude and drive lasers parameters), and to what extent the phase can be nudged into position near the bottom of the well and kept there (determined by the initial conditions and the adiabaticity of the chirped forcing).

Numerical simulations suggest that, as one would expect, the extent or depth of this trapping is improved by using a stronger drive (at least up to some moderately strong value), starting the drive frequency further above resonance, and chirping more slowly. In practice, of course, each of these strategies involves trade-offs. Increasing the drive strength increases the growth rates for laser-plasma instabilities that might disrupt the forcing. Either increasing the initial frequency up-shift or decreasing the chirp rate decreases the final amplitude that can be reached during a fixed interaction time.

3.5 Experimental Considerations

Unfortunately, as has been alluded to previously, the PBWA does not have unlimited time to be excited, as deleterious instabilities will eventually destroy wave coherence. For the parameters of interest, the oscillating two-stream (also referred to as modulational) instability limits the lifetime of coherent Langmuir waves to the ion time-scale, i.e., for times of the order of a few $1/\omega_i$. Although it is possible that the growth of this instability may be mitigated somewhat by the use of a chirped laser, in this paper we will use as a conservative figure the results of Mora *et. al.* [46] to set the time limit during which we can excite a coherent plasma wave suitable for accelerator applications.

For the relatively cold plasmas and moderately intense lasers we consider, it is shown in [46] that the growth rate of the oscillating two-stream instability is approximately equal to ω_i, and that this instability impedes plasma wave excitation and destroys coherence after about 5 e-foldings. Thus, we see that the drive lasers should have time duration $T \lesssim 5/\omega_i$. If one chirps the drive frequency leading to a total shift $\delta\omega$ during the autoresonant excitation, then the normalized chirp rate is limited to

$$\alpha \gtrsim 0.2 \, \frac{\omega_p}{\omega_i} \frac{\delta\omega}{\omega_p}, \tag{3.70}$$

or approximately $\alpha \gtrsim 2.3 \times 10^{-3} \, \delta\omega/\omega_p$ for singly-ionized Helium. Below, we choose two experimentally relevant parameter sets, one corresponding to a 10 μm CO_2 laser; the other, to a 800 nm Chirped Pulse Amplification (CPA) [77, 78] Ti:Sapphire laser system. We demonstrate how, beginning with the laser frequency above the linear resonance and then slowly decreasing it, one can robustly excite plasma waves to amplitudes larger than the cold, linear wave-breaking limit in times commensurate with onset of the oscillating two-stream instability.

3.5.1 CO_2 Laser at 10 μm

We consider parameters roughly corresponding to the most recently published UCLA upgrade [51]. We assume two pulses of duration $T = 100$ ps which enter the plasma at $t = 0$, central wavelengths near $\lambda = 10$ μm, and normalized intensities $a_1 = a_2 = 0.14$, corresponding to a normalized drive strength $\epsilon = 0.02$, so the threshold condition (3.62) implies that the normalized chirp rate should satisfy $\alpha(t) = -\frac{d}{dt}\Delta\omega(t)/\omega_p < 0.0009$. We choose a linear chirp, so that in physical units the beat frequency is given by $\Delta\omega(t) = (\mu_0 + \alpha\,\omega_p t)$, where $\mu_0 = 1.15$ and $\alpha = 0.00065$, with a total frequency sweep from $\Delta\omega(t =$

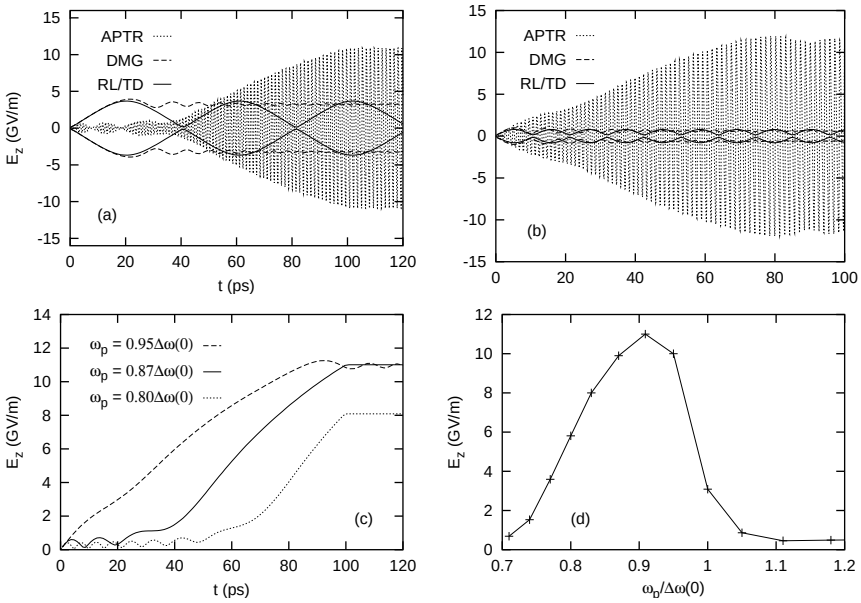

Figure 3.3. Plasma wave excitation for a 10μm CO_2 laser with intensity 2.7×10^{14}W/cm^2 ($\epsilon = 0.02$). In (a), the APTR case has $\omega(0) = 1.15\omega_p$, and $\omega = \omega_p$ at $t \approx 40$ps ($\alpha = 0.00065$). Total chirp is 1.5% of laser frequency. For comparison, we include the envelopes with no chirp, and DMG chirp starting on resonance. (b) is the same as (a), with ω_p changed by 10%. In (c), we show uniform accelerating fields via APTR for variations in ω_p of order 10%. (d) demonstrate robust field excitations of 5-10 GeV/m for density errors of order $\pm$20%.

$0) = 1.15$ to $\Delta\omega(t = T) = 0.74$. For these parameters, we collect the results from simulations integrating the quasi-static equation of motion (3.41) in Fig. 3.3. 3.3(a) demonstrates the excitation of a uniform accelerating field E_z of 10 GV/m, which is above the linear cold wave-breaking limit of $E_0 \approx 8.8$ GV/m, but below the cold relativistic limit $E_{\rm WB} \approx 61$ GV/m. The total chirp is modest, only about 1.5% of the laser carrier frequency $2\pi c/\lambda$.

For comparison, we also plot the simulated envelopes of the longitudinal field for the resonant RL/TD case $\Delta\omega(t) = \Delta\omega(0) = 1$, and for the chirped DMG scheme starting on linear resonance, $\Delta\omega(t) = (1 - \alpha\,\omega_p t)$, but using the same chirp rate as above. The resonant case demonstrates the characteristic RL limit of $E_z \leq E_{\rm RL} = (16\epsilon/3)^{1/3} E_0 \approx$ 4.2 GV/m, whereas the DMG scheme fails to achieve appreciable dynamic phase-locking, and the final plasma wave amplitude is about the same as in the resonant (unchirped) case. Using approximately these parameters, UCLA experiments have inferred accelerating amplitudes up to 2.8 GV/m [50] over short regions of plasma. More recently, plasma density variations corresponding to $E_z \approx 0.2 - 0.4$ GV/m have been directly measured with Thomson scattering [79].

Perhaps more important than the higher amplitude field in the autoresonant APTR case, is the fact that excitation is very robust with respect to mismatches between the beat frequency and the plasma frequency. In practice, these mismatches inevitably result from limited diagnostic accuracy or shot-to-shot jitter in the plasma or laser parameters. Because one sweeps over a reasonably broad frequency range and one only needs to pass through the resonance at some indeterminate point during the chirp history, no precise matching is required, and the exact value of the plasma density need not be accurately known. This robustness is demonstrated in Fig. 3.3(b)-(d). Plots (b) and (c) show the longitudinal field profile attained when there are variations in the density, and we see that APTR yields uniform, large amplitude fields over a wide range of densities. In Fig. 3.3(d), we plot the final accelerating gradient achieved via APTR when we vary the value of ω_p over a range of $\pm 10\%$, from its "design" value, while keeping the laser parameters fixed. We see large levels of excitation for a wide range in plasma variation, corresponding roughly to density mismatch/errors up to 20%. Thus, not only is autoresonant plasma wave excitation effective in avoiding saturation from detuning, it also mitigates experimental uncertainties in or shot-to-shot variations of plasma density.

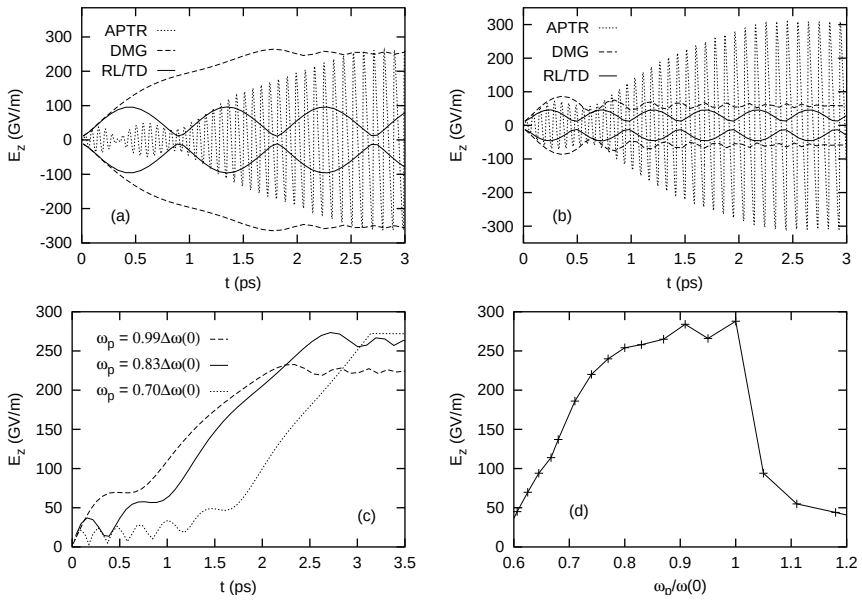

Figure 3.4. Plasma wave excitation for a 800 nm Ti:sapphire laser with $\bar{\omega}/\omega_p = 25$, intensity 2×10^{17} W/cm^2 ($\epsilon = 0.09$). APTR parameters are $\Delta\omega(0) = 1.2$, and $\Delta\omega = 1$ at $t \approx 1.4$ ps ($\alpha = 0.00225$). Total chirp is 3% of the laser frequency. In (a), we see that the DMG and APTR scheme give approximately the same final fields, about three times that for on-resonance. In (b), we changed ω_p by 10%, and the excitation for the DMG and the on-resonant scheme drop considerably, while APTR maintains its large excitation. (c) demonstrates that APTR gives robust, uniform accelerating fields for wide variations in density, while (d) indicates the insensitivity of APTR to errors of $\pm 35\%$ in the plasma density, while still yielding large amplitude plasma waves.

3.5.2 Ti:Sapphire Laser at 800 nm

Here, we analyze a representative case for a Ti:Sapphire CPA laser in a singly-ionized He plasma, with $n_0 = 1.4 \times 10^{18}$ cm^{-3}, so that $\omega_p/\bar{\omega} = 1/25$, and a laser duration $T = 3.2$ ps, chosen to correspond to the modulational instability limit. If we consider two 1 J pulses compressed to this time duration and focused to a waist of $w_0 = 6$ μm, this implies intensities of $I_0 = 2.0 \times 10^{17}$ W/cm^2, so that, with $\lambda \approx 800$ nm, we have $a_1 = a_2 = 0.3$, and $\epsilon = 0.09$. We choose $\Delta\omega(t = 0) = 1.2$, $\Delta\omega(t = T) = 0.5$, so that $\alpha = 0.0025$. The resulting plasma wave excitation is shown in Fig. 3.4(a). Here, we see maximum longitudinal electric fields $E_z \approx 260$ GV/m, corresponding to $\sim 1.6\, E_0 \approx 0.25 E_{\rm WB}$. For comparison, we also plot the resonant case, for which detuning results in maximum fields corresponding to the familiar RL limit $(16\epsilon/3)^{1/3} E_0 \approx 125$ GV/m, and the DMG case (starting on resonance), which yields results similar to the APTR case. The distinction between passing though resonance and starting on resonance can be seen, however, in Fig. 3.4(b), which indicates that a small change in the value of the plasma frequency has a dramatic effect when starting on resonance, but little effect when passing through resonance. The uniform, robust acceleration fields obtained via APTR are shown in Fig. 3.4(c) for density variations $\pm 20\%$. Finally, Fig. 3.4(d) shows the robustness of autoresonant excitation, for which density imperfections of $\pm 35\%$ have little effect on the accelerating gradients achieved.

3.6 Discussion: Comparisons, Scalings, and Extensions

In comparison to other PBWA schemes, including the fixed beat-frequency approach, either at (RL/TD) or below (TSS) linear resonance, the chirped (DMG) scheme, involving downward chirping from resonance, or the non-resonant PBWA, scheme, recently proposed by Filip *et al.* [79, 80], involving strongly forced waves at frequency shifts well below resonance in a marginally underdense plasma, the autoresonant/APTR PBWA enjoys a number of advantages, in terms of plasma wave amplitude, robustness, and quality. In previous sections we have seen how, for given drive laser intensity, autoresonant excitation yields longitudinal fields that can be considerably higher than the RL limit set by relativistic detuning of the plasma wave. We have also seen how APTR, i.e., slowly sweeping the frequency downward through resonance, provides a much greater degree of robustness to density mismatches, since neither the final amplitude nor frequency of the plasma wave is very sensitive to the precise location of the actual linear resonance, or to the precise chirp history.

Direct comparison with the chirped DMG scheme is slightly more complicated, as revealed in Fig. 3.5, because both schemes rely on autoresonance. In principle, given unlimited excitation time in the APTR case, and for any fixed drive laser intensity, the plasma wave can be autoresonantly excited to any amplitude at or below the nonlinear wave-breaking limit by choosing a sufficiently slow chirp rate. This is not always true for the DMG case starting at resonance, where performance appears to peak at some intermediate chirp rate, with rapid detuning for significantly faster chirp rates, and excessive ringing for significantly slower chirp. Without time constraints, the autoresonant APTR approach can always produce higher longitudinal fields for the same drive laser intensity or comparable longitudinal fields with smaller intensity. But in practice, the time allowed for excitation is inevitably limited, typically by ion instabilities as previously addressed, or even if these are somehow controlled, then by laser scattering or modulational instabilities, or ultimately by Landau or collisional damping of the Langmuir wave or hydrodynamic expansion of the plasma.

If the DMG and the APTR schemes are compared for realistic parameters using the same drive laser intensity and chirp rate, then autoresonant APTR excitation consistently results in larger electric fields for excitation times on the order of few ω_i^{-1}. But if the DMG scheme is started precisely on resonance, but using a slower chirp rate so as to achieve roughly the same final drive frequency as the APTR approach (which started above resonance), then the DMG approach can experience phase-locking over a longer time and can achieve slightly higher fields in the finite excitation time. Essentially, by using a sufficiently strong drive, chirp rates that are just slow enough to be adiabatic, and an initial beat frequency precisely at resonance, the DMG trajectory can become autoresonantly phase-locked, but without having to waste time by chirping down from some point well above the resonance as is done in the APTR case. The catch is that such phase-locking behavior starting from resonance is quite sensitive to the initial conditions, and DMG will fail to achieve persistent phase-locking if the resonance is missed by just a few percent, so robust phase-locking behavior is unlikely to be experimentally reproducible, unless one begins with the beat frequency safely above resonance and then adiabatically sweeps through it, rather than trying to start precisely on resonance.

In order to excite larger electric fields at a given plasma density, autoresonant/APTR PBWA requires either more intense drive lasers or longer excitation times or both. Even if technologically feasible, moving to larger a^2 introduces its own problems (as discussed below), and in any event tends to undercut one of the primary comparative advantage of the PBWA over either the standard short-pulse or the self-modulated LWFA. Since excitation

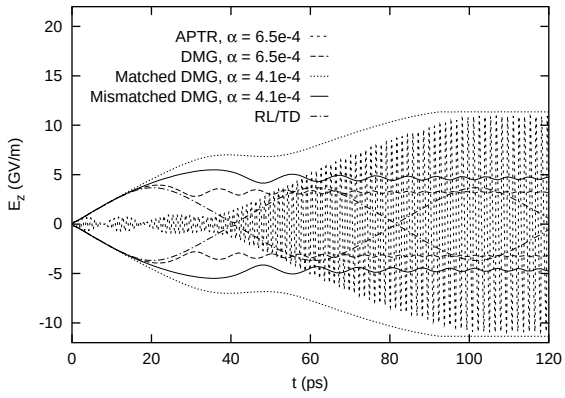

Figure 3.5. Comparison of three plasma beat-wave schemes for the CO_2 laser of Sec. 3.5.1. Autoresonant/APTR excitation is compared to the DMG scheme with identical chirp, and with a chirp chosen to maximize excitation. For identical chirp, DMG results in a field of approximately 50% less amplitude. The optimized DMG is comparable in amplitude to APTR, but a mismatch in plasma density of 2% has a sizable effect, unlike the autoresonant/APTR robustness to density fluctuations of order 10% to 20%.

times in the PBWA tend to be limited by deleterious ion instabilities (oscillating two-stream instabilities and perhaps ion acoustic instabilities) growing on the ω_i^{-1} time-scale, while for fixed drive laser intensity both the maximum chirp rate and the wave-breaking field scale with the (electron-dominated) plasma frequency ω_p, from the ratio

$$\frac{\omega_i}{\omega_p} = \left(\frac{mZ_i}{M_i}\right)^{1/2}, \tag{3.71}$$

we observe that, for given plasma density, the time available for plasma wave excitation may be increased simply by (singly) ionizing a gas comprised of a higher-mass atomic species. With heavier ions, one has more time to excite larger fields with the same laser power, or comparable fields with less laser power before ion motion disrupts coherence. In addition to ionic instabilities, laser instabilities are also of concern. Raman backscatter, with growth rate $\Gamma_{\rm RBS} \sim \frac{1}{2}\bar{a}\sqrt{\bar{\omega}\omega_p}$ and Raman forward scatter, with growth rate $\Gamma_{\rm RFS} \sim \bar{a}\frac{\omega_p}{\bar{\omega}}\frac{1}{\sqrt{8}\gamma_0}$, as well as related 2D self-focusing or modulational instabilities [52], are almost certainly not negligible for parameter regimes of interest, and their effects should be investigated in a more realistic simulations. There is some hope even that Raman scattering might be turned to our advantage: it may also be possible to exploit autoresonant excitation in counter-propagating or colliding-pulse schemes [81, 82, 83], where the enhanced growth of the plasma waves associated with Raman scatter may allow shorter excitation times for fields approaching the wave-breaking limit.

Strictly speaking, our simplified dynamical model has demonstrated the robustness of autoresonant excitation only with respect to global density mismatches, and to small

changes in the global chirp rate, drive intensity, or initial detuning. But we believe that some of this robustness should persist in the presence of moderate spatio-temporal variations and non-uniformity within the plasma and of complex spatio-temporal structure and dynamics in the lasers, features which are not captured by the idealized model used here. Effects such as ionization dynamics, thermal fluctuations, hydrodynamic expansion, electron and ion plasma instabilities, ponderomotive blow-out, and of course the nonlinear plasma excitation itself all lead to highly nontrivial plasma density profiles that vary in time and space and change the local conditions for resonance. Likewise, diffraction, tunneling ionization and the resulting ionization-induced depletion, refraction, and blue-shifting, as well as nonlinear focusing effects due to inhomogeneities and nonlinear back-action of the plasma wave on the EM fields, as well as Raman, self-modulation, and other instabilities, all can lead to appreciable distortion of the driving laser fields during the excitation process. The importance of these details for autoresonant wake excitation need to be further studied in the context of more detailed laser-plasma models.

Any such variation, inhomogeneity, or complex dynamical structure in the evolving laser fields or plasma is likely to prove deleterious in the absence of autoresonance with adiabatic passage through resonance, drastically reducing wake excitation and uniformity. Longitudinal fields in some spatial regions may be resonantly excited to moderate or large amplitudes while other regions may be mismatched in density and experience very little excitation, or may be first excited and then subsequently de-excited as a result of phase slippage and energy exchange back to the lasers. As a result,we expect the final plasma wave to be a highly irregular accelerating structure. But our autoresonant approach enjoys an intrinsic insensitivity and persistence, due to the local nature of the phase-locking. Provided only that the magnitude and scales for the non-uniformity are such as to allow an eikonal treatment of the waves, we expect [64] that at each position, the local plasma response will be autoresonantly excited by the drive, based on the local, slowly-varying values of the plasma frequency, drive amplitude, beat frequency, and chirp rate. Once phase-locked, the local Langmuir wave will grow monotonically to large amplitude, then saturate with very little ringing. Not all spatial regions will reach precisely the same final amplitude, but eventually the wave can grow monotonically more or less everywhere until local saturation ensues, so the final variations should be considerably less than in standard approaches. Although plausible, given what has been demonstrated in previous analyses of autoresonant phenomena, this expectation should also be verified in more realistic simulations.

We have also neglected all thermal effects, but, in realistic situations, tunneling ion-

ization by and inverse bremsstrahlung of the drive lasers will typically result in plasma temperatures of $T_e \gtrsim O(10\text{ eV})$. Damping of the high phase-velocity plasma waves should remain small, but thermal effects can lead to increased particle-trapping and lower thresholds for wave-breaking as well as induce changes in the Langmuir dispersion relation, while also resulting in some background of random short-scale density fluctuations and electrostatic oscillations which may impact the initial stages of autoresonant phase-locking, which assumes a suitably quiescent plasma. This too may be further investigated via numerical simulation.

The suitability of the excited plasma waves for relativistic particle acceleration depends not only on the magnitude and uniformity of the peak electric fields fields but even more crucially on the uniformity of the phase and phase velocity, and on the degree of phase-locking to the external drive. The Langmuir phase velocity v_p approximately matches the laser group velocity v_g, so first one should consider variability in the latter. Diffractive effects can substantially lower the longitudinal group velocity of the laser [52], but once accounted for, additional variation due to the finite bandwidth $\delta\omega \sim \omega_p$ are typically quite small in underdense plasmas, $\frac{\delta v_g}{v_g} \sim \frac{\omega_p^2}{\bar{\omega}^2}\frac{\delta\omega}{\bar{\omega}}$, as are variations due to moderate density variations: $\frac{\delta v_g}{v_g} \sim \frac{\omega_p^2}{\bar{\omega}^2}\frac{\delta n}{n_0}$. The relative change in the dephasing length L_d (for relativistic electrons with initial velocities very near c) is correspondingly small: $\frac{\delta L_\mathrm{d}}{L_\mathrm{d}} \sim \frac{\delta v_g}{v_g}$.

With variations in the group velocity v_g expected to be small, phase coherence of the plasma wave will depend on how closely v_p follows the essentially constant $v_g = \bar{v}_g$ of the laser. Particle-in Cell (PIC) simulations of Filip et al. [80] suggest that for the RL scheme, the effective phase velocity of the nonlinear plasma wave (measured in terms of the progression of the field maximum) can vary appreciably, i.e., 10% to 20%, reflecting phase slippage of the Langmuir wave primarily as a result of ponderomotive blowout and relativistic detuning, while the plasma wave produced in their non-resonant PBWA scheme exhibits substantially less phase slippage. By working within the QSA, in 1D geometry without transverse density variation, we cannot independently assess any such slippage effects for the present scheme, but we anticipate that it will be similarly small by virtue of the autoresonant phase-locking. That is, in the resonant RL/TD scheme, plasma inhomogeneities can lead to accelerating buckets that have a changing phase, so that electrons do not experience a constant accelerating field. With autoresonance, however, the laser can phase lock to a range of densities, creating an accelerating field of uniform phase that is everywhere directly related to the local phase of the driving beat wave.

An appealing feature of both the non-resonant PBWA and autoresonant PBWA is this

ostensible ability to phase-lock the plasma wave to the beat-wave of the applied drive lasers, with the implied hope that the electron injection, whether based on external cathode injection [84] or internal optical injection [85, 86] can also be phase-locked to the same lasers. Because of its potential importance, this phase-locking deserves careful investigation. An obvious worry with the non-resonant PBWA is that entrainment is achieved by brute force, in contrast to our self-trapping using adiabatic passage through resonance. Because of the intrinsic inefficiency of non-resonant forcing, it must rely on large driving fields and denser plasma to achieve large accelerating fields. But increasing ω_p and a^2 also increases group-velocity dispersion, which can inhibit phase-locking to an externally-known reference phase, and increases the growth rates for Raman and self-focusing instabilities which can modulate the laser envelope. One runs the risk of turning the non-resonant PBWA into a self-modulated LWFA. That is, such modulation can actually enhance the production of plasma waves, but the plasma response can then become entrained not to the initial laser envelope as applied to the plasma, with a prescribed shape and phase, but to the envelope after it has undergone some uncontrolled nonlinear evolution and modulation.

One must also carefully distinguish phase-locking from frequency-locking, however much all forms of oscillatory entrainment tend to be conflated under the former name. Perfect phase-locking implies perfect frequency-locking, and conversely (at least up to some constant but perhaps unknown phase), but in the case of only partial or imperfect entrainment, it is possible to achieve good frequency-locking without adequate phase-locking, or the converse. Whether the relative error in the phase-matching or in frequency-matching between driving and driven oscillation is greater depends on whether the Fourier content of fluctuations in the phase is primarily at higher or lower frequencies than the drive frequency itself.

For the purpose of matched particle injection, it is phase-locking which is desired, yet some caution is warranted in claims of true phase-locking in either the non-resonant or autoresonant PBWA. In any frequency-locked PBWA scheme, the nonlinear frequency of the Langmuir wave may be closely entrained to the precisely known drive frequency, but this does not necessarily imply that the absolute phase of the Langmuir wave may be precisely known. As a mechanism for phase-locking in the non-resonant PBWA, the authors appeal to the claim that an harmonic oscillator, strongly driven off resonance, remains synchronous with the driving force, and then suggest that this should extend to nonlinear oscillators. But this intuition holds only for damped linear oscillators after the transient is allowed to decay. If a linear oscillator, with natural frequency ω_n, and negligible damping on the observational time-scales, is forced by a constant-amplitude, sinusoidal drive with frequency

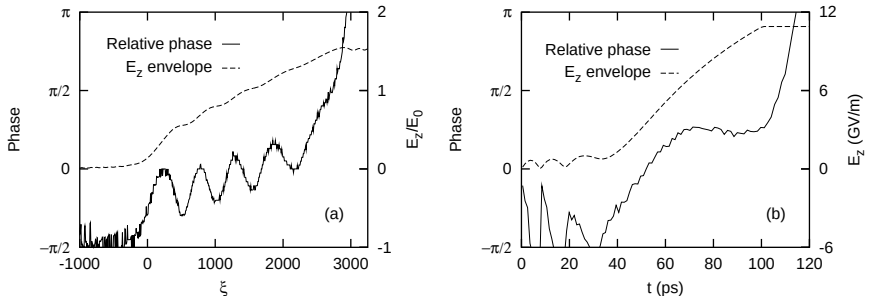

Figure 3.6. Evolving phase difference between the maxima of the laser beat and the longitudinal electric E_z (which may differ from Ψ in the text by a constant). Plot (a) has $\epsilon = 0.005$, $\alpha = 0.0001$, with the phase slowly evolving during autoresonance for $0 < \xi < 2700$, where E_z saturates. (b) uses the CO_2 laser parameters of Fig. 3.3, which demonstrates nearly constant phase locking during autoresonant excitation for $40 < \xi < 100$, permitting good control of electron injection.

$\omega_d < \omega_n$, then, independent of the strength of the drive, the driven oscillator will exhibit persistent, oscillating variations in phase relative to the drive phase $\psi_d = \omega_d t + \psi_d(0)$ which are $\sim O(\omega_d/\omega_n)$, i.e., only first-order in their frequency ratio. There is no obvious reason to expect in general that nonlinearities will improve matters.

In the autoresonant case, we may also encounter seemingly good frequency-locking either with comparably good or disappointingly inferior phase-locking, as shown in Fig. 3.6, where we plot, for two of our previous examples, the phase lag between the beat-wave of the lasers and the plasma wave, numerically estimated by the offset between corresponding relative maxima. Fig. 3.6(a) exhibits oscillating phase variation which is slowly varying compared to ω_p, but of order $\lesssim \pi/2$ in absolute (not relative) magnitude, even within the fully nonlinear regime of autoresonance prior to saturation. On any one shot, an injected electron bunch, if sufficiently relativistic (with electron trajectories $\xi_e(\tau) \approx \xi_e(0)$) and sufficiently short (bunch length $\Delta z_e \ll c\beta_p/c$), would experience uniform acceleration and little additional energy spread. But the exact phase of the maximal accelerating field could not be reliably predicted from one shot to the next, so there would be considerable shot-to-shot variation in average energy gain of the bunch and no way to reproducibly exploit the peak field, even if the arrival time of the bunch was perfectly correlated with the known beat-phase of the drive lasers in vacuum.

In contrast, in the second case of Fig. 3.6(b), as the oscillator passes through resonance, the plasma wave phase tightly locks with the drive phase and remains so until adiabaticity is lost and the excitation saturates. In this case, phase-locked injection scheme could reliably place electron bunches near the maximum acceleration gradient. It appears that the extent

of this phase error is determined by how deeply the oscillator phase is trapped in its effective potential, which improves with greater drive strength, slower chirp rate, and initial detuning further above resonance.

The most obvious drawback to the autoresonant APTR scheme, in comparison to the standard RL/TD approach, is the added complication of a chirped drive. But the total chirp required is modest, on the order of only ω_p in order to guarantee, in the face of some uncertainty as to the density, adiabatic passage through resonance from above. In typical strongly underdense plasmas this corresponds to only a few percent of the laser carrier frequency. In solid-state CPA laser systems, this should be relatively easy to achieve, because the laser pulse already undergoes optical stretching/chirping and compression/dechirping in a series of gratings, and the gain bandwidth of the crystals is intrinsically large ($\sim$ 20% near $\lambda \sim 1$ μm). One need only adjust the CPA optics so as to only partially re-compress the final chirped, amplified pulse, in order to leave some residual chirping.

In CO_2 lasers (or other gas laser systems), the gain bandwidths tend to be narrower, but can effectively be increased by operating at sufficiently high pressures to doppler-broaden the rotational lines into an overlapping quasi-continuum. If the final intensities and fluences needed are below the damage threshold for the required optics, one can simply add a pair of gratings similar to those used in CPA systems in order to chirp one of the amplified beams [87]. If the final intensity is too large, then with somewhat more difficulty one might arrange a multi-stage system, where an initial seed pulse is chirped in this manner before passing through the final amplifier stage, involving a gas under sufficient pressure to cover the bandwidth of the chirped seed. More exotically, one might imagine using nonlinear optical effects in a gas or plasma cell to achieve the required frequency shifts.

3.7 Conclusions

We have introduced and analyzed a straightforward, seemingly minor, modification of the DMG scheme for the chirped-pulse PBWA, based on the nonlinear phenomenon of autoresonance with adiabatic passage through resonance (APTR), which nevertheless enjoys certain advantages over previous approaches. Rather than starting at the linear resonance and chirping downward at some intermediate rate expected to match, on average, the beat frequency to the plasma wave frequency corresponding to the growing value of the plasma wave amplitude, we start above resonance and sweep the beat frequency downward past the resonance, at any sufficiently slow chirp rate, such that the plasma wave frequency beat

frequency automatically self-locks to the drive frequency, and the plasma wave amplitude automatically adjusts itself consistent with this frequency.

This new scheme is designed to overcome some of the well-known limitations of previous approaches, namely relativistic detuning and nonlinear modulation or other non-uniformity or non-stationarity in the driven Langmuir wave amplitude, and sensitivity to frequency mismatch due either to measurement uncertainties or to true density fluctuations and inhomogeneities. As in previous schemes, modulational instabilities of the ionic background ultimately limit the useful interaction time, but nevertheless peak electric fields at or approaching the wave-breaking limit seem readily attainable. Compared to traditional approaches, the autoresonant scheme achieves larger accelerating electric fields for given laser intensity, or comparable fields for less laser power; the plasma wave excitation appears to be much more robust to inevitable uncertainties and variations in plasma and laser parameters; it is largely insensitive to the precise choice of chirp rate, provided only that chirping is sufficiently slow; and the quality and uniformity of the resulting plasma wave and its suitability for accelerator applications may be superior.

In underdense plasmas, the total frequency shift required is only of the order of a few percent of the laser carrier frequency, and for possible experimental proofs-of-principle, the scheme might be implemented with relatively little additional modification to existing systems based on either solid-state amplifiers and Chirped Pulse Amplification techniques, or, with somewhat greater technological effort, using a CO_2 or other gas laser. Preliminary analysis has been performed within a simplified analytic and numerical model, and wake excitation has been studied using realistic parameters for realistic Ti:Sapphire and CO_2 laser systems. The results are very encouraging, and warrant extending investigation to higher-dimensional geometries, more realistic plasma inhomogeneities, and self-consistent laser evolution, via numerical solution with fluid and particle-in-cell (PIC) codes.

Acknowledgements

The research reported throughout this chapter was pursued in particularly close collaboration with R.R. Lindberg, and grew out of ideas on potential laser-plasma-based applications of the principle of autoresonance whose formulations and explanations have been pioneered by Lazar Friedland. We also benefited greatly from personal interactions with Professor Friedland while he visited U.C. Berkeley on sabbatical from the Racah Institute of Physics at the Hebrew University in Jerusalem.

Chapter 4

Hilbert-Space Variational Principle for Spontaneous Wiggler and Synchrotron Radiation

Mehr Licht!
(More light!)

Johann Wolfgang Von Goethe
(attributed last words)

4.1 Introduction

Variational principles are ubiquitous in electromagnetism [88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98]. Thomson's and Dirichlet's Principles allow the determination of potentials or capacitances in electrostatics, or the determination of steady-state currents and voltages in electrical networks. General reciprocity relations lead to variational estimates for eigenmodes of cavities or waveguides, impedances of cavities, apertures, or other structures, and even certain scattering coefficients. In magnetohydrodynamics, plasma kinetic theory, and certain equilibrium or non-equilibrium thermodynamic situations, stability or dynamical accessibility of charged particle distributions has been derived from minimum energy/free energy, maximum entropy, or minimum entropy production or minimum heat production principles [99, 100, 101, 102, 103, 104, 105, 106, 107]. Density functional theory, an elaboration of the famous Thomas-Fermi and Hartree-Fock methods in quantum mechanics,

has also been applied to classical and quantum Coulomb gases [108, 109]. Many numerical methods and approximation techniques [110, 111, 112, 95, 113, 114, 115] for mechanical or manual computation, associated with the names Raleigh, Ritz, Galerkin, Finite Elements, Minimum Residuals, Method of Moments, etc., may all be derived from or interpreted as variational principles. As is well known, all of ray optics may be derived from Fermat's Principle of Least Time, and, ultimately, all of classical electrodynamics may be developed via Hamilton's Principle, a variational formulation demanding stationarity of the action functional.

Variational techniques or approaches enjoy many advantages. They provide unified theoretical treatments and compact mathematical descriptions of many physical phenomena, which allow for straightforward change of coordinates and incorporation of additional constraints. Conventional equations of motion and conservation laws may be derived in a concise manner. Connections between classical and quantum limits are often more readily apparent. Variational principles often suggest appealing physical interpretations of the behaviors governed by them. They provide a unified, compact framework for performing eikonal, ponderomotive/oscillation-center, or other averaging techniques. Perhaps most importantly from a practical standpoint, they offer starting points for efficient approximation or numerical computation, whereby complicated partial differential equations or integro-differential equations may be replaced with more tractable mathematical beasts, namely quadratures, ordinary differential equations, systems of algebraic or even linear equations, or ordinary function minimization. Although approximate, these simplified, projected, or parameterized variational solutions may be more easily interpreted or more easily computed or offer more insight than the exact forms.

Here, we derive another principle, which we will call the *Maximum-Power Variational Principle* (MPVP), together with some surrounding mathematical formalism. Although relatively simple in its statement and scope, somewhere between intuitively plausible and obvious, depending upon one's point of view, the MPVP may be of some use in the classical theory of radiation, in particular in the analysis of light sources relying on spontaneous radiation from relativistic electron beams in undulators or other magnetic insertion devices, or more generally, for the approximation of features of other forms of synchrotron or "magnetic Bremsstrahlung" radiation. After some suitable generalization, these ideas may also be applicable to the cases of Čerenkov, transition, wave-guide/cavity, Smith-Purcell, or other types of radiation emitted by classical particle beams traveling through certain metallic or dielectric media or structures.

This work arose out of a recent treatment of coherent X-ray generation via a harmonic cascade in an electron beam which radiates in the low-gain regime while traveling through a series of undulators. Penn, *et al.* [116, 117, 118, 119, 120] approximate the modal structure of the radiation in terms of a paraxial mode described by certain adjustable parameters. Some of these parameters are subsequently determined directly by dynamical considerations, but some remain free, and are determined at the end of the calculation under the assumption that the "proper" or "natural" criterion is to choose the values of these parameters which maximize the resulting radiated power assuming that mode shape. This procedure seems so physically reasonable, appealing, and plausible that intuition suggests that, in some sense, it must be the correct approach under the circumstances. However, our intuition failed to immediately suggest a precise derivation or rigorous justification of this power-maximization idea. Then again, it does lead us to expect that the governing principles or concepts should be quite fundamental, and applicable to more general radiation problems.

In this chapter we pursue these ideas, developing various Hilbert-space approaches to the spontaneous radiation from prescribed current sources, culminating in the Maximum-Power Variational Principle (MPVP) applicable to such problems under very general assumptions. Although the final results are quite simple and intuitive, the mathematics are developed in some detail in a deliberate and pedagogical fashion; much can probably be skipped with impunity by all but the most interested readers.

We first consider the approximate but commonly-encountered case of paraxial fields, where the structure and dynamics of the fields and the development of the mathematical results are greatly simplified. Motivated by the formal parallels between the Schrödinger equation in non-relativistic quantum mechanics and the paraxial wave equation of classical physical optics, we present an elementary derivation of a convenient Hilbert-space formalism for the spontaneous radiation produced by prescribed classical sources in the paraxial limit, including a derivation of the MPVP. Guided by these results, we then employ Green function and spherical wave treatments of the general three-dimensional radiation fields to generalize these results to the case of non-paraxial fields propagating in free space. Specifically, we have in mind applications to the spontaneous synchrotron emission from a reasonably well-collimated relativistic electron bunch in a wiggler or similar insertion device, although the results remain valid more generally.

By *spontaneous* emission, we mean that the trajectories of the charged particles constituting the source for the radiation are considered prescribed functions of time, *independent* of the actual radiation fields generated. That is, the particle trajectories are assumed to

be determined by initial conditions, externally applied wiggler or other guiding fields, and possibly even space-charge effects (quasi-static self-fields, either exact including collisions, or in a mean-field/Vlasov approximation), while any back-action of the radiation itself on the particles – either recoil, absorption, or multiple scattering effects – is neglected. Thus the MPVP provides an approximate alternative to calculation of the fields via the usual Liénard-Weichart potentials or related expressions of Jeffimenko, Feynman, Heaviside, or others [92].

For the case of a wiggler or undulator, this implies that any gain due to ponderomotive feedback and dynamic bunching remains small, so the resulting self-consistent stimulated emission component of the radiation can be neglected compared to the spontaneous component. The spontaneously-emitted radiation and its measurable properties (power, angular distribution, coherence, etc.) can then in principle be expressed as deterministic functions of the incoming beam phase space profile (including emittance effects and any pre-bunching) and the prescribed external fields, without having to solve the equations of motion including the radiation to obtain completely self-consistent trajectories for the particles.

Although the radiation is treated classically, this terminology involving spontaneous/stimulated emission is standard in the Free Electron Laser (FEL) literature, which, of course, intentionally borrowed it from the quantum theory of lasers. The *stimulated* emission is that which requires (for ponderomotive bunching) the presence of both the static wiggler field and additional radiation, either from an external seed (FEL amplifier), from emission earlier in the same wiggler (SASE, or Self-Amplified Spontaneous Emission), or from a previous pass through the wiggler in the presence of a resonant cavity or mirror system (conventional FEL oscillator). The *spontaneous* emission is, by definition, that which occurs in the absence of additional (applied) radiation fields, but of course it still requires the static wiggler field. Truly "free" electrons do not radiate because they are not accelerated, and in fact energy-momentum considerations would forbid even a solitary accelerated electron from radiating if the interaction providing the acceleration did not also allow for a momentum exchange with some external matter (ultimately the wiggler magnets or other electrons, in the present case).

Therefore, the description of the wiggler radiation as spontaneous is similar to the atomic case in that it occurs in the absence of applied radiation (i.e., real photons), but it differs in that electrons bound in atoms, if somehow excited, can then radiate in the absence of any external fields. The static (i.e., virtual photon) wiggler field is therefore playing a role analogous to the nuclear Coulomb field. In the average rest frame of a

sufficiently relativistic beam, the Weizsacker-Williams method of virtual quanta may be used, since the Lorentz-transformed static wiggler field begins to resemble an oncoming beam of actual photons. Spontaneous radiation occurs when a backwardly-moving, virtual wiggler photon is Compton-scattered by an electron into a forward-moving, real photon. Stimulated emission occurs in a "three-body collision" where an electron simultaneously scatters a virtual wiggler photon and real radiation photon into two real photons, both in the same mode as the incident photon.

Throughout this analysis, we also assume that the charge and current densities are not only prescribed, but remain *localized* in space (so that the far-field, or wave or radiation zone, is defined) and time (so that certain Fourier transforms exist), in a manner more precisely specified in the course of our derivations. We could generalize our treatment to additionally include the effects of a uniform background conductive or dielectric media, but for simplicity we here assume that, apart from its generation by the prescribed microscopic sources in a bounded spacetime region, the emitted radiation otherwise propagates in vacuum. Generalizations to allow for non-uniform dielectric tensors, representing wave-guides, lenses, windows, or other optical devices, might also be possible, but are not pursued here. As we are interested in the causal (outgoing) radiation produced by the given sources, we neglect any incident source-free fields or fields from other, remote sources. Because we are working in the fully linear, spontaneous limit, any such fields may be added in superposition at the very end of the calculation.

4.2 Fundamental Equations

Throughout this chapter, we denote in boldface type any real or complex vectors, e.g., $\boldsymbol{a}$ or $\boldsymbol{b}$, and denote with carets any real or complex unit vectors such as $\hat{\boldsymbol{a}}$ which satisfy

$$\|\hat{\boldsymbol{a}}\| \equiv (\hat{\boldsymbol{a}}^* \cdot \hat{\boldsymbol{a}})^{1/2} = 1, \tag{4.1}$$

where for any N-dimensional, complex-valued vectors $\boldsymbol{a}$, $\boldsymbol{b} \in \mathbb{C}^N$,

$$\boldsymbol{a} \cdot \boldsymbol{b} = \sum_{n=1}^{N} a_n b_n \tag{4.2}$$

is the usual Euclidean dot product (without any implicit complex conjugation) of the components of $\boldsymbol{a}$ and $\boldsymbol{b}$ in some orthogonal basis. Additional notation will be introduced as needed.

Conforming to a widespread, if not universal, convention in beam physics, we work in the (non-covariant) Coulomb, transverse, or radiation gauge, where the vector potential $\boldsymbol{A} = \boldsymbol{A}_\perp = \boldsymbol{A}(\boldsymbol{x}, t)$ is chosen to be purely solenoidal (i.e., divergenceless), or everywhere transverse in the sense of Helmholtz's Theorem:

$$\boldsymbol{\nabla} \cdot \boldsymbol{A}(\boldsymbol{x}, t) = 0. \tag{4.3}$$

Here, $\boldsymbol{x} \in \mathbb{R}^3$ denotes the three-dimensional position and $t \in \mathbb{R}$ the time, in any one convenient Lorentz frame (typically the lab frame, or possibly the average electron beam frame, for wiggler radiation problems). In Gaussian units, the scalar potential $\Phi = \Phi(\boldsymbol{x}, t)$ then satisfies the instantaneous (i.e., unretarded) Poisson's equation

$$\nabla^2 \Phi(\boldsymbol{x}, t) = -4\pi\rho(\boldsymbol{x}, t) \tag{4.4}$$

given the charge density ρ, while the vector potential the vector potential $\boldsymbol{A}$ satisfies the inhomogeneous wave equation

$$\nabla^2 \boldsymbol{A}(\boldsymbol{x}, t) - \tfrac{1}{c^2} \tfrac{\partial^2}{\partial t^2} \boldsymbol{A}(\boldsymbol{x}, t) = -\tfrac{4\pi}{c} \boldsymbol{J}_\perp(\boldsymbol{x}, t), \tag{4.5}$$

where c is the speed of light *in vacuo*, and $\boldsymbol{J}_\perp$ is the transverse (i.e., solenoidal, or divergenceless) component of the full current density $\boldsymbol{J}$, and which, using (4.4) and local charge conservation, can be shown to be given by

$$\boldsymbol{J}_\perp(\boldsymbol{x}, t) = \boldsymbol{J}(\boldsymbol{x}, t) - \boldsymbol{J}_\parallel(\boldsymbol{x}, t) = \boldsymbol{J}(\boldsymbol{x}, t) - \tfrac{1}{4\pi} \boldsymbol{\nabla} \tfrac{\partial}{\partial t} \Phi(\boldsymbol{x}, t). \tag{4.6}$$

The physical fields are, as usual, related to the potentials by

$$\boldsymbol{E} = \boldsymbol{E}_\perp + \boldsymbol{E}_\parallel = -\tfrac{1}{c} \tfrac{\partial}{\partial t} \boldsymbol{A} - \boldsymbol{\nabla} \Phi, \tag{4.7}$$

and

$$\boldsymbol{B} = \boldsymbol{B}_\perp = \boldsymbol{\nabla} \times \boldsymbol{A}. \tag{4.8}$$

Note again that we are describing only the fields generated by the prescribed sources ρ and $\boldsymbol{J}$. By assumption, any externally applied fields (due ultimately to some additional, remote sources) such as undulator or bend magnets are not included in the potentials $\boldsymbol{A}$ or ϕ, but their effects on the trajectories of the charge particles constituting the actual radiation sources are assumed to be either already included in $\boldsymbol{J}(\boldsymbol{x}, t)$ and $\rho(\boldsymbol{x}, t)$, or else are neglected.

Without some prior expectation for the typical magnitudes of the dominant radiation frequency ω_{rad} or characteristic transverse spot size w_0 of a radiation beam, it is convenient

to re-scale the space-time coordinates using the remaining spatio-temporal scales, namely those determined by the speed of light c and the plasma frequency ω_p associated with the characteristic charge density of the known source. That is, choosing some reference number density $\bar{n}$, ideally characteristic of that of the actual particle bunch, and defining the corresponding linear plasma frequency

$$\omega_p = \left(\frac{4\pi \bar{n} e^2}{m_e} \right)^{1/2}, \tag{4.9}$$

were m_e is the electron mass and e the magnitude of the electron charge, we define normalized (dimensionless) variables: a normalized time $\tau = \omega_p t$, normalized longitudinal position $\xi = \frac{\omega_p}{c} z$, normalized transverse coordinates $\boldsymbol{r} = r_x \hat{\boldsymbol{x}} + r_y \hat{\boldsymbol{y}} = \frac{\omega_p}{c} \left(x \hat{\boldsymbol{x}} + y \hat{\boldsymbol{y}} \right)$, normalized three-dimensional coordinates $\boldsymbol{\zeta} = \xi \hat{\boldsymbol{z}} + \boldsymbol{r}$, the normalized vector potential $\boldsymbol{a}(\boldsymbol{\zeta}; \tau) = \boldsymbol{a}(\boldsymbol{r}, \xi; \tau) = \frac{e}{mc^2} \boldsymbol{A}(\boldsymbol{x}, t)$, normalized scalar potential $\phi(\boldsymbol{\zeta}; \tau) = \phi(\boldsymbol{r}, \xi; \tau) = \frac{e}{mc^2} \Phi(\boldsymbol{x}, t)$, normalized charge density $\mu(\boldsymbol{\zeta}; \tau) = \mu(\boldsymbol{r}, \xi; \tau) = \frac{1}{e\bar{n}} \rho(\boldsymbol{x}, t)$, and a normalized current density $\boldsymbol{j}(\boldsymbol{\zeta}; \tau) = \boldsymbol{j}(\boldsymbol{r}, \xi, \tau) = \frac{1}{e\bar{n}c} \boldsymbol{J}(\boldsymbol{x}, t)$, with transverse (solenoidal) part $\boldsymbol{j}_\perp = \frac{1}{e\bar{n}c} \boldsymbol{J}_\perp = \boldsymbol{j} - \left(\hat{\boldsymbol{z}} \frac{\partial}{\partial \xi} + \boldsymbol{\nabla}_\perp \right) \frac{\partial}{\partial \tau} \phi$, where $\boldsymbol{\nabla}_\perp \equiv \frac{\partial}{\partial \boldsymbol{r}} = \hat{\boldsymbol{x}} \frac{\partial}{\partial r_x} + \hat{\boldsymbol{y}} \frac{\partial}{\partial r_y}$ is the scaled, (geometrically) transverse gradient operator, and we also define $\boldsymbol{\partial} = \frac{\partial}{\partial \boldsymbol{\zeta}} = \hat{\boldsymbol{z}} \frac{\partial}{\partial \xi} + \boldsymbol{\nabla}_\perp$ as the scaled gradient operator in the full three-dimensional space.

In these variables, Poisson's equation becomes

$$\partial^2 \phi(\boldsymbol{\zeta}; \tau) = \left(\frac{\partial^2}{\partial \xi^2} + \nabla_\perp^2 \right) \phi(\boldsymbol{r}, \xi; \tau) = \mu(\boldsymbol{r}, \xi; \tau), \tag{4.10}$$

where $\nabla_\perp^2 = \boldsymbol{\nabla}_\perp \cdot \boldsymbol{\nabla}_\perp$ is the scaled (geometrically) transverse Laplacian, and $\partial^2 = \boldsymbol{\partial} \cdot \boldsymbol{\partial}$ is the scaled three-dimensional Laplacian. The solutions to this equation just consist of the well-known unretarded (i.e., instantaneous) Coulomb fields,

$$\phi(\boldsymbol{\zeta}; \tau) = -\frac{1}{4\pi} \int d^3 \boldsymbol{\zeta}' \, \frac{\mu(\boldsymbol{\zeta}', \tau)}{\| \boldsymbol{\zeta} - \boldsymbol{\zeta}' \|} \tag{4.11}$$

which fall off like the inverse-square distance from the instantaneous positions of the charges (i.e., as $O(\frac{1}{\| \boldsymbol{\zeta} - \boldsymbol{\zeta}' \|^2})$), so they do not contribute to the radiation fields, and can be neglected for our purposes. (These unretarded Coulomb fields are "attached" to the charges and move with them even under acceleration, so do not include or in any way constitute radiation.) The scaled wave equation for the vector potential may be written as

$$\left(\partial^2 - \frac{\partial^2}{\partial \tau^2} \right) \boldsymbol{a}(\boldsymbol{\zeta}; \tau) = \left(\frac{\partial^2}{\partial \xi^2} + \nabla_\perp^2 - \frac{\partial^2}{\partial \tau^2} \right) \boldsymbol{a}(\boldsymbol{r}, \xi; \tau) = -\boldsymbol{j}_\perp(\boldsymbol{r}, \xi; \tau), \tag{4.12}$$

where $\boldsymbol{a}$ also satisfies the gauge condition

$$\boldsymbol{\partial} \cdot \boldsymbol{a}(\boldsymbol{\zeta}; \tau) = \left(\hat{\boldsymbol{z}} \frac{\partial}{\partial \xi} + \boldsymbol{\nabla}_\perp \right) \cdot \boldsymbol{a}(\boldsymbol{r}, \xi; \tau) = 0. \tag{4.13}$$

All of what we conventionally consider radiation is associated with the fields derived from this vector potential, although these fields also include (in the near-field and intermediate field regions) non-radiative components needed to cancel the non-retarded features of the Coulomb fields, as well as properly-retarded but non-radiative "velocity" fields.

To separately analyze each spectral component of the radiation, we perform Fourier transforms in (scaled) time, obtaining, for each (scaled) frequency ω, the frequency-domain wave equation

$$\left(\partial^2 + \omega^2\right) \boldsymbol{a}(\boldsymbol{r}, \xi; \omega) = \left(\frac{\partial^2}{\partial \xi^2} + \nabla_\perp^2 + \omega^2\right) \boldsymbol{a}(\boldsymbol{r}, \xi; \omega) = -\boldsymbol{j}_\perp(\boldsymbol{r}, \xi; \omega), \tag{4.14}$$

and frequency-domain gauge condition

$$\boldsymbol{\partial} \cdot \boldsymbol{a}(\boldsymbol{\zeta}; \omega) = \left(\hat{\boldsymbol{z}} \frac{\partial}{\partial \xi} + \boldsymbol{\nabla}_\perp\right) \cdot \boldsymbol{a}(\boldsymbol{r}, \xi; \omega) = 0, \tag{4.15}$$

where

$$\boldsymbol{a}(\boldsymbol{\zeta}; \omega) = \frac{1}{\sqrt{2\pi}} \int d\tau \, e^{i\omega\tau} \boldsymbol{a}((\boldsymbol{\zeta}; \tau) \tag{4.16}$$

and

$$\boldsymbol{j}_\perp(\boldsymbol{\zeta}; \omega) = \frac{1}{\sqrt{2\pi}} \int d\tau \, e^{i\omega\tau} \boldsymbol{j}_\perp(\boldsymbol{\zeta}; \tau) \tag{4.17}$$

denote the usual Fourier transforms in scaled time, and are in general complex-valued. Because the physical, time-domain fields are real-valued, we necessarily have $\boldsymbol{a}(\boldsymbol{\zeta}; -\omega) = \boldsymbol{a}(\boldsymbol{\zeta}; \omega)^*$, so we can restrict attention to the analytic signal, or positive frequency ($\omega > 0$), components of the radiation fields. We assume that the source charge density ρ and current density $\boldsymbol{j}$ are at least weakly localized in time, to the extent that all needed Fourier transforms of the sources and the resulting fields actually exist (at least as generalized functions).

Corresponding to the scaled, frequency-domain vector potential $\boldsymbol{a}(\boldsymbol{\zeta}; \omega)$, we define the scaled solenoidal electric field:

$$\boldsymbol{\varepsilon}_{\boldsymbol{a}} = \boldsymbol{\varepsilon}_{\boldsymbol{a}}(\boldsymbol{\zeta}; \omega) = i\omega \, \boldsymbol{a}(\boldsymbol{\zeta}; \omega); \tag{4.18}$$

and the scaled frequency-domain magnetic field

$$\boldsymbol{b}_{\boldsymbol{a}} = \boldsymbol{b}_{\boldsymbol{a}}(\boldsymbol{\zeta}; \omega) = \partial \times \boldsymbol{a}(\boldsymbol{\zeta}; \omega). \tag{4.19}$$

The subscript may be dropped when it is clear from which vector potential these fields are derived. For later convenience, we also define the scaled, frequency-domain Poynting vector

$$\boldsymbol{s}_{\boldsymbol{a}}(\boldsymbol{\zeta}; \omega) = \boldsymbol{\varepsilon}_{\boldsymbol{a}}(\boldsymbol{\zeta}; \omega) \times \boldsymbol{b}_{\boldsymbol{a}}(\boldsymbol{\zeta}; \omega)^* \tag{4.20}$$

associated with the *solenoidal* fields only. For any orientable surface Σ with unit outward normal $\hat{\boldsymbol{n}}$, and positive frequency interval $0 < \omega_0 \leq \omega \leq \omega_1$, the quantity

$$\mathcal{P}[\boldsymbol{a}](\Sigma;\omega_0,\omega_1) \equiv \int\limits_{\omega_0}^{\omega_1} d\omega \ \mathrm{Re}\Big[\int\limits_{\Sigma} d^2\sigma \, \hat{\boldsymbol{n}}\cdot\boldsymbol{s}_{\boldsymbol{a}}\Big] \tag{4.21}$$

may be taken as the scaled, time-averaged (over an optical period) net electromagnetic power, in the bandwidth $[\omega_0,\omega_1]$, which passes outward through the surface Σ.

In order that the far-field, or radiation zone, even be well-defined, we assume that the physical source $\boldsymbol{j}(\boldsymbol{\zeta};\tau)$ remains localized in space throughout the relevant emission process, at least in the direction(s) of observed radiation propagation. In the general, three-dimensional case, this means that $\boldsymbol{j}(\boldsymbol{\zeta};\omega)$ is non-negligible only in a finite neighborhood of some fixed point $\boldsymbol{\zeta}_s$, which through a translation of coordinate axes generally will be taken to coincide with the origin. However, from (4.6) and (4.11), it then follows that the solenoidal component $\boldsymbol{j}_\perp(\boldsymbol{\zeta};\tau)$ will not be strictly localized, but must possess decaying tails which asymptotically fall off like $O(1/\|\boldsymbol{\zeta}-\boldsymbol{\zeta}_s\|^2)$ at sufficiently large distances from $\boldsymbol{\zeta}_s$. That is, $\boldsymbol{j}$ and $\boldsymbol{j}_\perp$ will not both be of compact support simultaneously. But the fields arising from any particular part of $\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\tau)$ will consist of the near-zone and intermediate zone fields which will fall off like like $O(1/\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|^2)$ or faster, as well as the true radiation fields which fall off more slowly, i.e. as $O(1/\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|)$. Because our interest resides in the latter, it follows that as long as $\boldsymbol{j}(\boldsymbol{\zeta};\tau)$ is appropriately localized, we can always choose some suitably large but finite region beyond which the contributions of the decaying tails of $\boldsymbol{j}_\perp$ to the radiation fields may be neglected to any desired level of accuracy, and we can then act as if both $\boldsymbol{j}(\boldsymbol{\zeta};\tau)$ and $\boldsymbol{j}_\perp(\boldsymbol{\zeta};\tau)$ have compact support. Said differently, in the estimates for the radiative fields, we can tolerate persistent errors of order $O(1/\|\boldsymbol{\zeta}-\boldsymbol{\zeta}_s\|^2)$, because by definition these errors correspond only to spurious near-zone or intermediate-zone fields.

4.3 Paraxial Case

By paraxial, we mean that the radiation from the electron bunch in the insertion device can be assumed to be highly mono-directional, in the sense that the relevant angular scales – namely any possible overall angular offset between the electron beam and optical axis of the insertion device, the angular spread of the particle beam, and the characteristic angle of diffraction – are assumed to be small, in a sense made precise below. If we restrict attention to radiation observed within a sufficiently small detection solid-angle, then radiation from relativistic particles in certain non-straight devices, i.e., synchrotron radiation

in circular rings, may also be treated within the paraxial approximation. The needed local directionality is provided by the assumed finite acceptance entendue of the detector and the characteristic $O(1/\gamma)$ angular spread in the radiation resulting from relativistic $O(1/\gamma^2)$ compression effects between emitter and observer times, where γ is the usual relativistic kinematic factor associated with the mean velocity of the source charges, which together limit the contributing portion of the electron trajectory to some small arc.

4.3.1 Paraxial Approximation to the Wave Equation

We have already seen how, in the Coulomb gauge, the scalar potential ϕ does not contribute to the radiation fields. The vector potential $\boldsymbol{a}$ contributes to near-field (quasi-static), intermediate (induction-zone), and far-field (radiation-zone) effects (and in fact part of it must cancel, in the resulting fields, the non-retarded effects of the scalar potential), but considerable simplification is possible if we effectively restrict our attention to those portions of the fields which, if allowed to evolve freely after propagating beyond the finite support of the sources, will actually survive into the far field and constitute radiation, and further assume that the radiation of interest propagates nearly along the $+\hat{\boldsymbol{z}}$ axis, with a small characteristic angle of diffraction $\theta_{\text{D}} \ll 1$ and any overall angular offset $\delta\theta_0 \lesssim O(\theta_d)$. Then we can make the so-called paraxial approximation to the wave-equation for the vector potential $\boldsymbol{a}$, which in effect consists of an asymptotic expansion in $\delta\theta_0$.

First, for any $\omega > 0$ of interest, we express the vector potential in enveloped form:

$$\boldsymbol{a}(\boldsymbol{r}, \xi; \omega) = \boldsymbol{\psi}(\boldsymbol{r}, \xi; \omega) e^{ik\xi}, \tag{4.22}$$

where the (scaled), axial carrier wavenumber $k = k(\omega) = +\omega$ satisfies the vacuum dispersion relation for a local plane wave traveling along $+\hat{\boldsymbol{z}}$. Similarly, we define an enveloped source term such that:

$$\boldsymbol{j}_\perp(\boldsymbol{r}, \xi; \omega) = 2k\, \boldsymbol{f}(\boldsymbol{r}, \xi; \omega) e^{ik\xi}. \tag{4.23}$$

Using the chain rule, the wave equation then becomes

$$\left(\tfrac{\partial^2}{\partial \xi^2} + 2ik \frac{\partial}{\partial \xi} + \nabla_\perp^2 \right) \boldsymbol{\psi}(\boldsymbol{r}, \xi; \omega) + 2k \boldsymbol{f}(\boldsymbol{r}, \xi; \omega) = 0, \tag{4.24}$$

while the radiation envelope $\boldsymbol{\psi}$ and source envelope $\boldsymbol{f}$ must both satisfy envelope-transversality conditions:

$$\left(\hat{\boldsymbol{z}} \tfrac{\partial}{\partial \xi} + ik\, \hat{\boldsymbol{z}} + \boldsymbol{\nabla}_\perp \right) \cdot \boldsymbol{\psi}(\boldsymbol{r}, \xi; \omega) = 0, \tag{4.25a}$$

$$\left(\hat{\boldsymbol{z}} \tfrac{\partial}{\partial \xi} + ik\, \hat{\boldsymbol{z}} + \boldsymbol{\nabla}_\perp \right) \cdot \boldsymbol{f}(\boldsymbol{r}, \xi; \omega) = 0. \tag{4.25b}$$

We assume that the charge density ρ and full current density $\boldsymbol{j}$ remain weakly localized transversely, in the sense that they are square-integrable in any transverse plane. Actually we will impose the slightly stronger constraint that $|\boldsymbol{f}(\boldsymbol{r};\xi;\omega)| \to 0$ more rapidly than $1/|\boldsymbol{r}|$, as $|\boldsymbol{r}| \to \infty$. We also assume that the physical currents are strongly localized longitudinally, in the sense that they are negligible outside some finite interval $\xi_0' \leq \xi \leq \xi_1'$ along the propagation axis (typically chosen to overlap the origin at $\xi = 0$), implying in general that the solenoidal part of the current density $\boldsymbol{j}_\perp$ will actually have a component with non-compact support, corresponding to the $-\left(\hat{\boldsymbol{z}}\frac{\partial}{\partial\xi} + \boldsymbol{\nabla}_\perp\right)\frac{\partial}{\partial\tau}\phi$ term. However, as we have argued, this contribution to the sources falls off no slower than the inverse-square of the distance from the source charges, so with arbitrarily small relative error in the radiation fields, which are linear in the sources and fall off with the inverse distance from them, we can safely truncate the source envelope $\boldsymbol{f}(\boldsymbol{r};\xi;\omega)$ outside of some sufficiently large but finite range $\xi_0 < \xi_0' \leq \xi \leq \xi_1' < \xi_1$. (Again, the neglected contributions are strictly necessary to cancel the unretarded part of the Coulomb fields to ensure that the full fields remain compatible with relativistic causality, and to supply the proper non-radiative velocity fields, but we are confining attention only to actual radiation.) Most of the final results will not actually depend on the size of this interval, so if necessary we may safely take $|\xi_1 - \xi_0| \to \infty$ at the end of the calculation, provided that the charge and current densities continue to fall off sufficiently rapidly to ensure convergence of the relevant integrals, and that we can somehow distinguish the far-fields from the local quasi-static fields when the latter may then extend to infinity.

The (scaled) waist of the radiation beam, or characteristic transverse spot size w_0, may be defined in terms of the minimal, power-weighted, root mean-square radius:

$$\tfrac{1}{2}w_0^2 = \inf_\xi \frac{\int d^2\boldsymbol{r}\ |\boldsymbol{r}|^2\,|\boldsymbol{\psi}(\boldsymbol{r},\xi;\omega)|^2}{\int d^2\boldsymbol{r}\ |\boldsymbol{\psi}(\boldsymbol{r},\xi;\omega)|^2}. \tag{4.26}$$

The corresponding characteristic diffraction angle θ_{D} may then be taken as

$$\theta_{\text{D}} = \frac{1}{2\pi}\frac{\lambda}{w_0} = \frac{1}{kw_0}, \tag{4.27}$$

and finally the (scaled) Rayleigh range ξ_{R}, or longitudinal length-scale for appreciable change in the transverse spot size due to diffraction, becomes:

$$\xi_{\text{R}} = \tfrac{1}{2}kw_0^2. \tag{4.28}$$

We now treat θ_{D} as a small parameter. Specifically, if ψ_j is any (non-zero) component of the radiation envelope $\boldsymbol{\psi}$, then we may consistently assume that: $\left|\frac{\partial}{\partial\xi}\psi_j\right| / |\psi_j| \sim O(\xi_{\text{R}}^{-1})$,

while $\|\boldsymbol{\nabla}_\perp \psi_j\| / |\psi_j| \sim O(w_0^{-1})$. If the spot size is sufficiently large compared to the carrier wavelength $\lambda = 2\pi/k$, such that $\theta_{\mathrm{D}} \ll 1$, then with a relative error of $O(\theta_{\mathrm{D}}^2)$, we may drop the higher-order longitudinal spatial derivative in equation (4.24), obtaining the forced paraxial wave equation

$$\left(i\tfrac{\partial}{\partial\xi} + \tfrac{1}{2k}\nabla_\perp^2\right)\boldsymbol{\psi}(\boldsymbol{r},\xi;\omega) + \boldsymbol{f}(\boldsymbol{r},\xi;\omega) = 0. \tag{4.29}$$

Note that we have not yet directly imposed the same ordering assumptions on the spatial variations in the envelope $\boldsymbol{f}(\boldsymbol{r};\xi;\omega)$, but as the latter is the actual source for the radiation, it is clear that the transverse scale-lengths of $\boldsymbol{\psi}$ must be consistent with those of $\boldsymbol{f}(\boldsymbol{r};\xi;\omega)$ under diffractive propagation (and therefore often, but not necessarily, comparable in magnitude in the vicinity of the sources). However, as long as we consider sufficiently small wavelengths, the scale-length for longitudinal variations in $\boldsymbol{f}(\boldsymbol{r};\xi;\omega)$ need not be comparable to ξ_{R} in order for the paraxial approximation to remain accurate for the fields of interest, i.e., for radiation observed in the far-field region beyond the sources. Specifically, we may imagine the paraxial fields as the superposition of those fields from each infinitesimal transverse slice of current, evolved forward according to the free-space paraxial propagation, so that each component will satisfy the paraxial conditions everywhere except for a discontinuous jump at its respective source slice, which can be replaced with an equivalent boundary condition applied to the homogeneous (source-free) equation. The paraxial solution is expected to remain accurate for any $\xi > \xi_1$ if the paraxial conditions both hold for the calculated fields in the region $\xi > \xi_1$, and also hold for $\xi < \xi_1$, when these fields are extrapolated backwards via the free-space (i.e., homogeneous, or source-free) paraxial propagation.

In general, for a source to then produce radiation-zone fields (but not necessarily near fields) described within the paraxial approximation, the longitudinal length-scale for the source size is essentially unconstrained, while the *effective* transverse source area should be sufficiently large compared to the wavelength λ. For a relativistic electron beam ($\gamma \gg 1$) radiating in a weak, planar undulator structure (normalized wiggler parameter $a_u < 1$), consisting of $N_u \gg 1$ periods of length λ_u, the emission angle $\Delta\theta$ for radiation observed in the far field within the homogeneous bandwidth $\Delta\omega = \frac{\omega}{\lambda N_u}$, centered on the resonant frequency $\omega = \frac{2\pi}{\lambda}$, where $\lambda = \frac{\lambda_u}{2\gamma^2}\left(1 + \frac{a_u^2}{2}\right)$, will be confined by relativistic effects and constructive interference over the multiple wiggler periods to be

$$\Delta\theta \sim \sqrt{\tfrac{\lambda}{N_u\lambda_u}} \approx \tfrac{1+\frac{1}{4}a_u^2}{\sqrt{2}\sqrt{N_u}\gamma}. \tag{4.30}$$

The effective transverse size Δx of this extended source is then not that of the individual

radiating electrons (either zero, for true point particles, or perhaps the classical electron radius r_e), nor the RMS transverse size σ_x of the electron beam, but roughly

$$\Delta x \sim \left[\sigma_x^2 + \tfrac{1}{2} N_u \lambda_u \Delta\theta\right]^{1/2} \geq \tfrac{1}{\sqrt{2}} \tfrac{\gamma\sqrt{N_u}}{\sqrt{1+\frac{1}{2}a_u^2}} \lambda \gg \lambda, \tag{4.31}$$

as can be seen by simple ray-tracing arguments. In fact, for sufficiently low-emittance electron beams where $\sigma_x \ll N_u \lambda_u \Delta\theta$, we have $\Delta_x \Delta\theta \sim \lambda$, and we expect that the radiation might be well described by a small number of diffraction-limited paraxial modes.

Strictly speaking, the paraxial wave equation is valid when the propagation direction is sufficiently close to the fiducial ($+\hat{\boldsymbol{z}}$) axis, and the scale-length for transverse variations in the radiation beam are large compared to the carrier wavelength and small compared to the scale-length for longitudinal envelope variation (i.e., residual variation after the faster carrier oscillations have been removed). Actually, although we have formally assumed that $\theta_{\rm D} \ll 1$, numerical solutions to the full wave equation typically reveal that the paraxial approximation actually remains surprisingly accurate for focused radiation beams provided only $\theta_{\rm D} \lesssim O(1)$. For still smaller spot sizes (tighter focus), the paraxial equation predicts a coherence volume $V_{\rm coh} < O(\lambda^3)$ for a single oscillatory mode, which would be in conflict with the diffraction limit dictated by the "Heisenberg" inequality constraining conjugate Fourier transform pairs.

From our above ordering assumptions and the transversality condition (4.25a), it follows that $|\psi_z| \,/ (|\psi_x|^2 + |\psi_y|^2)^{1/2} \sim O(\theta_{\rm D})$, so the axial component ψ_z can in certain circumstances be neglected outright, but we will retain it here in order to explicitly satisfy the gauge constraint. However, whenever convenient, the gauge condition for fully paraxial fields may be simplified, within the same $O(\theta_{\rm D}^2)$ relative accuracy as the paraxial wave equation itself, by dropping the longitudinal derivative:

$$(ik\,\hat{\boldsymbol{z}} + \boldsymbol{\nabla}_\perp) \cdot \boldsymbol{\psi}(\boldsymbol{r}, \xi; \omega) \approx 0. \tag{4.32}$$

So if the (geometrically) transverse components $\boldsymbol{\psi}_\perp = \boldsymbol{\psi} - \hat{\boldsymbol{z}}\big(\hat{\boldsymbol{z}} \cdot \boldsymbol{\psi}\big)$ are specified, then in this approximation $\psi_z(\boldsymbol{r}, \xi; \omega)$ may be taken as

$$\hat{\boldsymbol{z}} \cdot \boldsymbol{\psi}(\boldsymbol{r}, \xi; \omega) = \tfrac{i}{k} \boldsymbol{\nabla}_\perp \cdot \boldsymbol{\psi}_\perp(\boldsymbol{r}, \xi; \omega), \tag{4.33}$$

or conversely, if $\psi_z(\boldsymbol{r}, \xi; \omega)$ is given, then a $\boldsymbol{\psi}_\perp$ can always be chosen to satisfy (4.33) and ensure paraxial Helmholtz-transversality. The solenoidal source envelope $\boldsymbol{f}$ must in general satisfy the complete envelope constraint (4.25b) including the longitudinal derivative, because its ξ-dependence may be otherwise arbitrary. Since the paraxial wave equation is

first-order in $\boldsymbol{\xi}$, it of course only describes radiation traveling nearly in one direction (rightward, in our convention). If we are also interested in radiation propagating leftward (i.e., in the $-\hat{\boldsymbol{z}}$ direction) from the sources, obviously we can handle those fields by a separate but exactly analogous procedure.

4.3.2 Hilbert Space Formalism

Mathematically, the homogeneous (i.e., free-space, or source-free) part of (4.29) is exactly analogous to the Schrödinger equation, in the position representation, for a quantum mechanical, spin-one particle of mass M, moving non-relativistically in a potential-free, two-dimensional, Euclidean space, where the longitudinal position ξ plays the role of the "temporal" evolution variable, $\boldsymbol{r}$ serves as the spatial coordinates, and polarization is analogous to spin angular momentum, all in units where $\hbar = 1$ and $M = k = \omega$.

In order to leverage the analogies to ordinary quantum mechanics, we introduce a Hilbert space, inner product, state vectors, operators, etc., using a Dirac-like notation. The full Hilbert space may be considered a tensor product of the Hilbert spaces for the (geometrically) transverse-spatial and the spin (polarization) degrees of freedom. (Our notation is similar, but not identical, to that in [121, 122]). We associate the ket (really a spinor) $|\boldsymbol{\psi}\rangle$ with the complex vector field $\boldsymbol{\psi}(\boldsymbol{r})$ defined on the (geometrically) transverse spatial plane, i.e., as a mapping $\boldsymbol{\psi} : \mathbb{R}^2 \to \mathbb{C}^3$ (possibly also parameterized by additional continuous variables such as ξ and ω and by discrete mode indices, all suppressed for the moment). The corresponding bra $\langle\boldsymbol{\psi}|$ is associated with the dual, or conjugate transpose, vector field, $\boldsymbol{\psi}(\boldsymbol{r})^\dagger$, and a combined $\mathscr{L}_2$/Euclidean inner product is defined by

$$\langle\boldsymbol{\psi}_1\,|\,\boldsymbol{\psi}_2\rangle = \int d^2\boldsymbol{r}\ \boldsymbol{\psi}_1(\boldsymbol{r})^* \cdot \boldsymbol{\psi}_2(\boldsymbol{r}). \tag{4.34}$$

We define a number of Hermitian operators acting on the spatial degrees of freedom by their action on these position-representation wave vector fields $\boldsymbol{\psi}$. Given any fixed unit vectors $\hat{\boldsymbol{e}}$, $\hat{\boldsymbol{e}}' \in \mathbb{C}^2$ in the complex linear span of $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$, the transverse position operators just act via multiplication by the corresponding transverse spatial coordinates

$$\hat{\boldsymbol{e}} \cdot \boldsymbol{Q}\boldsymbol{\psi}(\boldsymbol{r}) \equiv (\hat{\boldsymbol{e}} \cdot \boldsymbol{r})\boldsymbol{\psi}(\boldsymbol{r}), \tag{4.35}$$

while the conjugate "momentum" operators act by differentiation:

$$\hat{\boldsymbol{e}}' \cdot \boldsymbol{P}\boldsymbol{\psi}(\boldsymbol{r}) \equiv \tfrac{1}{i}\left(\hat{\boldsymbol{e}}' \cdot \boldsymbol{\nabla}_\perp\right)\boldsymbol{\psi}(\boldsymbol{r}), \tag{4.36}$$

so these operators then satisfy the usual canonical commutation relations

$$\left[\hat{\boldsymbol{e}} \cdot \boldsymbol{Q}\,,\, \hat{\boldsymbol{e}}' \cdot \boldsymbol{P}\right] = i\left(\hat{\boldsymbol{e}} \cdot \hat{\boldsymbol{e}}'\right) I_r, \tag{4.37}$$

where $I_{\boldsymbol{r}}$ is the identity operator on transverse position-state space. We stress that these operators act only on the transverse spatial coordinates, so that $\hat{z} \cdot \boldsymbol{Q} = \hat{\boldsymbol{z}} \cdot \boldsymbol{P} = 0$. The "kinetic energy" or "free-particle Hamiltonian" operator is then defined as:

$$H \equiv \tfrac{1}{2M}\boldsymbol{P} \cdot \boldsymbol{P} = -\tfrac{1}{2k}\nabla_\perp^2, \tag{4.38}$$

which is just the infinitesimal generator of free-space, diffractive propagation along ξ. With somewhat more mathematical care about operator domains than is justified for this presentation, these Hermitian operators may be extended to fully self-adjoint operators without difficulty.

In order to include the polarization, we also define the complex helicity basis in $\mathbb{C}^3$:

$$\hat{\boldsymbol{\epsilon}}_s = -s\tfrac{1}{\sqrt{2}}\left(\hat{\boldsymbol{x}} - is\hat{\boldsymbol{y}}\right) + (1-s^2)\hat{\boldsymbol{z}} = -\tfrac{1}{2}s^2(1+s)\hat{\boldsymbol{e}}_- + \tfrac{1}{2}s^2(1-s)\hat{\boldsymbol{e}}_+ + (1-s^2)\hat{\boldsymbol{z}}, \tag{4.39}$$

for $s = -1, 0, +1$, satisfying

$$\hat{\boldsymbol{\epsilon}}_s^* \cdot \hat{\boldsymbol{\epsilon}}_{s'} = \delta_{s\,s'}, \tag{4.40}$$

where $\delta_{s\,s'}$ is the usual Kronecker delta symbol and $\hat{\boldsymbol{e}}_\pm \equiv \frac{1}{\sqrt{2}}(\hat{\boldsymbol{x}} \pm i\hat{\boldsymbol{y}}) = \hat{\boldsymbol{e}}_\mp^*$, and then define various Hermitian "spin" operators by their action on the polarization components of the fields $\boldsymbol{\psi}$. In particular, we define a vector of mutually orthogonal polarization projection operators

$$\boldsymbol{\Pi} = \sum_{s=-1,0,+1} \hat{\boldsymbol{\epsilon}}_s \Pi_s, \tag{4.41}$$

where

$$\Pi_s \boldsymbol{\Psi} = \hat{\boldsymbol{\epsilon}}_s \left(\hat{\boldsymbol{\epsilon}}_s^* \cdot \boldsymbol{\Psi}\right), \tag{4.42}$$

such that

$$\Pi_s \Pi_{s'} = \Pi_{s'} \Pi_s = \delta_{s\,s'} \Pi_s, \tag{4.43}$$

and

$$\sum_{s=-1,0,+1} \Pi_s = I_3, \tag{4.44}$$

where

$$I_3 = \sum_{s=-1,0,+1} \hat{\boldsymbol{\epsilon}}_s \hat{\boldsymbol{\epsilon}}_s^\dagger \tag{4.45}$$

is the identity operator on the spin degrees of freedom. We also define the helicity, or circular polarization operator by

$$S_z = \Pi_1 - \Pi_{-1}, \tag{4.46}$$

so that

$$S_z \boldsymbol{\psi} = \hat{\boldsymbol{e}}_- \hat{\boldsymbol{e}}_-^* \cdot \boldsymbol{\psi} - \hat{\boldsymbol{e}}_+ \hat{\boldsymbol{e}}_+^* \cdot \boldsymbol{\psi}. \tag{4.47}$$

and

$$S_z^2 = I_3 - \Pi_0. \tag{4.48}$$

For an enlightening discussion of the orbital and spin components of angular momentum in electromagnetic fields, see [123].

Clearly, all of these polarization operators commute with the spatial operators, so we may define generalized (i.e., non-normalizable) simultaneous eigenkets of position and spin $|\boldsymbol{r}; s\rangle$ for any $\boldsymbol{r} \in \mathbb{R}^2$ and $s \in \{-1, 0, 1\}$:

$$\boldsymbol{Q} \, |\boldsymbol{r}; s\rangle = \boldsymbol{r} \, |\boldsymbol{r}; s\rangle \,, \tag{4.49}$$

$$e^{i\boldsymbol{r}' \cdot \boldsymbol{P}} \, |\boldsymbol{r}; s\rangle = \left|\boldsymbol{r} - \boldsymbol{r}'; s\right\rangle, \tag{4.50}$$

and

$$S_z \, |\boldsymbol{r}; s\rangle = s \, |\boldsymbol{r}; s\rangle \,. \tag{4.51}$$

We similarly define generalized simultaneous eigenkets of spin and "momentum" (transverse wavevector) by taking the Fourier transforms of these position kets:

$$|\boldsymbol{p}; s\rangle = \tfrac{1}{2\pi} \int d^2\boldsymbol{r} \, e^{i\boldsymbol{p}\cdot\boldsymbol{r}} \, |\boldsymbol{r}; s\rangle \,, \tag{4.52}$$

which satisfy

$$\boldsymbol{P} \, |\boldsymbol{p}; s\rangle = \boldsymbol{p} \, |\boldsymbol{p}; s\rangle \,, \tag{4.53}$$

and

$$S_z \, |\boldsymbol{p}; s\rangle = s \, |\boldsymbol{p}; s\rangle \,. \tag{4.54}$$

These generalized eigenkets satisfy the orthogonality relations

$$\left\langle \boldsymbol{r}; s \,\middle|\, \boldsymbol{r}'; s' \right\rangle = \delta_{s\,s'} \, \delta^{(2)}(\boldsymbol{r} - \boldsymbol{r}'), \tag{4.55}$$

and

$$\left\langle \boldsymbol{p}; s \,\middle|\, \boldsymbol{p}'; s' \right\rangle = \delta_{s\,s'} \, \delta^{(2)}(\boldsymbol{p} - \boldsymbol{p}'), \tag{4.56}$$

and the completeness relation

$$\sum_s \int d^2\boldsymbol{r}\ |\boldsymbol{r};s\,\rangle\langle\,\boldsymbol{r};s| = \sum_s \int d^2\boldsymbol{p}\ |\boldsymbol{p};s\,\rangle\langle\,\boldsymbol{p};s| = I, \tag{4.57}$$

where $\delta^{(2)}(\boldsymbol{r})$ is the two-dimensional Dirac delta function, and $I = I_3 \otimes I_{\boldsymbol{r}}$ is the identity operator on the full Hilbert space $\mathcal{H}$ consisting of all $|\boldsymbol{\psi}\rangle$ which are normalizable, i.e., for which $\langle\boldsymbol{\psi}\,|\,\boldsymbol{\psi}\rangle < \infty$, implying that the corresponding paraxial radiation fields $\boldsymbol{\psi}(\boldsymbol{r})$ are of finite power (but not necessarily of finite energy). The allowed radiation envelopes actually reside in the Hilbert sub-space $\mathcal{H}_\perp \subset \mathcal{H}$ consisting of vector fields which also satisfy the transverse envelope gauge constraint when evolved in ξ according to free-space paraxial evolution.

Defining the coupled spin-momentum operator

$$T(\delta) = \left(k - i\delta\frac{\partial}{\partial\xi}\right)\Pi_0 + \boldsymbol{P}\cdot\boldsymbol{\Pi} \tag{4.58}$$

for a real parameter δ, and the constant reference ket $|\boldsymbol{\psi}_0\rangle \in \mathcal{H}$ such that

$$\boldsymbol{\psi}_0(\boldsymbol{r}) = \tfrac{1}{\sqrt{3}}\left(\hat{\boldsymbol{\epsilon}}_{-1} + \hat{\boldsymbol{\epsilon}}_0 + \hat{\boldsymbol{\epsilon}}_{+1}\right), \tag{4.59}$$

the full envelope-transversality constraints (4.25) on the radiation and source may be written in our quantum notation as:

$$\langle\boldsymbol{\psi}_0\,|T(\delta=1)|\,\boldsymbol{\psi}(\xi;\omega)\rangle = 0, \tag{4.60a}$$

$$\langle\boldsymbol{\psi}_0\,|T(\delta=1)|\,\boldsymbol{f}(\xi;\omega)\rangle = 0, \tag{4.60b}$$

while the approximate paraxial gauge constraint (4.32) becomes:

$$\langle\boldsymbol{\psi}_0\,|T(\delta=0)|\,\boldsymbol{\psi}(\xi;\omega)\rangle = 0. \tag{4.61}$$

Again, within the paraxial approximation, either of these constraints may be used interchangeably for the radiation, but the source must satisfy the full transversality constraint because its ξ dependence is arbitrary.

It is straightforward to see how the full envelope gauge constraint (4.60a) (i.e., including the longitudinal derivative) will be automatically maintained by the evolution, and how the approximate paraxial constraint (4.61) (i.e., neglecting the longitudinal derivative) will be preserved by free-space propagation or by propagation in the region of a fully paraxial source. Assuming that $|\boldsymbol{\psi}(\xi;\omega)\rangle$ satisfies the paraxial wave equation, with an initial condition satisfying

$$\lim_{\xi\to-\infty}\langle\boldsymbol{\psi}_0\,|T(\delta)|\,\boldsymbol{\psi}(\xi;\omega)\rangle = 0, \tag{4.62}$$

(our specific choice of $|\boldsymbol{\psi}(\xi;\omega)\rangle = 0$ for all $\xi < \xi_0$ clearly works), and that the source $|\boldsymbol{f}(\xi;\omega)\rangle$ satisfies

$$\langle \boldsymbol{\psi}_0 | T(\delta) | \boldsymbol{f}(\xi;\omega)\rangle = 0 \qquad \forall \xi \in \mathbb{R}, \tag{4.63}$$

and noting that

$$[T(\delta), H(\omega)] = \left[T(\delta), \tfrac{\partial}{\partial \xi}\right] = 0, \tag{4.64}$$

as well as

$$\tfrac{\partial}{\partial \xi} H(\omega) = 0, \tag{4.65}$$

and

$$H(\omega) |\boldsymbol{\psi}_0\rangle = 0, \tag{4.66}$$

then:

$$\begin{aligned} \tfrac{\partial}{\partial \xi} \langle \boldsymbol{\psi}_0 | T(\delta) | \boldsymbol{\psi}(\xi;\omega)\rangle &= \left\langle \boldsymbol{\psi}_0 \left| T(\delta) \tfrac{\partial}{\partial \xi} \right| \boldsymbol{\psi}(\xi;\omega)\right\rangle \\ &= -i \langle \boldsymbol{\psi}_0 | T(\delta) H(\omega) | \boldsymbol{\psi}(\xi;\omega)\rangle + i \langle \boldsymbol{\psi}_0 | T(\delta) | \boldsymbol{f}(\xi;\omega)\rangle \\ &= -i \langle \boldsymbol{\psi}_0 | H(\omega) T(\delta) | \boldsymbol{\psi}(\xi;\omega)\rangle + 0 = 0. \end{aligned} \tag{4.67}$$

Using the initial condition, we then have $\langle \boldsymbol{\psi}_0 | T(\delta) | \boldsymbol{\psi}(\xi;\omega)\rangle = 0$ for all ξ. This shows that the full ($\delta = 1$) envelope transversality constraint (4.60a) on the fields, including the longitudinal derivative, is preserved everywhere, if it holds initially for the fields and holds everywhere for the sources. The approximate (i.e., $\delta = 0$) paraxial gauge constraint (4.61) is preserved by the evolution in regions free of sources and wherever the solenoidal source also satisfies this same, stronger condition.

4.3.3 Green Function Solution

In this quantum-like notation, the paraxial wave equation may be written as

$$i \tfrac{\partial}{\partial \xi} |\boldsymbol{\psi}(\xi;\omega)\rangle = H |\boldsymbol{\psi}(\xi;\omega)\rangle - |\boldsymbol{f}(\xi;\omega)\rangle \tag{4.68}$$

Given some "initial" state $|\boldsymbol{\psi}(\xi';\omega)\rangle$, the solution to the homogeneous part of the equation (i.e., for $|\boldsymbol{f}(\xi;\omega)\rangle = 0$) representing free space propagation of the radiation fields, may be written in terms of the unitary evolution operator as

$$|\boldsymbol{\psi}(\xi;\omega)\rangle = U(\xi,\xi';\omega) \left|\boldsymbol{\psi}(\xi';\omega)\right\rangle, \tag{4.69}$$

where $U(\xi,\xi';k)$ satisfies the operator-valued Schrödinger equation

$$i \tfrac{\partial}{\partial \xi} U(\xi,\xi';\omega) = H U(\xi,\xi';\omega) \tag{4.70}$$

with initial condition

$$U(\xi', \xi'; \omega) = I, \tag{4.71}$$

and possesses the group composition properties

$$U(\xi, \xi'; \omega)^{-1} = U(\xi, \xi'; \omega)^\dagger = U(\xi', \xi; \omega), \tag{4.72}$$

and

$$U(\xi, \xi''; \omega) U(\xi'', \xi'; \omega) = U(\xi, \xi'; \omega). \tag{4.73}$$

Here $H = H(\omega)$ (with $\omega = k$) is the Hamiltonian, or diffraction operator, defined above. Because H is ξ-independent, we can immediately integrate (4.70) to find

$$U(\xi, \xi'; \omega) = e^{-i(\xi - \xi')H(\omega)}. \tag{4.74}$$

The solution to the full inhomogeneous wave equation (4.68) may be written in terms of the propagator, or causal Green function operator $G(\xi, \xi'; \omega)$ satisfying

$$\left(i\tfrac{\partial}{\partial \xi} - H\right) G(\xi, \xi'; \omega) = \delta(\xi - \xi') \tag{4.75}$$

and

$$G(\xi, \xi'; \omega) = 0 \quad \text{for} \quad \xi < \xi'. \tag{4.76}$$

It can easily be verified that this operator is given by

$$G(\xi, \xi'; \omega) = -i\Theta(\xi - \xi') U(\xi, \xi'; \omega). \tag{4.77}$$

Here $\Theta(\xi)$ is the usual Heaviside step function, and $\delta(\xi) = \frac{d}{d\xi}\Theta(\xi)$ is the one-dimensional Dirac delta function. Assuming no initial (incoming or outgoing) radiation, the formal solution to the inhomogeneous equation (4.68) (i.e., with the source term) is then

$$\begin{aligned} |\boldsymbol{\psi}(\xi; \omega)\rangle &= -\int_{-\infty}^{\infty} d\xi' \, G(\xi, \xi'; \omega) \left| \boldsymbol{f}(\xi'; \omega) \right\rangle \\ &= i \int_{\xi_0}^{\min(\xi, \xi_1)} d\xi' \, U(\xi, \xi'; \omega) \left| \boldsymbol{f}(\xi'; \omega) \right\rangle, \end{aligned} \tag{4.78}$$

where we have used the assumption that the support of $\boldsymbol{f}(\xi; \boldsymbol{r}; \omega)$ is confined to $\xi_0 < \xi < \xi_1$. For any $\xi < \xi_0$, we then have $|\boldsymbol{\psi}(\xi; \omega)\rangle = 0$ consistent with our assumed initial condition, while for all $\xi \geq \xi_1$, this just reduces to free-space propagation: $|\boldsymbol{\psi}(\xi; \omega)\rangle = U(\xi, \xi_1; \omega) \left| \boldsymbol{\psi}(\xi_1; \omega) \right\rangle$. In between we must actually solve the full expression (4.78).

4.3.4 Energetics

In our dimensionless variables, recall that the normalized, solenoidal electric field is given by

$$\boldsymbol{\varepsilon}_{\boldsymbol{a}\perp}(\boldsymbol{r};\xi;\tau) = -\tfrac{\partial}{\partial\tau}\boldsymbol{a}(\boldsymbol{r};\xi;\tau) = -\tfrac{\partial}{\partial\tau}\boldsymbol{\psi}(\boldsymbol{r};\xi;\tau)e^{ik\xi} \tag{4.79}$$

in the (scaled) time domain, or equivalently

$$\boldsymbol{\varepsilon}_{\boldsymbol{a}\perp}(\boldsymbol{r};\xi;\omega) = i\omega\,\boldsymbol{a}(\boldsymbol{r};\xi;\omega) = i\omega\,\boldsymbol{\psi}(\boldsymbol{r};\xi;\omega)e^{ik\xi} \tag{4.80}$$

in the (scaled) frequency domain. Within the paraxial approximation, the inner product $\langle\boldsymbol{\psi}(\xi;\omega)\,|\,\boldsymbol{\psi}(\xi;\omega)\rangle$ is then proportional to the (time-averaged) power spectral density for radiation of frequency ω passing through the transverse plane at the longitudinal position ξ. In fact, in scaled units we may define the (normalized) power spectral density as

$$\begin{aligned}\mathscr{P}'_{\mathrm{EM}}[\boldsymbol{a}](\xi;\omega) &= \omega^2\,\langle\boldsymbol{\psi}(\xi;\omega)\,|\,\boldsymbol{\psi}(\xi;\omega)\rangle = \omega^2\int d^2\boldsymbol{r}\;\;|\boldsymbol{\psi}(\boldsymbol{r};\xi;\omega)|^2 \\ &= \omega^2\int d^2\boldsymbol{r}\;\;|\boldsymbol{a}(\boldsymbol{r};\xi;\omega)|^2 = \int d^2\boldsymbol{r}\;\;|\boldsymbol{\varepsilon}_\perp(\boldsymbol{r};\xi;\omega)|^2\,.\end{aligned} \tag{4.81}$$

Strictly speaking, this yields the time-averaged Poynting flux only to $O(\theta_{\mathrm{D}})$ at any finite distance from the sources, but it is exact as $\xi\to\infty$ and we consider only the truly radiative transport of energy to spatial infinity. The unitary (inner-product preserving) nature of the paraxial evolution in vacuum then corresponds to energy conservation for paraxial fields, where the (time-averaged) power crossing any transverse plane in any frequency bandwidth remains constant (with respect to propagation distance ξ) during free-space evolution. Note, however, that while it is particularly convenient to formulate paraxial propagation in terms of the transformation of the fields in successive transverse planes, the actual wavefronts (level surfaces of phase) of a general paraxial beam are not planar, except at the focus.

From the actual solution $|\boldsymbol{\psi}(\xi_2;\omega)\rangle$ in any conveniently-chosen transverse plane $\xi_2\geq\xi_1$ beyond (i.e., to the right of) the sources, we can construct the extrapolated free-space field

$$|\boldsymbol{\chi}(\xi;\omega)\rangle = U(\xi,\xi_2;\omega)\,|\boldsymbol{\psi}(\xi_2;\omega)\rangle = i\int_{\xi_0}^{\xi_1} d\xi'\,U(\xi,\xi';\omega)\,\big|\boldsymbol{f}(\xi';\omega)\big\rangle\,, \tag{4.82}$$

which represents the post-source paraxial radiation envelope extrapolated throughout all space via free-space propagation, including in the backward or "upwind" direction into the region where sources are actually present. Thus $|\boldsymbol{\chi}(\xi;\omega)\rangle$ and $|\boldsymbol{\psi}(\xi;\omega)\rangle$ coincide for all $\xi>\xi_1$, but not in general for $\xi<\xi_1$, where $|\boldsymbol{\chi}(\xi;\omega)\rangle$ satisfies the homogeneous (source-free) Schrödinger equation with no incoming radiation as $\xi\to-\infty$, whereas $|\boldsymbol{\psi}(\xi;\omega)\rangle$ satisfies the

inhomogeneous equation everywhere, so that all rightward outgoing radiation (i.e., fields propagating away from the sources, to the right toward $\xi \to \infty$) is matched by incoming radiation (i.e., fields propagating left to right, from $\xi \to -\infty$ toward the sources.) For any particular ξ, these extrapolated fields are those fields which would have been present if the fields actually observed beyond the sources at some $\xi_2 > \xi_1$ were instead produced by some effective source $\boldsymbol{g}(\boldsymbol{r};\xi;\bar{\xi};\omega)$ located in some sufficiently remote region $\xi_0 - \bar{\xi} < \xi < \xi_1 - \bar{\xi}$, instead of by the actual sources within $\xi_0 < \xi < \xi_1$:

$$|\boldsymbol{\chi}(\xi;\omega)\rangle = i \int_{\xi_0-\bar{\xi}}^{\xi_1-\bar{\xi}} d\xi' \; U(\xi,\xi';\omega) \left|\boldsymbol{g}(\xi';\bar{\xi};\omega)\right\rangle, \tag{4.83}$$

where the effective source is determined from the actual source simply by longitudinal translation and compensation for diffraction:

$$\left|\boldsymbol{g}(\xi';\bar{\xi};\omega)\right\rangle = U(\xi',\xi+\bar{\xi}) \left|\boldsymbol{f}(\xi'+\bar{\xi};\omega)\right\rangle, \tag{4.84}$$

and $\bar{\xi} = \bar{\xi}(\xi;\xi_1) \geq 0$ may any positive constant satisfying

$$\bar{\xi} > \xi_1 - \xi. \tag{4.85}$$

Choosing any $\bar{\xi} > \xi_1 - \xi_0$, the expression (4.83) is then valid for any $\xi \geq \xi_0$, or we may actually take $\bar{\xi} \to +\infty$ if convenient, moving the effective sources to an arbitrarily remote "upwind" location, leading to source-free solutions at any $-\infty < \xi \leq +\infty$. (Again, any left-going, outward radiation from the sources should be handled by an analogous but separate paraxial solution. For the particular case of relativistic electron beams, we have seen that the Lorentz transformation of the electron's dipole radiation fields from the average rest frame to the lab frame confine the radiation substantially to a narrow solid angle of $\lesssim O\left(\frac{1}{\gamma}\right)$ in the forward direction.)

Using the Green function solution (4.78), and taking advantage of the assumed finite support of $\boldsymbol{f}$, we find that for any $\xi > \xi_1$,

$$\begin{aligned}
\langle \boldsymbol{\psi}(\xi;\omega) \,|\, \boldsymbol{\psi}(\xi;\omega)\rangle &= i \int_{\xi_0}^{\min(\xi,\xi_1)} d\xi' \; \left\langle \boldsymbol{\psi}(\xi;\omega) \left| U(\xi,\xi';\omega) \right| \boldsymbol{f}(\xi';\omega) \right\rangle \\
&= i \int_{-\infty}^{\infty} d\xi' \; \left\langle \boldsymbol{\chi}(\xi';\omega) \,\middle|\, \boldsymbol{f}(\xi';\omega) \right\rangle \\
&= i \int d\xi' \int d^2\boldsymbol{r} \; \boldsymbol{\chi}(\boldsymbol{r};\xi';\omega)^* \cdot \boldsymbol{f}(\boldsymbol{r};\xi';\omega),
\end{aligned} \tag{4.86}$$

which is just a three-dimensional inner product, or overlap integral, between the source envelope $\boldsymbol{f}(\boldsymbol{r};\xi';\omega)$ and the free-space radiation envelope $\boldsymbol{\chi}(\boldsymbol{r};\xi';\omega)$ which has been extrapolated via free-space propagation backwards into the region where sources are present.

While solenoidal and irrotational vector fields are not in general locally orthogonal in the geometric sense, they are Hilbert-space orthogonal when their Euclidean inner product is integrated over all space. Because $|\boldsymbol{\psi}(\xi;\omega)\rangle$ and hence $|\boldsymbol{\chi}(\xi;\omega)\rangle$ satisfy the transverse gauge constraint, this overlap integral may be simplified to

$$\begin{aligned}
\langle \boldsymbol{\psi}(\xi;\omega) \,|\, \boldsymbol{\psi}(\xi;\omega)\rangle &= i \int d\xi' \int d^2\boldsymbol{r}\; \boldsymbol{\chi}(\boldsymbol{r};\xi';\omega)^* \cdot \boldsymbol{f}(\boldsymbol{r};\xi';\omega) \\
&= -\tfrac{1}{2\omega^2} \int d\xi' \int d^2\boldsymbol{r}\; \left[i\omega\, \boldsymbol{\chi} e^{ik\xi} \right]^* \cdot \left[2k\, \boldsymbol{f} e^{ik\xi} \right] \\
&= -\tfrac{1}{2\omega^2} \int d\xi' \int d^2\boldsymbol{r}\; \boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\boldsymbol{r};\xi';\omega)^* \cdot \boldsymbol{j}_\perp(\boldsymbol{r};\xi';\omega) \\
&= -\tfrac{1}{2\omega^2} \int d\xi' \int d^2\boldsymbol{r}\; \boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\boldsymbol{r};\xi';\omega)^* \cdot \boldsymbol{j}(\boldsymbol{r};\xi';\omega) \\
&= -\tfrac{1}{2\omega^2} \int d\xi' \, \left\langle \boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\xi';\omega) \,\middle|\, \boldsymbol{j}(\xi';\omega) \right\rangle ,
\end{aligned} \tag{4.87}$$

where $\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\boldsymbol{r};\xi';\omega)$ is the scaled, solenoidal electric field associated with the extrapolated radiation envelope $\boldsymbol{\chi}$. In order to evaluate the power, we therefore need never explicitly decompose the current density into its solenoidal and irrotational components, nor actually worry about the precise spatial cutoff beyond which the solenoidal component may be neglected.

As one would expect from energy conservation considerations, this overlap integral may be interpreted in terms of the rate of energy transfer between the charges constituting the sources and the fields. We may define the scaled power spectral density associated with the rate of (time-averaged) mechanical work done by scaled electric fields $\boldsymbol{\varepsilon}(\boldsymbol{r};\xi;\omega)$ on the charges associated with the scaled current density $\boldsymbol{j}(\boldsymbol{r};\xi';\omega)$ as

$$\begin{aligned}
\mathscr{P}'_{\mathrm{mech}}[\boldsymbol{\varepsilon};\boldsymbol{j}](\omega) &= \mathrm{Re} \int d\xi' \, \left\langle \boldsymbol{\varepsilon}(\xi';\omega) \,\middle|\, \boldsymbol{j}(\xi';\omega) \right\rangle \\
&= \mathrm{Re} \int d\xi' \int d^2\boldsymbol{r}\; \boldsymbol{\varepsilon}(\boldsymbol{r};\xi';\omega)^* \cdot \boldsymbol{j}(\boldsymbol{r};\xi';\omega).
\end{aligned} \tag{4.88}$$

From (4.87), we see that $\int d\xi' \, \langle \boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\xi';\omega) \,|\, \boldsymbol{j}(\xi';\omega)\rangle$ must be purely real and non-positive, because $\langle \boldsymbol{\psi} \,|\, \boldsymbol{\psi}\rangle = \|\boldsymbol{\psi}\|^2$ is necessarily real and non-negative. Physically, this reflects the fact that the back-extrapolated fields, if actually present, would lead to no reactive energy transfer to local fields associated with re-arrangement of the source charges, and mechanical energy would be transferred monodirectionally from the charges to the radiation fields.

Therefore (4.87) is equivalent to:

$$\mathscr{P}'_{\mathrm{EM}}[\boldsymbol{a}](\xi;\omega) = -\tfrac{1}{2}\mathscr{P}'_{\mathrm{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp};\boldsymbol{j}](\omega) \tag{4.89}$$

That is, the average power seen in the far field is exactly one-half of that which would be delivered by the sources to the extrapolated homogeneous fields if these fields were those actually present in the region of the sources.

The presence of the extra factor of $\frac{1}{2}$ is intuitively plausible. The actual paraxial fields from each transverse current slice consist only of the forward half of the corresponding extrapolated fields from this slice, and hence the former fields could interact only with "downstream" current slices while propagating in the forward ($+\hat{\boldsymbol{z}}$) direction, while the extrapolated source-free fields from any current slice impinge on all other current slices, both upstream and downstream. If we adopt the symmetric convention $\Theta(0) = \frac{1}{2}$ for the Heaviside step function (which is anyway the natural choice when Fourier transforms are involved), each current slice also appears to interact only with one-half of its own field. Assuming a reciprocity arising from Newton's Third Law (magnetic forces can violate this law, but do no mechanical work, and radiation reaction forces can violate the third law as well, but are ignored here), we then anticipate an over-counting of the work by exactly a factor of 2 when we use the extrapolated fields and integrate over all current slices. This is a dynamical, electromagnetic analog of the well known electrostatic result that the potential energy associated with a given charge distribution ρ in a given external potential Φ_{ext} is given by $U = \int d^3\boldsymbol{x}\,\rho\Phi_{\mathrm{ext}}$, while the self-energy (potential energy of formation) of a charge distribution resulting in a potential Φ is $U = \frac{1}{2}\int d^3\boldsymbol{x}\,\rho\Phi$.

Below, we will provide an elementary proof of this result in the most general case, but it is informative to verify it with an explicit calculation for paraxial fields. By using the

Green function solution and inserting a resolution of the identity, we have, for any $\xi > \xi_1$,

$$\begin{aligned}
\mathscr{P}'_{\text{EM}}[\boldsymbol{a}](\xi;\omega) &= \omega^2 \left\langle \boldsymbol{\psi}(\xi;\omega) \,|\, \boldsymbol{\psi}(\xi;\omega) \right\rangle = i\omega^2 \int_{\xi_0}^{\xi_1} d\xi' \;\; \left\langle \boldsymbol{\chi}(\xi';\omega) \,\middle|\, \boldsymbol{f}(\xi';\omega) \right\rangle \\
&= \omega^2 \int_{\xi_0}^{\xi_1} d\xi' \int_{\xi_0}^{\xi_1} d\xi'' \;\; \left\langle \boldsymbol{f}(\xi'';\omega) \left| U(\xi'',\xi';\omega) \right| \boldsymbol{f}(\xi';\omega) \right\rangle \\
&= \omega^2 \sum_s \int d^2\boldsymbol{r} \left| \int_{\xi_0}^{\xi_1} d\xi' \;\; \left\langle \boldsymbol{r}; s \left| U(\xi_0,\xi';\omega) \right| \boldsymbol{f}(\xi';\omega) \right\rangle \right|^2 \\
&= \omega^2 \sum_s \int d^2\boldsymbol{r} \left| \int_{\xi_0}^{\xi_1} d\xi' \, \boldsymbol{g}(\boldsymbol{r};\xi_0;\xi'-\xi_0;\omega) \right|^2
\end{aligned} \tag{4.90}$$

which explicitly demonstrates its reality and non-negativity. Using the same sort of manipulations, we then have

$$\begin{aligned}
\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\psi}\perp};\boldsymbol{j}](\omega) &= \text{Re}\left[\int_{\xi_0}^{\xi_1} d\xi' \;\; \left\langle \boldsymbol{\varepsilon}_{\psi\perp}(\xi';\omega) \,\middle|\, \boldsymbol{j}(\xi';\omega) \right\rangle \right] \\
&= \text{Re}\left[-2i\omega k \int_{\xi_0}^{\xi_1} d\xi' \;\; \left\langle \boldsymbol{\psi}(\xi';\omega) \,\middle|\, \boldsymbol{f}(\xi';\omega) \right\rangle \right] \\
&= -2\omega^2 \,\text{Re}\left[\int_{\xi_0}^{\xi_1} d\xi' \int_{\xi_0}^{\xi'} d\xi'' \;\; \left\langle \boldsymbol{f}(\xi'';\omega) \left| U(\xi'',\xi';\omega) \right| \boldsymbol{f}(\xi';\omega) \right\rangle \right] \\
&= -2\omega^2 \sum_s \int d^2\boldsymbol{r} \;\, \text{Re}\left[\int_{\xi_0}^{\xi_1} d\xi' \, \boldsymbol{g}(\boldsymbol{r};\xi_0;\xi'-\xi_0;\omega) \int_{\xi_0}^{\xi'} d\xi'' \, \boldsymbol{g}(\boldsymbol{r};\xi_0;\xi''-\xi_0;\omega)^* \right]
\end{aligned} \tag{4.91}$$

But using integration by parts, this simplifies to

$$\begin{aligned}
&\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\psi}\perp};\boldsymbol{j}](\omega) \\
&\quad = (-2\omega^2) \left[\tfrac{1}{2} \sum_s \int d^2\boldsymbol{r} \left| \int_{\xi_0}^{\xi_1} d\xi' \, \boldsymbol{g}(\boldsymbol{r};\xi_0;\xi'-\xi_0;\omega) \right|^2 \right] \\
&\quad = -\omega^2 \left\langle \boldsymbol{\psi}(\xi;\omega) \,|\, \boldsymbol{\psi}(\xi;\omega) \right\rangle = -\mathscr{P}'_{\boldsymbol{a}}(\xi;\omega).
\end{aligned} \tag{4.92}$$

So, indeed it is the case that:

$$\mathscr{P}'_{\text{EM}}[\boldsymbol{a}](\xi;\omega) = -\tfrac{1}{2}\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp};\boldsymbol{j}](\omega) = -\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{a}\perp};\boldsymbol{j}](\omega), \tag{4.93}$$

consistent with expectations of energy conservation. Note that while the extrapolated source-free solution $\boldsymbol{\chi}$ coincides, by construction, with $\boldsymbol{a}$ in the downstream region $\xi > \xi_1$,

and so exhibits identical out-flowing power, in order to actually satisfy the source-free equation, it clearly must have an equal magnitude of power flowing into the interaction region through any upstream plane at $\xi < \xi_0$, which accounts for the extra factor of two.

4.3.5 Basis-Set Approximation

As we have seen, equation (4.78) in principle gives the exact solution to the paraxial wave equation with prescribed sources, without any further approximations, and is compatible with energy conservation for paraxial electromagnetic fields. For given sources, the field at an arbitrary longitudinal plane ξ can be determined via a convolution integral over the longitudinal source position ξ' and a two-dimensional integral over either the transverse position-dependence (real space) or transverse wavevector-dependence (reciprocal space) of the solenoidal component of the current. In position space, we have

$$|\boldsymbol{\psi}(\xi;\omega)\rangle = \sum_s \int d^2\boldsymbol{r} \int d^2\boldsymbol{r}' \int_{\xi_0}^{\min(\xi,\xi_1)} d\xi' \, \frac{k\, e^{i\frac{k|\boldsymbol{r}-\boldsymbol{r}'|^2}{2(\xi-\xi')}}}{2\pi i(\xi-\xi')} \left\langle \boldsymbol{r}'; s \,\middle|\, \boldsymbol{f}(\xi';\omega) \right\rangle \, |\boldsymbol{r}; s\rangle \,, \tag{4.94}$$

where we have used the well-known result for the two-dimensional, position-space, free-particle propagator in quantum mechanics. In momentum space representation, the propagator is diagonal, and we have

$$|\boldsymbol{\psi}(\xi;\omega)\rangle = i \sum_s \int d^2\boldsymbol{p} \int_{\xi_0}^{\min(\xi,\xi_1)} d\xi' \, e^{-i(\xi-\xi')\frac{|\boldsymbol{p}|^2}{2k}} \left\langle \boldsymbol{p}; s \,\middle|\, \boldsymbol{f}(\xi';\omega) \right\rangle \, |\boldsymbol{p}; s\rangle \,, \tag{4.95}$$

where $\langle \boldsymbol{p}; s \,|\, \boldsymbol{f}(\xi';\omega)\rangle$ is just the Fourier transform (with respect to the scaled transverse wavevector) of the solenoidal current density slice at longitudinal position ξ'. Now, using either either of these approaches, we require something like ten integrations to determine a given frequency component of the field at a given spatial location, starting from the full current density as a function of space and time coordinates. In practice, we often only know the source probabilistically, so we will typically have to perform additional averages over the particle distribution function, involving up to six more integrations over the full particle phase space.

A complementary approach is to decompose the paraxial radiation fields into a complete, orthogonal set of modes. Because we are actually only interested in the radiation observed beyond the region of interaction with the localized sources, these modes are naturally chosen to satisfy the homogeneous (source-free) paraxial wave equation for all longitudinal positions $\xi > \xi_1$, but we may then also choose or extrapolate these modes to be free-space solutions

everywhere, including within and behind the support of the sources. That is, we consider a countable set of explicitly ξ-dependent, solenoidal, envelope modes $|\boldsymbol{u}_n(\xi;\omega)\rangle \in \mathcal{H}_\perp$, where n denotes some set of integral transverse-spatial and polarization modal indices, such that each mode envelope satisfies

$$i\tfrac{\partial}{\partial\xi}\,|\boldsymbol{u}_n(\xi;\omega)\rangle = H(\omega)\,|\boldsymbol{u}_n(\xi;\omega)\rangle\,, \tag{4.96}$$

or equivalently

$$|\boldsymbol{u}_n(\xi;\omega)\rangle = U(\xi,\xi';\omega)\,\big|\boldsymbol{u}_n(\xi';\omega)\big\rangle \tag{4.97}$$

together with the gauge constraint (4.60a), for all longitudinal positions $-\infty < \xi < \infty$. Because free-space propagation is unitary, these modes may be chosen to be orthonormal in each transverse plane, i.e.,

$$\langle \boldsymbol{u}_n(\xi;\omega)\,|\,\boldsymbol{u}_{n'}(\xi;\omega)\rangle = \delta_{n\,n'} \tag{4.98}$$

and complete in each transverse plane, in the sense that

$$\text{span}\,\big[\,|\boldsymbol{u}_n(\xi;\omega)\rangle\,\big] \cong \mathcal{H}_\perp. \tag{4.99}$$

or equivalently

$$\sum_n |\boldsymbol{u}_n(\xi;\omega)\,\rangle\langle\,\boldsymbol{u}_n(\xi;\omega)| = I_\perp, \tag{4.100}$$

where $I_\perp$ is the identity operator restricted to the solenoidal sub-space $\mathcal{H}_\perp$, or equivalently the Hermitian projection from $\mathcal{H}$ into $\mathcal{H}_\perp$. Familiar and convenient choices are the usual free-space Gauss-Hermite modes or Gauss-Laguerre modes in paraxial optics. The spatial profiles of these modes may be characterized by specifying the longitudinal location of the focal plane and the eigenvalues of certain Hermitian operators of which they are eigenfunctions in that plane. For example, in the focal plane, the components of the Gauss-Hermite modes are simultaneous number states of two harmonic oscillator Hamiltonians (one for each transverse Cartesian coordinate), while the Gauss-Laguerre modes are simultaneous eigenstates of a radial harmonic oscillator Hamiltonian and the longitudinal component of orbital angular momentum [122, 121]. Modes with slightly off-axis propagation directions can be defined in terms of the coherent states associated with these quadratic Hamiltonians, and many other generalizations are possible to suit particular problems.

Any linear combinations of the $\boldsymbol{u}_n$ modes satisfy the homogeneous paraxial wave equation, so cannot possibly represent the near fields in the region actually containing the sources, but they can exactly reproduce any fields in the "downstream" vacuum region beyond the sources, and then be extrapolated upstream. That is, for all $\xi > \xi_1$, the actual

radiation envelope $|\boldsymbol{\psi}(\xi;\omega)\rangle$ satisfies the homogeneous wave equation, lies entirely in the solenoidal sub-space $\mathcal{H}_\perp$, and may be decomposed as

$$\begin{aligned} |\boldsymbol{\psi}(\xi;\omega)\rangle &= \left[\sum_n |\boldsymbol{u}_n(\xi;\omega)\,\rangle\langle\,\boldsymbol{u}_n(\xi;\omega)|\right] |\boldsymbol{\psi}(\xi;\omega)\rangle \\ &= \sum_n \langle \boldsymbol{u}_n(\xi;\omega)\,|\,\boldsymbol{\psi}(\xi;\omega)\rangle\; |\boldsymbol{u}_n(\xi;\omega)\rangle\,, \end{aligned} \tag{4.101}$$

and both the modulus and argument of the complex expansion coefficients appearing in the sum have a simple interpretation. Using the orthonormality of the modes, the normalized power spectral density at a given transverse plane $\xi > \xi_1$ is given by

$$\mathscr{P}'_{\text{EM}}[\boldsymbol{a}](\xi;\omega) = \omega^2\,\langle \boldsymbol{\psi}(\xi;\omega)\,|\,\boldsymbol{\psi}(\xi;\omega)\rangle = \omega^2 \sum_n |\langle \boldsymbol{u}_n(\xi;\omega)\,|\,\boldsymbol{\psi}(\xi;\omega)\rangle|^2\,; \tag{4.102}$$

and, individually, each

$$\tfrac{\partial}{\partial\omega}\tilde{\mathscr{P}}_{\text{EM}}[\boldsymbol{u}_n](\xi;\omega) \equiv \omega^2\,|\langle \boldsymbol{u}_n(\xi;\omega)\,|\,\boldsymbol{\psi}(\xi;\omega)\rangle|^2 \geq 0 \tag{4.103}$$

provides the normalized power-spectral-density contribution from the nth mode, while

$$\theta_n(\xi;\omega) = \arg\big[\langle \boldsymbol{u}_n(\xi;\omega)\,|\,\boldsymbol{\psi}(\xi;\omega)\rangle\big] \tag{4.104}$$

determines the slowly-varying envelope phase (i.e., phase apart from the fast $k\xi - \omega\tau$ dependence) of the nth mode. From unitarity, it follows that $\mathscr{P}'_{\text{EM}}[\boldsymbol{u}_n](\xi;\omega)$, $\theta_n(\xi;\omega)$ and $\mathscr{P}'_{\text{EM}}[\boldsymbol{a}](\xi;\omega) = \sum_n \tfrac{\partial}{\partial\omega}\tilde{\mathscr{P}}_{\text{EM}}[\boldsymbol{u}_n](\xi;\omega)$ are all independent of ξ for any $\xi > \xi_1$. With perhaps a slight abuse of notation, we have used $\tfrac{\partial}{\partial\omega}\tilde{\mathscr{P}}_{\text{EM}}[\boldsymbol{u}_n]$ with a tilde to denote the scaled power (spectral density) contribution from the nth normalized mode; i.e., the power spectral density associated through Poynting's theorem with the ket $\langle \boldsymbol{u}_n(\xi;\omega)\,|\,\boldsymbol{\psi}(\xi;\omega)\rangle\; |\boldsymbol{u}_n(\xi;\omega)\rangle$, and not the power spectral density $\mathscr{P}'_{\text{EM}}[\boldsymbol{u}_n]$ literally associated with the ket $|\boldsymbol{u}_n(\xi;\omega)\rangle$, which is fixed at the value ω^2 by our normalization conventions.

In general, a countably infinite number of orthogonal modes are required to exactly describe the radiation fields, but if the modes are chosen appropriately, a finite basis set $\mathcal{B}_N$ of size N, $1 \leq N < \infty$, such that attention is confined to $n \in \mathcal{B}_N$, may provide a sufficiently accurate representation. For example, if the mode shape is approximately Gaussian in cross section, then it should be accurately represented by the fundamental and perhaps a few higher-order Gauss-Hermite modes, with proper choice of the waist size and location. Formally, the expansion coefficients may be determined by again using the Green function solution (4.78), taking advantage of the finite support of $|\boldsymbol{f}(\omega)\rangle$. In a derivation exactly

analogous to that in the previous section, we find that for any $\xi > \xi_1$,

$$\begin{aligned}\langle \boldsymbol{u}_n(\xi;\omega)\,|\,\boldsymbol{\psi}(\xi;\omega)\rangle &= i\int_{\xi_0}^{\min(\xi,\xi_1)} d\xi'\ \big\langle \boldsymbol{u}_n(\xi;\omega)\,\big|U(\xi,\xi';\omega)\big|\,\boldsymbol{f}(\xi';\omega)\big\rangle \\ &= i\int_{-\infty}^{\infty} d\xi'\ \big\langle \boldsymbol{u}_n(\xi';\omega)\,\big|\,\boldsymbol{f}(\xi';\omega)\big\rangle \\ &= i\int d\xi' \!\int d^2\boldsymbol{r}\ \boldsymbol{u}_n(\boldsymbol{r};\xi';\omega)^*\cdot\boldsymbol{f}(\boldsymbol{r};\xi';\omega),\end{aligned} \tag{4.105}$$

which apart from an overall constant is just a three-dimensional inner product, or overlap integral, between the solenoidal source envelope $\boldsymbol{f}(\boldsymbol{r};\xi';\omega)$ and the free-space mode envelope $\boldsymbol{u}_n(\boldsymbol{r};\xi';\omega)$. Equivalently, we can write this as the overlap integral between the full current density $\boldsymbol{j}(\boldsymbol{r};\xi';\omega)$, and the normalized electric field $\boldsymbol{\varepsilon}_{n\perp}(\xi';\omega) = i\omega\,\boldsymbol{u}_n(\boldsymbol{r};\xi';\omega)$ associated with the nth mode:

$$\begin{aligned}\langle \boldsymbol{u}_n(\xi;\omega)\,|\,\boldsymbol{\psi}(\xi;\omega)\rangle &= -\tfrac{1}{2\omega^2}\int d\xi' \!\int d^2\boldsymbol{r}\ \boldsymbol{\varepsilon}_{n\perp}(\boldsymbol{r};\xi;\omega)^*\cdot\boldsymbol{j}(\boldsymbol{r};\xi;\omega) \\ &= -\tfrac{1}{2\omega^2}\int d\xi'\ \big\langle \boldsymbol{\varepsilon}_{n\perp}(\xi';\omega)\,\big|\,\boldsymbol{j}(\xi';\omega)\big\rangle\,.\end{aligned} \tag{4.106}$$

So, again, we do not need to explicitly decompose the current density into its solenoidal and irrotational components in order to calculate these expansion coefficients. The factor of $\frac{1}{2}$ is analogous to that appearing in (4.89), and in fact (4.106) reduces to the full energy conservation result if we are lucky enough to choose a basis which contains the extrapolated solution $\boldsymbol{\chi}(\boldsymbol{r};\xi;\omega)$ in its linear span. More generally, it is clear that the finite-basis set approximation is just the orthogonal projection of the actual extrapolated free-space solution into the subspace $\operatorname*{span}_{n\in\mathcal{B}_N}\big[\,|\boldsymbol{u}_n(\xi;\omega)\rangle\,\big]$ spanned by the modes in $\mathcal{B}_N$:

$$|\boldsymbol{\nu}(\xi;\omega;\mathcal{B}_N)\rangle = \sum_{n\in\mathcal{B}_N} |\boldsymbol{u}_n(\xi;\omega)\,\rangle\langle\,\boldsymbol{u}_n(\xi;\omega)|\,|\boldsymbol{\chi}(\xi;\omega)\rangle\,. \tag{4.107}$$

From linearity, it immediately follows that if $|\boldsymbol{\nu}_m(\xi;\omega)\rangle$ is the basis-set approximation to $|\boldsymbol{\chi}_m(\xi;\omega)\rangle$ resulting form a source $|\boldsymbol{f}_m(\xi;\omega)\rangle$ then $\sum_m c_m(\omega)\,|\boldsymbol{\nu}_m(\xi;\omega)\rangle$ is the approximation to $\sum_m c_m(\omega)\,|\boldsymbol{\chi}_m(\xi;\omega)\rangle$ resulting from the source envelope $\sum_m c_m(\omega)\,|\boldsymbol{f}_m(\xi;\omega)\rangle$, where the $c_m(\omega)\in\mathbb{C}$ are arbitrary coefficients. So we may also determine the basis set approximation for any source by first finding the basis set approximation for the impulse source $|\boldsymbol{r};s\rangle$ and then performing a convolution over the actual source. That is, we can use as an approximate propagator, the restriction of the full Green operator to the subspace $\operatorname*{span}_{n\in\mathcal{B}_N}\big[\,|\boldsymbol{u}_n(\xi;\omega)\rangle\,\big]$:

$$|\boldsymbol{\nu}(\xi;\omega;\mathcal{B}_N)\rangle = i\int_{\xi_0}^{\xi_1} d\xi' \sum_{n\in\mathcal{B}_N} \big|\boldsymbol{u}_n(\xi;\omega)\,\big\rangle\big\langle\,\boldsymbol{u}_n(\xi';\omega)\big|\,\big|\boldsymbol{f}(\xi';\omega)\big\rangle\,, \tag{4.108}$$

which clearly agrees with previous results after exchanging the order of the integration and summation.

It follows from the properties of orthogonal projections in a Hilbert space that the projected basis-set approximation is also uniquely determined by the following constrained minimum-distance criteria:

$$|\boldsymbol{\nu}(\xi;\omega;\mathcal{B}_N)\rangle = \arg\min_{\boldsymbol{\nu}}\Big[\,\|\,|\boldsymbol{\nu}(\xi;\omega)\rangle - |\boldsymbol{\chi}(\xi;\omega)\rangle\,\|\Big] \tag{4.109a}$$

$$\text{s.t.}\;\; |\boldsymbol{\nu}(\xi;\omega)\rangle \in \operatorname*{span}_{n\in\mathcal{B}_N}\Big[|\boldsymbol{u}_n(\xi;\omega)\rangle\Big]. \tag{4.109b}$$

The estimated radiated power (spectral density) is then always a lower bound on the actual power (spectral density) in any plane $\xi > \xi_1$:

$$\mathscr{P}'_{\text{EM}}[\boldsymbol{\nu}](\xi;\omega;\mathcal{B}_N) = \sum_{n\in\mathcal{B}_N} \tfrac{\partial}{\partial\omega}\tilde{\mathscr{P}}_{\text{EM}}[\boldsymbol{u}_n](\xi;\omega) \leq \mathscr{P}'_{\text{EM}}[\boldsymbol{\chi}](\xi;\omega) = \mathscr{P}'_{\text{EM}}[\boldsymbol{a}](\xi;\omega), \tag{4.110}$$

with equality if and only if $\boldsymbol{\nu}(\boldsymbol{r};\xi;\omega;\mathcal{B}_N) = \boldsymbol{\chi}(\boldsymbol{r};\xi;\omega)$. In fact, it is straightforward to establish that, at the constrained optimum (4.109), the squared-distance between the basis-set approximation and the actual extrapolated fields is just proportional to their difference in power spectral density:

$$\omega^2\|\,|\boldsymbol{\nu}(\xi;\omega;\mathcal{B}_N)\rangle - |\boldsymbol{\chi}(\xi;\omega)\rangle\|^2 = \mathscr{P}'_{\text{EM}}[\boldsymbol{\chi}](\xi;\omega) - \mathscr{P}'_{\text{EM}}[\boldsymbol{\nu}](\xi;\omega;\mathcal{B}_N) \geq 0. \tag{4.111}$$

4.3.6 Variational Approximation

Such connections revealed in the basis-set approach between the orthogonal projection, minimum distance, maximum radiated power, minimal negative mechanical work, and maximal source/field overlap suggest that an approximate field profile for the radiation and a lower bound on the radiated power may be obtained directly through a general variational principle. Consider a parameterized family of trial envelope modes $\boldsymbol{v}(\boldsymbol{r};\xi;\omega;\boldsymbol{\alpha})$ depending on some set of adjustable parameters denoted by the vector $\boldsymbol{\alpha}$, where, for any fixed choice of allowed parameter values, we assume the trial mode satisfies the free-space wave equation, is solenoidal, and is normalized in any transverse plane. In principle, the trial mode can be written formally as a linear combination of the above basis modes:

$$|\boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha})\rangle = \sum_n c_n(\omega;\boldsymbol{\alpha})\,|\boldsymbol{u}_n(\xi;\omega)\rangle \tag{4.112}$$

for some complex expansion coefficients $c_n = c_n(\omega;\boldsymbol{\alpha}) \in \mathbb{C}$ which satisfy the constraint

$$\sum_n |c_n(\omega;\boldsymbol{\alpha})|^2 = 1, \tag{4.113}$$

so that

$$\langle \boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha}) \,|\, \boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha})\rangle = 1, \tag{4.114}$$

but are otherwise arbitrary. Note that the expansion coefficients are here assumed independent of ξ, but may in general depend on the adjustable parameters in some arbitrary or complicated, and perhaps nonlinear, fashion. (More generally, we could allow dependence on ξ in the c_n, leading to a variational problem involving the solution of a set of coupled ODEs rather than minimization of a scalar-valued function. This extra generalization is not really needed for the case of spontaneous radiation.) Using the Cauchy-Schwarz inequality, we have, for any ξ,

$$|\langle \boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha}) \,|\, \boldsymbol{\psi}(\xi;\omega)\rangle|^2 \leq \langle \boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha}) \,|\, \boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha})\rangle \, \langle \boldsymbol{\psi}(\xi;\omega) \,|\, \boldsymbol{\psi}(\xi;\omega)\rangle \,, \tag{4.115}$$

which using (4.114) simplifies to

$$\omega^2 \, |\langle \boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha}) \,|\, \boldsymbol{\psi}(\xi;\omega)\rangle|^2 \leq \omega^2 \big(\, \langle \boldsymbol{\psi}(\xi;\omega) \,|\, \boldsymbol{\psi}(\xi;\omega)\rangle \, \big), \tag{4.116}$$

or equivalently

$$\tfrac{\partial}{\partial\omega} \tilde{\mathscr{P}}_{\text{EM}}[\boldsymbol{v}](\xi;\omega;\boldsymbol{\alpha}) \leq \mathscr{P}'_{\text{EM}}[\boldsymbol{a}](\xi;\omega), \tag{4.117}$$

with strict equality if and only if $\boldsymbol{v}(\boldsymbol{r};\xi;\omega;\boldsymbol{\alpha})$ exactly mirrors the shape and polarization of the actual paraxial solution $\boldsymbol{\psi}(\boldsymbol{r};\xi;\omega)$ in the plane ξ, up to some overall phase; i.e.,

$$|\boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha})\rangle = e^{i\theta_{\boldsymbol{v}}(\omega;\boldsymbol{\alpha})} \big(\langle \boldsymbol{v}(\xi;\omega) \,|\, \boldsymbol{\psi}(\xi;\omega)\rangle \big)^{-1/2} \, |\boldsymbol{\psi}(\xi;\omega)\rangle \tag{4.118}$$

for some real angle $\theta_{\boldsymbol{v}}(\omega;\boldsymbol{\alpha})$.

We have indeed arrived at a variational principle for the paraxial radiation, where we may approximate both the relative spatial/polarization profile and overall amplitude of the radiation fields by maximizing the power-spectral density in a normalized trial mode $\boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha})$ measured in some post-source plane $\xi > \xi_1$, as a function of the variational parameters $\boldsymbol{\alpha}$ determining the trial mode's shape and polarization. That is, we take as the approximate radiation envelope

$$|\boldsymbol{\nu}(\xi;\omega;\tilde{\boldsymbol{\alpha}})\rangle = \big(\langle \boldsymbol{v}(\xi;\omega;\tilde{\boldsymbol{\alpha}}) \,|\, \boldsymbol{\psi}(\xi;\omega)\rangle \big) \, |\boldsymbol{v}(\xi;\omega;\tilde{\boldsymbol{\alpha}})\rangle \,, \tag{4.119}$$

where the optimal parameter vector $\tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\alpha}}[\boldsymbol{j}](\omega)$ is chosen so as to maximize

$$\mathscr{P}'_{\text{EM}}[\boldsymbol{\nu}](\xi;\omega;\boldsymbol{\alpha}) = \tfrac{\partial}{\partial\omega} \tilde{\mathscr{P}}_{\text{EM}}[\boldsymbol{v}](\xi;\omega;\boldsymbol{\alpha}) = \omega^2 \, |\langle \boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha}) \,|\, \boldsymbol{\psi}(\xi;\omega)\rangle|^2 \,. \tag{4.120}$$

The optimized mode shape is then the best guess, within the manifold of possibilities allowed by the shapes parameterized by $\boldsymbol{\alpha}$, of the actual paraxial field profile beyond the sources,

and the squared-norm yields a lower bound on the actual paraxial power spectral density of the radiation at the frequency under consideration.

However, as written, this variational principle appears vacuous at best, since if we knew the actual $\boldsymbol{\psi}(\boldsymbol{r};\xi;\omega)$ so as to be able to calculate its overlap with the trial field $\boldsymbol{v}(\boldsymbol{r};\xi;\omega;\boldsymbol{\alpha})$, there would of course be no need for a variational approximation in the first place. But from (4.106), we may also write

$$\begin{aligned}\langle \boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha}) \,|\, \boldsymbol{\psi}(\xi;\omega)\rangle &= -\tfrac{1}{2\omega^2}\int d\xi' \int d^2\boldsymbol{r}\; \boldsymbol{\varepsilon}_{\boldsymbol{v}\perp}(\boldsymbol{r};\xi;\omega;\boldsymbol{\alpha})^* \cdot \boldsymbol{j}(\boldsymbol{r};\xi;\omega) \\ &= -\tfrac{1}{2\omega^2}\int_{\xi_0}^{\xi_1} d\xi' \,\left\langle \boldsymbol{\varepsilon}_{\boldsymbol{v}\perp}(\xi';\omega;\boldsymbol{\alpha}) \,\middle|\, \boldsymbol{j}(\xi';\omega)\right\rangle\end{aligned} \tag{4.121}$$

where $\boldsymbol{\varepsilon}_{\boldsymbol{v}\perp}(\boldsymbol{r};\xi;\omega;\boldsymbol{\alpha}) = i\omega\,\boldsymbol{v}(\boldsymbol{r};\xi;\omega;\boldsymbol{\alpha})e^{ik\xi}$ is the normalized solenoidal frequency-domain electric field associated with the unit-norm trial vector potential envelope $\boldsymbol{v}(\boldsymbol{r};\xi;\omega;\boldsymbol{\alpha})$. In terms of the trial envelope $\boldsymbol{\nu}(\boldsymbol{r};\xi;\omega;\tilde{\boldsymbol{\alpha}})$, we see that

$$\begin{aligned}\mathscr{P}'_{\boldsymbol{\nu}}(\xi;\omega;\boldsymbol{\alpha}) &= \omega^2\,|\langle \boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha}) \,|\, \boldsymbol{\psi}(\xi;\omega;\boldsymbol{\alpha})\rangle|^2 = \omega^2\,\langle \boldsymbol{\nu}(\xi;\omega;\boldsymbol{\alpha}) \,|\, \boldsymbol{\nu}(\xi;\omega;\boldsymbol{\alpha})\rangle \\ &= -\tfrac{1}{2}\int d\xi' \,\left\langle \boldsymbol{\varepsilon}_{\boldsymbol{\nu}\perp}(\xi';\omega;\boldsymbol{\alpha}) \,\middle|\, \boldsymbol{j}(\xi';\omega)\right\rangle,\end{aligned} \tag{4.122}$$

where $\boldsymbol{\varepsilon}_{\boldsymbol{\nu}\perp}(\boldsymbol{r};\omega;\boldsymbol{\alpha}) = i\omega\,\boldsymbol{\nu}(\xi;\omega;\boldsymbol{\alpha})$ is the solenoidal electric field associated with the unnormalized trial envelope. Because the left-hand side is purely real and non-negative, the right hand side must already be so as well, so in fact:

$$\begin{aligned}\mathscr{P}'_{\boldsymbol{\nu}}(\xi;\omega;\boldsymbol{\alpha}) &= -\tfrac{1}{2}\,\mathrm{Re}\left[\int d\xi' \,\left\langle \boldsymbol{\varepsilon}_{\boldsymbol{\nu}\perp}(\xi';\omega;\boldsymbol{\alpha}) \,\middle|\, \boldsymbol{j}(\xi';\omega)\right\rangle\right] \\ &= -\tfrac{1}{2}\mathscr{P}'_{\mathrm{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\nu}\perp};\boldsymbol{j}](\omega;\boldsymbol{\alpha}).\end{aligned} \tag{4.123}$$

From the inequality (4.117) and the energy-conservation result (4.89) we arrive at an alternative formulation of the variational principle:

$$\begin{aligned}\mathscr{P}'_{\boldsymbol{\nu}}(\xi;\omega;\boldsymbol{\alpha}) &\le \mathscr{P}'_{\boldsymbol{\nu}}(\xi;\omega;\tilde{\boldsymbol{\alpha}}) = -\tfrac{1}{2}\mathscr{P}'_{\mathrm{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\nu}\perp};\boldsymbol{j}](\omega;\boldsymbol{\alpha}) \\ &\le \mathscr{P}'_{\boldsymbol{a}}(\xi;\omega) = -\mathscr{P}'_{\mathrm{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{a}\perp};\boldsymbol{j}](\omega) \\ &= -\tfrac{1}{2}\mathscr{P}'_{\mathrm{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp};\boldsymbol{j}](\omega).\end{aligned} \tag{4.124}$$

In addition, since

$$\langle \boldsymbol{\nu}(\xi;\omega;\boldsymbol{\alpha}) \,|\, \boldsymbol{\nu}(\xi;\omega;\boldsymbol{\alpha})\rangle = |\langle \boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha}) \,|\, \boldsymbol{\chi}(\xi;\omega)\rangle|^2 = \langle \boldsymbol{\nu}(\xi;\omega;\boldsymbol{\alpha}) \,|\, \boldsymbol{\chi}(\xi;\omega)\rangle\,, \tag{4.125}$$

for all ξ, it follows that the unnormalized trial envelope $\boldsymbol{\nu}(\boldsymbol{r};\xi;\omega;\tilde{\boldsymbol{\alpha}})$ (i.e., with the absolute amplitude included as an adjustable parameter) also minimizes the Hilbert-space distance

between the trial solution and the actual extrapolated paraxial solution in any transverse plane ξ :

$$\begin{aligned} \| \, |\boldsymbol{\nu}(\xi;\omega;\boldsymbol{\alpha})\,\rangle - |\boldsymbol{\chi}(\xi;\omega)\rangle \| &\geq \| \, |\boldsymbol{\nu}(\xi;\omega;\tilde{\boldsymbol{\alpha}})\rangle - |\boldsymbol{\chi}(\xi;\omega)\rangle \, \| \\ &= \omega^{-2} \left[\mathscr{P}'_{\boldsymbol{\chi}}(\xi;\omega) - \mathscr{P}'_{\boldsymbol{\nu}}(\xi;\omega;\tilde{\boldsymbol{\alpha}}) \right] \geq 0. \end{aligned} \tag{4.126}$$

The results (4.124) and (4.126) are obviously closely related to the results (4.110) and (4.109) derived in the basis-set expansion, which first led us to the variational formulation. Equivalently, we can express our variational principle most succinctly in terms of the unnormalized (amplitude-included) trial field which solves a constrained power-maximization problem for all parameters $\boldsymbol{\alpha}$ defining the trial-solution $\boldsymbol{\nu}(\boldsymbol{r};\xi;\omega;\tilde{\boldsymbol{\alpha}})$:

$$\tilde{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \left[\mathscr{P}'_{\boldsymbol{\nu}}(\xi;\omega;\boldsymbol{\alpha}) \right] \tag{4.127a}$$

$$\text{s.t. } \mathscr{P}'_{\boldsymbol{\nu}}(\xi;\omega;\boldsymbol{\alpha}) = -\tfrac{1}{2} \mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\nu}\perp};\boldsymbol{j}](\omega;\boldsymbol{\alpha}). \tag{4.127b}$$

These several equivalent formulations of the maximum-power variational principle (MPVP) are intuitively appealing, perhaps so much so that they might be guessed immediately, without the entire build-up of mathematical machinery. The variational approximation, which maximizes the resemblance (in the Hilbert-space metric), within the considered family of possibilities, to the actual paraxial field in the post-source region, or equivalently, to the extrapolated source-free paraxial field everywhere in space, can be obtained by maximizing the source/field spatial overlap $\left| \int d\xi' \, \langle \boldsymbol{\varepsilon}_{\boldsymbol{v}\perp}(\xi';\omega;\boldsymbol{\alpha}) \, | \, \boldsymbol{j}(\xi';\omega) \rangle \right|$, i.e., the magnitude of the three-dimensional inner product between the normalized source-free trial field and the actual current density. That is to say, our best approximation to the free-space radiation fields beyond the sources is found by maximizing the physical resemblance of these fields, when extrapolated backward into the region of the sources assuming free-space propagation, to the spatial profile of the sources.

Equivalently, we can say the optimal profile is that which, if it actually were to be incident on the sources, would experience the maximal small-signal gain (i.e., neglecting saturation effects or indeed any back-action on the sources), due to energy absorbed from those sources, and in this case the virtual "gross" power gain (i.e., due to stimulated emission, but neglecting the stimulated absorption) is numerically proportional to the estimated spectral density of spontaneously-radiated power. This approximation also optimizes the total radiated power spectral density in the full variational radiation envelope $\boldsymbol{\nu}(\boldsymbol{r};\xi;\omega;\boldsymbol{\alpha})$, consistent with the constraint, arising from energy-conservation, that this power could have arisen from work done by the sources, with appropriate compensation (i.e., scaling by $\frac{1}{2}$)

for the fact that we have replaced the inhomogeneous fields with their homogeneous extrapolation in the region of these sources.

Note that this variational principle is reminiscent of the Rayleigh-Ritz variational principle familiar from ordinary quantum mechanics, but despite the analogies between paraxial radiation propagation and non-relativistic quantum mechanics, these variational principles are actually distinct. In quantum mechanics, "the" Rayleigh-Ritz method is applicable to approximating the low-lying energy eigenvalues and corresponding stationary states of the homogeneous, time-independent Schrödinger equation, whereas the present variational principle is associated with solutions to an inhomogeneous, time-dependent Schrödinger equation.

However, the MPVP does share with the quantum-mechanical Raleigh-Ritz technique many of the same features and limitations common to variational approximations and optimization problems. Because we are determining a stationary point (in fact, a maximum) of a power (spectral density) functional, the actual value obtained for the power spectrum at the extremum is relatively insensitive to the precise shape of the trial field – generally, if the characteristic relative error in the latter is $O(\epsilon)$ in some sense, then the relative error in the former is $O(\epsilon^2)$, because the first-order variations must vanish. In order to estimate the power to say 1%, we need only match the variational parameters describing the field shape to about 10%. This is, of course, great news if we are actually most interested in estimating the power, but, conversely, much less satisfying if we are interested in determining, say, the spot size w_0, or some characterization of radiation spatial profile or the angular spectrum. This is analogous to the situation in quantum mechanics, where it is well known that the Raleigh-Ritz approximation estimates the ground-state energy eigenvalue more accurately than its corresponding wavefunction.

The variational approximation for the field profile is expected to be accurate to the extent that the trial mode can overlap, or mimic, the actual mode in any transverse plane $\xi > \xi_1$ in the free-space region beyond the sources, and obviously the approximation is expected to improve as the number of functionally-independent shape parameters is increased. Since $\boldsymbol{\nu}(\xi;\omega;\tilde{\boldsymbol{\alpha}})$ satisfies the free-space equation, it is actually an approximation throughout all ξ to the extrapolated envelope $\boldsymbol{\chi}(\boldsymbol{r};\xi;\omega)$. The corresponding power yields a lower bound for the actual power radiated, but in the absence of an upper bound, one does not have any foolproof means to estimate the closeness of this approximation to the actual power. We can consider a so-called "minimizing" sequence of variational problems where at each stage we expand the parametric family of trial envelopes to include additional possible

details of the field profiles which we anticipate might be present in the actual radiation but which could not be captured in the previous trial fields. If the extended family continues to include as a possibility the previous variational solution, then the estimated power will monotonically approach the actual (unknown) power from below, and the Hilbert-space distance between the trial envelope and the actual (unknown) envelope in any post-source plane will monotonically approach zero from above. In practice, we can stop when the difference (in power or Hilbert-space distance) upon adding finer details to the trial envelope becomes sufficiently small, or else when we have expended as much computational effort as we can spare. From above, we know that an additional parameter leading to an $O(\epsilon)$ relative increase in the power is expected to produce only an $O(\sqrt{\epsilon})$ relative improvement in the local field profile.

Although we formally expressed the normalized ket $|\boldsymbol{v}(\xi;\omega;\boldsymbol{\alpha})\rangle$ as a sum over the orthonormal modes $|\boldsymbol{u}_n(\xi;\omega;\boldsymbol{\alpha})\rangle$, we stress that the variational envelope can be explicitly written in any arbitrary form, with adjustable parameters which appear linearly or non-linearly, as long as it satisfies the homogeneous wave equation, gauge constraint, and normalization constraint. If the parameters $\boldsymbol{\alpha}$ represent the expansion coefficients in some finite linear sum of orthonormal modes, which is the assumption of the classic Raleigh-Ritz and Ritz-Galerkin techniques in the calculus of variations, then the variational approach formally reduces to the basis-set approach, but it is often more convenient or efficient in practice to use an explicitly nonlinear family of trial solutions where the variational parameters do not appear linearly. For example, one might use a Gaussian beam with indeterminate waist size and location, as discussed in the introduction, or perhaps even additional parameters to include skew or eccentricity or higher-order structure in the spot profile.

4.4 Non-Paraxial Generalization

When translated from the quantum mechanical into conventional electromagnetic notation, none of the energy-balance and variational results derived above in the paraxial limit depend in any essential way on the assumed paraxial nature of the fields. This suggests that analogous results should hold more generally, beyond the paraxial approximation, for any radiation emitted spontaneously by charges moving along prescribed trajectories. The essential features which emerged in the paraxial case, which we now abstract to more general problems, are: approximation of the actual radiation fields arising from the sources by homogeneous (i.e., free-space, or source-free) fields everywhere in space; a variational prin-

ciple, derived from the Cauchy-Schwarz inequality in a Hilbert space picture, mandating the maximization of (only) outgoing power spectral density; energy exchange between fields and charges which can be characterized as the three-dimensional inner product, or overlap integral, between the extrapolated, solenoidal electric field and either the full or solenoidal current density, as convenient; and, finally, a functional relationship between the mechanical work performed on/by the charges by/on the extrapolated free-space fields within a region and the outgoing radiation flux through a closed boundary surface, arising, physically, from energy conservation, and, mathematically, either explicitly from the formal solution or else from some form of the Fundamental Theorem of Calculus (e.g., Gauss's theorem or Green's Identities in multiple dimensions.)

4.4.1 Vector Spherical Harmonics and the Spherical Wave Basis

In order to generalize our results to the non-paraxial case, we first introduce a bit more mathematical notation and machinery. We use the scaled spatial position $\boldsymbol{\zeta}$, and introduce spherical coordinates: the scaled radius $\zeta = \|\boldsymbol{\zeta}\| \geq 0$, the polar angle θ, and the azimuthal angle φ, so that $\boldsymbol{\zeta} = \zeta\hat{\boldsymbol{\zeta}} = \zeta\hat{\boldsymbol{x}}$. We let R denote any (closed) spatial region in $\mathbb{R}^3$, and ∂R will denote its (orientable) boundary. For any position $\boldsymbol{\zeta}_0$ and any nonnegative $\zeta' \geq 0$, we define the closed ball of radius ζ' centered at $\boldsymbol{\zeta}_0$ as $B(\zeta';\boldsymbol{\zeta}_0) \equiv \{\boldsymbol{\zeta} \mid \|\boldsymbol{\zeta} - \boldsymbol{\zeta}_0\| \leq \zeta'\}$, and we will take $V(\zeta';\boldsymbol{\zeta}_0) \subset \mathbb{R}^3$ to denote a closed, simply-connected region satisfying $B(\zeta';\boldsymbol{\zeta}_0) \subseteq V(\zeta';\boldsymbol{\zeta}_0)$. For simplicity, whenever the central position coincides with the origin ($\boldsymbol{\zeta}_0 = \mathbf{0}$), the dependence will be suppressed in the notation; i.e., $B(\zeta') \equiv B(\zeta';\mathbf{0})$ and $V(\zeta') \equiv V(\zeta';\mathbf{0})$.

As in the paraxial case, we take the (scaled, frequency-domain) current density $\boldsymbol{j}(\boldsymbol{\zeta};\omega)$, assumed known, to be localized in space, vanishing identically outside some finite-radius ball $B(\zeta_1')$, for some $0 < \zeta_1' < \infty$. Again, the solenoidal component $\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega)$ will therefore not have have compact support in general, but because of the rapid falloff of its spatial dependence ($O(\zeta^{-2})$ as $\zeta \to \infty$), we may, with arbitrary small error in the calculated radiation-zone fields, neglect it beyond some sufficiently large but finite radius $\zeta_1 \geq \zeta_1'$. (Again, at the end of the calculation, we could attempt to take this support to be infinite, provided that the full current density continues to fall off sufficiently rapidly to ensure convergence of the integrals, and to enable somehow the unambiguous characterization of the "far-zone" versus non-radiative fields, when the sources extend to infinity. We will see one specific procedure to accomplish this below.)

It will prove very convenient to have some explicit basis in which to decompose the vector potential, at least for source-free solutions or in the radiation-zone of the actual sources. The plane-wave (spatial-Fourier) basis immediately comes to mind, but certain cumbersome singularities inevitably arise in this representation. Namely, whenever the vacuum dispersion relation is satisfied, i.e., $\|\boldsymbol{k}\| = \omega$, solutions to the homogeneous Helmholtz equation possess delta-function singularities, while solutions to the inhomogeneous solution exhibit second-order divergences whenever the corresponding Fourier components $\boldsymbol{j}_\perp(\boldsymbol{k},\omega) \neq \boldsymbol{0}$. In order to avoid such difficulties, we will instead use the spherical wave basis [92, 124, 125]. The conventional vector spherical harmonics are defined as

$$\boldsymbol{X}_{\ell m} = \boldsymbol{X}_{\ell m}(\hat{\boldsymbol{\zeta}}) = \boldsymbol{X}_{\ell m}(\theta,\varphi) = \tfrac{1}{\sqrt{\ell(\ell+1)}}\boldsymbol{L}Y_{\ell m}(\theta,\varphi), \tag{4.128}$$

where, in analogy to quantum mechanics, we have defined the Hermitian (scaled) orbital angular momentum operator as

$$\boldsymbol{L} \equiv \boldsymbol{Q}_{3\mathrm{D}} \times \boldsymbol{P}_{3\mathrm{D}} \equiv \boldsymbol{\zeta} \times \left(\tfrac{1}{i}\boldsymbol{\partial}\right) = -i\left(\boldsymbol{\zeta} \times \boldsymbol{\partial}\right), \tag{4.129}$$

which acts only on the angular degrees of freedom (i.e., not on the radial or the polarization degrees-of-freedom),

$$\boldsymbol{L} = \tfrac{1}{\sqrt{2}}\hat{\boldsymbol{e}}_+ e^{-i\varphi}\left(-\tfrac{\partial}{\partial\theta} + i\cot(\theta)\tfrac{\partial}{\partial\varphi}\right) + \tfrac{1}{\sqrt{2}}\hat{\boldsymbol{e}}_- e^{+i\varphi}\left(+\tfrac{\partial}{\partial\theta} + i\cot(\theta)\tfrac{\partial}{\partial\varphi}\right) - i\hat{\boldsymbol{z}}\tfrac{\partial}{\partial\varphi}; \tag{4.130}$$

and the functions $Y_{\ell m}(\theta,\varphi)$ denote the usual scalar spherical harmonics for nonnegative integers $\ell = 0, 1, 2, \ldots$, and integer $m = -\ell, -\ell+1, \ldots, 0, \ldots, \ell-1, \ell$. In addition, we define two other related sets of basis vector fields:

$$\boldsymbol{Z}_{\ell m} = \boldsymbol{Z}_{\ell m}(\hat{\boldsymbol{\zeta}}) = \boldsymbol{Z}_{\ell m}(\theta,\varphi) = \hat{\boldsymbol{\zeta}} \times \boldsymbol{X}_{\ell m}, \tag{4.131}$$

and

$$\boldsymbol{N}_{\ell m} = \boldsymbol{N}_{\ell m}(\hat{\boldsymbol{\zeta}}) = \boldsymbol{N}_{\ell m}(\theta,\varphi) = \hat{\boldsymbol{\zeta}}\, Y_{\ell m}. \tag{4.132}$$

(By definition, we take $\boldsymbol{X}_{0\,0} = \boldsymbol{Z}_{0\,0} \equiv \boldsymbol{0}$ identically, ultimately reflecting the fact that spherically-symmetric solutions to the free-space Maxwell equations can exist only in the non-radiative $\omega \to 0$ limit.)

The orbital angular momentum operator $\boldsymbol{L}$ effects infinitesimal (inverse) rotations of the spatial degrees of freedom, i.e., observation position or field point, and satisfies the useful

identities:

$$\hat{\boldsymbol{\zeta}} \cdot \boldsymbol{L} \ = 0, \tag{4.133a}$$

$$\boldsymbol{L} \times \boldsymbol{L} = i\boldsymbol{L}, \tag{4.133b}$$

$$\left[L^2, \boldsymbol{L}\right] = \boldsymbol{0}, \tag{4.133c}$$

$$\left[P_{\text{3D}}^2, \boldsymbol{L}\right] = \boldsymbol{0}, \tag{4.133d}$$

$$L^2 = r\tfrac{\partial^2}{\partial r^2}\left(r \quad\right) + r^2 P_{\text{3D}}^2, \tag{4.133e}$$

$$\boldsymbol{L} \cdot (\ \ ) = i\boldsymbol{\partial} \cdot [\boldsymbol{\zeta} \times (\ \ )] = -i\boldsymbol{\zeta} \cdot [\boldsymbol{\partial} \times (\ \ )]\,. \tag{4.133f}$$

The scalar spherical harmonics themselves may be interpreted as simultaneous eigenstates of L^2 and L_z. Each of the basis vector fields may then be interpreted as a simultaneous eigenstate of $J^2 = J_x^2 + J_y^2 + J_z^2$ and J_z for some spin-one object, where $\boldsymbol{J} = \boldsymbol{L} + \boldsymbol{S}$ is the total (spatial plus spin/polarization) angular momentum operator. For fixed ℓ and fixed $\boldsymbol{W} = \boldsymbol{X}$, $\boldsymbol{Z}$, or $\boldsymbol{N}$, the members $\boldsymbol{W}_{\ell m}(\theta, \varphi)$ of each of these three families therefore transform amongst themselves under total spatial rotations (of both spatial and polarization degrees of freedom). The spin angular momentum operator $\boldsymbol{S}$ generates rotations of the polarization vector, and in the Cartesian basis, $\boldsymbol{S} = \hat{\boldsymbol{x}} S_1 + \hat{\boldsymbol{y}} S_2 + \hat{\boldsymbol{z}} S_3$ may be defined as a vector of 3×3 Hermitian matrices S_j for $j = 1, 2, 3$, whose components are given by

$$(S_j)_{kn} = -i\epsilon_{jkn}, \tag{4.134}$$

where ϵ_{jkn} is the completely antisymmetric Levi-Civita tensor in three dimensions. From this definition one can prove the useful operator identity

$$\boldsymbol{\partial} \times \ \equiv \boldsymbol{P}_{3D} \cdot \boldsymbol{S}. \tag{4.135}$$

Using various vector identities, these spherical vector harmonic basis fields also can be shown to satisfy:

$$\boldsymbol{L} \cdot \boldsymbol{X}_{\ell m} = \sqrt{\ell(\ell+1)}\, Y_{\ell m}, \tag{4.136a}$$

$$\boldsymbol{L} \cdot \boldsymbol{Z}_{\ell m} = 0, \tag{4.136b}$$

$$\boldsymbol{L} \cdot \boldsymbol{N}_{\ell m} = 0; \tag{4.136c}$$

and

$$\boldsymbol{L} \times \boldsymbol{X}_{\ell m} = i\boldsymbol{X}_{\ell m}, \tag{4.137a}$$

$$\boldsymbol{L} \times \boldsymbol{Z}_{\ell m} = \sqrt{\ell(\ell+1)}\, \boldsymbol{N}_{\ell m}, \tag{4.137b}$$

$$\boldsymbol{L} \times \boldsymbol{N}_{\ell m} = \ \boldsymbol{0}; \tag{4.137c}$$

as well as

$$\hat{\boldsymbol{\zeta}} \cdot \boldsymbol{X}_{\ell m} = 0, \tag{4.138a}$$

$$\hat{\boldsymbol{\zeta}} \cdot \boldsymbol{Z}_{\ell m} = 0, \tag{4.138b}$$

$$\hat{\boldsymbol{\zeta}} \cdot \boldsymbol{N}_{\ell m} = Y_{\ell m}; \tag{4.138c}$$

and

$$\hat{\boldsymbol{\zeta}} \times \boldsymbol{X}_{\ell m} = \quad \boldsymbol{Z}_{\ell m}, \tag{4.139a}$$

$$\hat{\boldsymbol{\zeta}} \times \boldsymbol{Z}_{\ell m} = -\boldsymbol{X}_{\ell m}, \tag{4.139b}$$

$$\hat{\boldsymbol{\zeta}} \times \boldsymbol{N}_{\ell m} = \quad \boldsymbol{0}; \tag{4.139c}$$

and also

$$\boldsymbol{\partial} \cdot \boldsymbol{X}_{\ell m} = \quad 0, \tag{4.140a}$$

$$\boldsymbol{\partial} \cdot \boldsymbol{Z}_{\ell m} = -\tfrac{i}{\zeta}\sqrt{\ell(\ell+1)}Y_{\ell m}, \tag{4.140b}$$

$$\boldsymbol{\partial} \cdot \boldsymbol{N}_{\ell m} = \quad \tfrac{2}{\zeta}Y_{\ell m}; \tag{4.140c}$$

together with

$$\boldsymbol{\partial} \times \boldsymbol{X}_{\ell m} = \quad \tfrac{1}{\zeta}\left(i\sqrt{\ell(\ell+1)}\boldsymbol{N}_{\ell m} + \boldsymbol{Z}_{\ell m}\right), \tag{4.141a}$$

$$\boldsymbol{\partial} \times \boldsymbol{Z}_{\ell m} = -\tfrac{1}{\zeta}\boldsymbol{X}_{\ell m}, \tag{4.141b}$$

$$\boldsymbol{\partial} \times \boldsymbol{N}_{\ell m} = -\tfrac{i}{\zeta}\sqrt{\ell(\ell+1)}\boldsymbol{X}_{\ell m}. \tag{4.141c}$$

For any allowed ℓ and m, the vector spherical harmonics are geometrically orthogonal as vectors at every point $\hat{\boldsymbol{\zeta}}$ on the unit sphere:

$$\boldsymbol{X}_{\ell m}(\hat{\boldsymbol{\zeta}})^* \cdot \boldsymbol{Z}_{\ell m}(\hat{\boldsymbol{\zeta}}) = \boldsymbol{X}_{\ell m}(\hat{\boldsymbol{\zeta}})^* \cdot \boldsymbol{N}_{\ell m}(\hat{\boldsymbol{\zeta}}) = \boldsymbol{Z}_{\ell m}(\hat{\boldsymbol{\zeta}})^* \cdot \boldsymbol{N}_{\ell m}(\hat{\boldsymbol{\zeta}}) = 0; \tag{4.142}$$

are (Hilbert-space) orthonormal when integrated over solid angle (except in the trivial case where both vector fields identically vanish for $\ell = 0$):

$$\int d\theta\, \sin\theta \int d\varphi\, \boldsymbol{W}_{\ell m}(\theta,\varphi)^* \cdot \boldsymbol{W}'_{\ell' m'}(\theta,\varphi) = \delta_{\boldsymbol{W}\boldsymbol{W}'}\delta_{\ell\ell'}\delta_{m m'}\left(1-\delta_{\boldsymbol{W}\boldsymbol{X}}\delta_{\ell 0}\right)\left(1-\delta_{\boldsymbol{W}\boldsymbol{Z}}\delta_{\ell 0}\right); \tag{4.143}$$

and also collectively constitute a complete set for vector fields defined on the unit sphere:

$$\sum_{\boldsymbol{W}=\boldsymbol{X},\boldsymbol{Z},\boldsymbol{N}} \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \boldsymbol{W}_{\ell m}(\theta,\varphi) \cdot \boldsymbol{W}_{\ell m}(\theta',\varphi')^\dagger = \delta(\cos\theta - \cos\theta')\delta(\varphi-\varphi')I_3, \tag{4.144}$$

where I_3 is the identity matrix on the polarization degrees of freedom.

Now, outside the effective support of the solenoidal sources (i.e., strictly speaking, asymptotically as $\|\boldsymbol{\zeta}\| \to \infty$, but in practice for any $\zeta \geq \zeta_2 > \zeta_1$), the actual Coulomb-gauge vector potential $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$ is assumed to satisfy the homogeneous Helmholtz equation, and the method of separation-of-variables may be used in spherical coordinates to show that there the solution may always be written in the form:

$$\boldsymbol{a}(\boldsymbol{\zeta};\omega) = \sum_{\ell,\,m} \tfrac{1}{\omega^2} a^{\mathrm{E}}_{\ell m}(\omega) \boldsymbol{\partial} \times \left[h^{(1)}_\ell(k\zeta) \boldsymbol{X}_{\ell m}(\theta,\varphi) \right] - \tfrac{i}{\omega} a^{\mathrm{M}}_{\ell m}(\omega) h^{(1)}_\ell(k\zeta) \boldsymbol{X}_{\ell m}(\theta,\varphi), \quad (4.145)$$

where the spherical Hankel functions of the first and second kind,

$$h^{(1)}_\ell(x) = j_\ell(x) + i\, n_\ell(x), \quad (4.146a)$$

$$h^{(2)}_\ell(x) = j_\ell(x) - i\, n_\ell(x), \quad (4.146b)$$

are defined in terms of the usual spherical Bessel functions,

$$j_\ell(x) = \quad (-x)^\ell \left(\frac{1}{x} \frac{d}{dx} \right)^\ell \left(\frac{\sin x}{x} \right), \quad (4.147)$$

(which should not be confused by notational similarity with any current density), and the spherical Neumann functions,

$$n_\ell(x) = -(-x)^\ell \left(\frac{1}{x} \frac{d}{dx} \right)^\ell \left(\frac{\cos x}{x} \right). \quad (4.148)$$

Any two of these four families of special functions may be taken as the linearly-independent solutions to the radial differential equation obtained through separation of variables. The appearance of only the $h^{(1)}_\ell(kr) \propto \frac{e^{ik\zeta}}{k\zeta}$ in the radial part of our formal solution reflects the facts that the solution holds in the extra-source region $\zeta > \zeta_1$ which excludes the origin, and that we have imposed the so-called Sommerfeld boundary conditions, requiring only outgoing radiation at spatial infinity.

Each term in the superposition (4.145) is everywhere divergenceless, so the total vector field is solenoidal, as required. The expansion coefficients $a^{\mathrm{E}}_{\ell m}(\omega) \in \mathbb{C}$ and $a^{\mathrm{M}}_{\ell m}(\omega) \in \mathbb{C}$ are, respectively, the (scaled) electric and magnetic multipole moments of the transverse current density $\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega)$ with respect to the origin, and are uniquely determined by quadratures over $\boldsymbol{j}$ and its derivatives, or equivalently, by field values on some bounding sphere in the far-field (or, in fact, by just the radial components of the electric and magnetic fields on any two spherical shells outside the sources.) Given just vector potential data on some spherical surface $\|\boldsymbol{\zeta}\| = \zeta_3 > \zeta_1$, these coefficients can be easily determined using the orthogonality

properties of the vector spherical harmonics:

$$a^{\mathrm{M}}_{\ell m}(\omega) = \quad \tfrac{i\omega}{h_\ell(k\zeta_3)} \int d\theta\, \sin\theta \int d\varphi\, bv X^*_{\ell m} \cdot \boldsymbol{a}(\zeta_3,\theta,\varphi;\omega), \tag{4.149a}$$

$$a^{\mathrm{E}}_{\ell m}(\omega) = -i\tfrac{\omega^2\zeta_3}{\sqrt{\ell(\ell+1)}h_\ell(k\zeta_3)} \int d\theta\, \sin\theta \int d\varphi\, bv N^*_{\ell m} \cdot \boldsymbol{a}(\zeta_3,\theta,\varphi;\omega). \tag{4.149b}$$

Using the far-field asymptotic behavior, angular momentum selection rules, certain Green function expansions discussed below, and a bit of algebra, one can express these coefficients directly in terms of the transverse current density:

$$a^{\mathrm{M}}_{\ell m}(\omega) = -\frac{\omega^2}{\sqrt{\ell(\ell+1)}} \sum_{m'=\min[0,m-1]}^{\max[m+1,\ell]} \boldsymbol{M}_{\ell m'}(\omega) \cdot \int d\theta\, \sin\theta \int d\varphi\, Y_{\ell m}(\theta,\varphi)^* \boldsymbol{L} Y_{\ell m'}(\theta,\varphi), \tag{4.150a}$$

$$a^{\mathrm{E}}_{\ell m}(\omega) = \quad i\frac{\omega}{\sqrt{\ell(\ell+1)}} \sum_{m'=\min[0,m-1]}^{\max[m+1,\ell]} \boldsymbol{R}_{\ell m'}(\omega) \cdot \int d\theta\, \sin\theta \int d\varphi\, Y_{\ell m}(\theta,\varphi)^* \boldsymbol{L} Y_{\ell m'}(\theta,\varphi); \tag{4.150b}$$

where we have defined

$$\boldsymbol{M}_{\ell m}(\omega) = \int d^3\boldsymbol{\zeta}\, \left[j_\ell(k\zeta) Y_{\ell m}(\hat{\boldsymbol{\zeta}})^* \right] \boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega), \tag{4.151a}$$

$$\boldsymbol{R}_{\ell m}(\omega) = \int d^3\boldsymbol{\zeta}\, \left[j_\ell(k\zeta) Y_{\ell m}(\hat{\boldsymbol{\zeta}})^* \right] \boldsymbol{\partial}\times\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega). \tag{4.151b}$$

Still simpler expressions can be obtained from these by direct manipulations, or else more efficiently, by recalling that $\boldsymbol{L}$ and ∂^2 commute, so that

$$\left(\partial^2 + \omega^2\right) \left[\boldsymbol{L}\cdot\boldsymbol{a}(\boldsymbol{\zeta};\omega)\right] = \boldsymbol{L}\cdot\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega). \tag{4.152}$$

Using spherical wave decompositions for the scalar $\boldsymbol{L}\cdot\boldsymbol{a}$ in addition to the vector $\boldsymbol{a}$ itself, and using various identities, it can be verified that

$$a^{\mathrm{M}}_{\ell m}(\omega) = -\frac{\omega^2}{\sqrt{\ell(\ell+1)}} \int d^3\boldsymbol{\zeta}\, j_\ell(k\zeta) Y_{\ell m}(\hat{\boldsymbol{\zeta}})^* \left[\boldsymbol{L}\cdot\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega)\right], \tag{4.153a}$$

$$a^{\mathrm{E}}_{\ell m}(\omega) = \quad i\frac{\omega}{\sqrt{\ell(\ell+1)}} \int d^3\boldsymbol{\zeta}\, j_\ell(k\zeta) Y_{\ell m}(\hat{\boldsymbol{\zeta}})^* \boldsymbol{L}\cdot\left[\boldsymbol{\partial}\times\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega)\right]. \tag{4.153b}$$

But $\boldsymbol{L}\cdot\boldsymbol{j}_\perp = i\boldsymbol{\zeta}\cdot[\boldsymbol{\partial}\times\boldsymbol{j}_{perp}]$ and $\boldsymbol{\partial}\times\boldsymbol{j}_\perp = \boldsymbol{\partial}\times\boldsymbol{j}$, so these last expressions for the multipole moments can be simplified further to obtain:

$$a^{\mathrm{M}}_{\ell m}(\omega) = -i\frac{\omega^2}{\sqrt{\ell(\ell+1)}} \int d^3\boldsymbol{\zeta}\, j_\ell(k\zeta) Y_{\ell m}(\hat{\boldsymbol{\zeta}})^* \boldsymbol{\zeta}\cdot\left[\boldsymbol{\partial}\times\boldsymbol{j}(\boldsymbol{\zeta};\omega)\right], \tag{4.154a}$$

$$a^{\mathrm{E}}_{\ell m}(\omega) = \quad \frac{\omega}{\sqrt{\ell(\ell+1)}} \int d^3\boldsymbol{\zeta}\, j_\ell(k\zeta) Y_{\ell m}(\hat{\boldsymbol{\zeta}})^* \boldsymbol{\zeta}\cdot\left[\boldsymbol{\partial}\times\boldsymbol{\partial}\times\boldsymbol{j}(\boldsymbol{\zeta};\omega)\right]. \tag{4.154b}$$

Various other equivalent forms are possible, depending on precisely what data are used and where, but explicit forms for the multipole expansion coefficients will not actually be

needed here. (After all, if they were known or easily calculable, we would not need to resort to any sort of approximation.) However, equations (4.153) reveal an important fact, namely that the multipole moments determining the Coulomb-gauge vector potential in the extra-source region can be expressed in terms of the total current density $\boldsymbol{j}$ and its curl, rather than just the solenoidal component $\boldsymbol{j}_\perp$. This means that, in principle, these expansion coefficients may be determined without explicit decomposition of $\boldsymbol{j}$ into its solenoidal and irrotational parts, and without any specific (but then somewhat arbitrary) choice for the precise spatial cutoff for the effective support of $\boldsymbol{j}_\perp$.

The expansion (4.145) for the vector potential, with coefficients formally determined by (4.154) or other means, represents the full contribution to the outgoing radiation from the given sources asymptotically as $\|\boldsymbol{\zeta}\| \to \infty$. At finite distances outside the effective support of the solenoidal source, i.e., for $\zeta_1 < \|\boldsymbol{\zeta}\| < \infty$, strictly speaking it converges to the Coulomb-gauge vector potential only up to "near-field" errors of size $O(1/\|\boldsymbol{\zeta}\|)$ (which can be made as small as we wish in a relative sense by choosing ζ_2 sufficiently large), but it still does contain all radiative contributions which will survive into the far field if actually allowed to evolve via free-space propagation, and contains no further, spurious radiative components. In fact, this expansion is a well-behaved, solenoidal solution to the source-free wave equation everywhere in space except for the origin, i.e., for any $\boldsymbol{\zeta} \neq \mathbf{0}$. It actually represents exactly (in all zones) the transverse part of the vector potential associated with an idealized singular source possessing the same multipole moments as the actual source $\boldsymbol{j}(\boldsymbol{\zeta};\omega)$, but concentrated into a single point at the origin $\boldsymbol{\zeta} = \mathbf{0}$.

As such, the expansion (4.145) is a natural (if not unique) extension, or extrapolation, of the outgoing radiation fields (unambiguously defined only for $\|\boldsymbol{\zeta}\| \to \infty$) to arbitrary non-zero but finite distances from the origin, i.e., for $0 < \|\boldsymbol{\zeta}\| \leq \infty$, based on imagining the actual radiation-zone fields of the given source are produced instead by a single equivalent point-source, so that the fields satisfy the source-free wave equation almost everywhere. This is the best one can do by extrapolating purely outgoing radiation fields according to free-space propagation – they must have some source somewhere. Conversely, as we will see in detail below, source-free solutions must have matching contributions of incoming and outgoing radiation. While convenient, this point-multipole extrapolation clearly is not unique, because we could choose this effective point-source to be essentially anywhere, but mathematically it is convenient to have it coincide with the origin of our coordinate system, and physically it is convenient to locate it somewhere near the geometric center of the actual source.

Note that time-reversed, or equivalently conjugate, or incoming-wave solutions can obviously be written in an analogous manner, with the $h_\ell^{(1)}(k\zeta)$ simply replaced with the corresponding $h_\ell^{(2)}(k\zeta) \propto \frac{e^{-ik\zeta}}{k\zeta}$ in the superposition of spherical waves. This expansion also admits a corresponding point-multipole extrapolation, although the singularity at the origin represents an absorber (sink) rather than emitter (source). Any solenoidal solution to the source-free wave equation, in particular a source-free "trial" vector potential $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha})$ to appear in the variational principle, satisfies the homogeneous equation everywhere in space, and may be expressed, for any $\boldsymbol{\zeta}$, in a similar spherical-wave form:

$$\boldsymbol{\chi}(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha}) = \sum_{\ell,\,m} \tfrac{1}{\omega^2}\chi^{\mathrm{E}}_{\ell\,m}(\omega;\boldsymbol{\alpha})\boldsymbol{\partial}\times\big[j_\ell(k\zeta)\boldsymbol{X}_{\ell\,m}(\theta,\varphi)\big] + -\tfrac{i}{\omega}\chi^{\mathrm{M}}_{\ell\,m}(\omega;\boldsymbol{\alpha})j_\ell(k\zeta)\boldsymbol{X}_{\ell\,m}(\theta,\varphi). \tag{4.155}$$

The use of $j_\ell(k\zeta)$ is the only choice for the radial dependence which solves the Helmholtz equation but is regular (i.e., non-singular) everywhere in space, including the origin. Each term in (4.155) is also automatically divergenceless everywhere, so their sum (when convergent) is solenoidal. The complex expansion coefficients $\chi^{\mathrm{E}}_{\ell\,m}(\omega;\boldsymbol{\alpha})$ and $\chi^{\mathrm{M}}_{\ell\,m}(\omega;\boldsymbol{\alpha})$ for this source-free solution may nevertheless be interpreted as some (scaled) multipole moments characterizing what we might term an "effective source/sink" $\boldsymbol{g}(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha})$ acting both in retarded fashion as an emitter, and in an advanced (i.e., time-reversed) fashion as an absorber, or otherwise may just be regarded as free expansion coefficients determined by (or even taken to be equal to) the actual variational parameters $\boldsymbol{\alpha}$ appearing in $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha})$. The familiar expansions

$$e^{i\boldsymbol{k}\cdot\boldsymbol{\zeta}} = 4\pi\sum_{\ell=0}^{\infty} i^\ell j_\ell(k\zeta)\sum_{m=-\ell}^{\ell} Y_{\ell\,m}(\hat{\boldsymbol{\zeta}})^* Y_{\ell\,m}(\hat{\boldsymbol{k}}) = \sum_{\ell=0}^{\infty} i^\ell\sqrt{4\pi(2\ell+1)}\,j_\ell(k\zeta)Y_{1\,0}(\arccos(\hat{\boldsymbol{\zeta}}\cdot\hat{\boldsymbol{k}}),0), \tag{4.156}$$

for any fixed (scaled) wavevector $\boldsymbol{k} = k\hat{\boldsymbol{k}}$, may be used to establish the connection between the plane wave and spherical wave representations, and to confirm that they both span the same space of transverse, source-free solutions.

4.4.2 Free-Space Green Functions

Alternatively, the exact solution to the inhomogeneous problem at any position $\boldsymbol{\zeta}$ may be expressed in terms of a convolution over the retarded Green function:

$$G^{\mathrm{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) = \frac{e^{+ik\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|}}{4\pi\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|}, \tag{4.157}$$

which is invariant under simultaneous translations, rotations, or reflections of both spatial arguments, and satisfies the scalar Helmholtz equation with impulsive source (and with

negative unit charge, conforming to convention for electromagnetic problems),

$$\left(\partial^2 + \omega^2\right) G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) = -\delta(\boldsymbol{\zeta}-\boldsymbol{\zeta}'), \tag{4.158}$$

together with the Sommerfeld asymptotic radiation conditions for outgoing waves:

$$G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) = O(1/\zeta) \text{ as } \zeta \to \infty \tag{4.159a}$$

$$\left(\tfrac{\partial}{\partial\zeta} - ik\right) G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) = o(1/\zeta) \text{ as } \zeta \to \infty, \tag{4.159b}$$

for any fixed emitter position $\boldsymbol{\zeta}'$. (Because we are here always working with solenoidal sources in otherwise free space, we can use a scalar rather than dyadic Green function [126].) The retarded Green function $G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)$, then represents an outgoing (scalar) spherical wave at the observation position $\boldsymbol{\zeta}$ emanating from a harmonic point source at $\boldsymbol{\zeta}'$. (By symmetry, analogous conditions and interpretations obviously hold true under permutation of the two spatial positions $\boldsymbol{\zeta}$ and $\boldsymbol{\zeta}'$, from which follow various well-known electromagnetic reciprocity results such as the Raleigh-Carson and Lorentz theorems.) The advanced Green function

$$G^{\text{adv}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) = \left[G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\right]^* = \frac{e^{-ik\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|}}{4\pi\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|}, \tag{4.160}$$

satisfies the same impulsive Helmholtz equation, and in fact may be interpreted as the time-reversal of the retarded impulse response, representing an incoming spherical wave, observed at $\boldsymbol{\zeta}$, and ultimately converging to a point absorber at $\boldsymbol{\zeta}'$. Given any current density $\boldsymbol{j}(\boldsymbol{\zeta}';\omega)$ with compact support in the neighborhood $V(\zeta_1)$ of the origin, and supposing no incoming radiation from distant sources by assumption, the solenoidal component $\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega)$ can result only in outgoing radiation asymptotically as $\zeta \to \infty$, so the Coulomb-gauge vector potential can always be expressed as:

$$\boldsymbol{a}(\boldsymbol{\zeta};\omega) = \int d^3\boldsymbol{\zeta}'\, G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega) = \int\limits_{V(\zeta_1)} d^3\boldsymbol{\zeta}'\, G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega). \tag{4.161}$$

Clearly, this satisfies the appropriate inhomogeneous Helmholtz equation:

$$\begin{aligned}(\partial^2+\omega^2)\boldsymbol{a}(\boldsymbol{\zeta};\omega) &= \int d^3\boldsymbol{\zeta}'\,(\partial^2+\omega^2)G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega)\\ &= -\int d^3\boldsymbol{\zeta}'\,\delta(\boldsymbol{\zeta}-\boldsymbol{\zeta}')\,\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega), = -\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega)\end{aligned} \tag{4.162}$$

and also shares with $G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)$ the appropriate outgoing-wave Sommerfeld boundary conditions. Because $G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)$ is radially symmetric, depending on the spatial coordinates only through $\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|$, and because $\boldsymbol{\partial}\cdot\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega) = 0$ by construction, we have

$$\begin{aligned}\boldsymbol{\partial}\cdot\left[G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega)\right] &= \boldsymbol{\partial}G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\cdot\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega) = -\boldsymbol{\partial}'G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\cdot\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega)\\ &= -\boldsymbol{\partial}'\cdot\left[G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega)\right].\end{aligned} \tag{4.163}$$

Since $G^{\rm ret}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) = O(1/\left|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\right|)$ and $\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega) \lesssim O(1/\zeta^2)$ as $\zeta \to \infty$, Gauss's law then implies

$$\begin{aligned}\boldsymbol{\partial}\cdot\boldsymbol{a}(\boldsymbol{\zeta};\omega) &= \int d^3\boldsymbol{\zeta}'\,\boldsymbol{\partial}\cdot[G^{\rm ret}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega)] = -\int d^3\boldsymbol{\zeta}'\,\boldsymbol{\partial}'\cdot[G^{\rm ret}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega)] \\ &= -\lim_{\zeta_2\to\infty}\int\limits_{V(\zeta_2)} d\sigma'\hat{\boldsymbol{n}}\cdot[G^{\rm ret}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega)] = 0\end{aligned} \tag{4.164}$$

so indeed $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$ is solenoidal, as required. (Subsequently, we will make use of a number of similar multidimensional "integrations by parts," relying on the sufficiently rapid fall-off of $\boldsymbol{j}_\perp$ and radiative fall-off of $G^{\rm ret}$, but for brevity we will not always provide the explicit demonstrations.)

Outside the effective support of sources, say for $\zeta \geq \zeta_2 > \zeta_1$, the solution to the inhomogeneous Helmholtz equation may also be written in terms of a surface integral involving the Green function and boundary data. Specifically, by applying Green's second identity to the region $\lim\limits_{\zeta_3\to\infty} V(\zeta_3) - V(\zeta_2)$, using the Helmholtz equations satisfied by the vector potential and the retarded Green function, and making use of the assumed Sommerfeld radiation conditions, we find:

$$\boldsymbol{a}(\boldsymbol{\zeta};\omega) = \int\limits_{\partial V(\zeta_2)} d\sigma'\left[\boldsymbol{a}(\boldsymbol{\zeta}';\omega)(\hat{\boldsymbol{n}}'\cdot\boldsymbol{\partial}')G^{\rm ret}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) - G^{\rm ret}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)(\hat{\boldsymbol{n}}'\cdot\boldsymbol{\partial}')\boldsymbol{a}(\boldsymbol{\zeta}';\omega)\right], \tag{4.165}$$

which is the vector form of the well-known Kirchoff diffraction integral, and may also be regarded as a quantitative form of Huygens' principle. Here the unit vector $\hat{\boldsymbol{n}}' = \hat{\boldsymbol{n}}(\boldsymbol{\zeta}')$ denotes an outward normal to the boundary surface at any point $\boldsymbol{\zeta}' \in \partial V(\zeta_2)$. At arbitrary points in space, within or beyond the sources, this convolution (4.165) is identical to the vector potential appearing in the spherical-wave expansion (4.145), and to it applies the same caveat about near-field errors for finite $\|\boldsymbol{\zeta}\|$. In particular, for any $\boldsymbol{\zeta} \neq \boldsymbol{0}$, including within the effective support of the solenoidal source, it is a solution of the source-free Helmholtz equation, equal to the point-multipole extrapolation discussed above.

By piecing together different spherical wave solutions to produce just the right discontinuity at $\boldsymbol{\zeta} = \boldsymbol{\zeta}'$, it is straightforward to establish the well-known Bessel decomposition of the retarded Green function:

$$G^{\rm ret}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) = i\omega\sum_{\ell=0}^{\infty} j_\ell(k\zeta_<)h_\ell^{(1)}(k\zeta_>)\sum_{m=-\ell}^{\ell} Y_{\ell m}(\hat{\boldsymbol{\zeta}})Y_{\ell m}(\hat{\boldsymbol{\zeta}}')^* \tag{4.166}$$

where

$$\zeta_< = \min[\zeta, \zeta'], \tag{4.167a}$$

$$\zeta_> = \max[\zeta, \zeta']. \tag{4.167b}$$

From linearity, we can see that the difference between the retarded and advanced Green functions,

$$D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) \equiv G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) - G^{\text{adv}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) = -\frac{\sin\left(k\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|\right)}{2\pi i\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|} = \frac{i}{2\pi}\,\text{sinc}\left(k\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|\right), \tag{4.168}$$

must be a solution (with respect to either spatial coordinate) to the source-free scalar Helmholtz equation everywhere in space. Substituting in the decomposition (4.166) for the Green function, this homogeneous solution $D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)$ may be written as:

$$D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) = 2i\omega \sum_{\ell,m} j_\ell(k\zeta) j_\ell(k\zeta')^* Y_{\ell m}(\hat{\boldsymbol{\zeta}}) Y_{\ell m}(\hat{\boldsymbol{\zeta}}')^*, \tag{4.169}$$

where we have used the facts that $\sum_{m=-\ell}^{\ell} Y_{\ell m}(\hat{\boldsymbol{\zeta}}) Y_{\ell m}(\hat{\boldsymbol{\zeta}}')^*$ and $j_\ell(k\zeta)$ are both real. This function will be referred to either as the fundamental (source-free) solution or as the homogeneous Green function.

Now, it turns out that, just as a general solution to the inhomogeneous equation with outgoing Sommerfeld radiation boundary conditions may be written as a convolution between the retarded Green function and the solenoidal portion of the current density, any arbitrary solenoidal solution to the homogeneous problem may be written as a convolution between the fundamental solution $D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)$ and the solenoidal component $\boldsymbol{g}_\perp(\boldsymbol{\zeta}';\omega)$, of some effective source/sink term $\boldsymbol{g}(\boldsymbol{\zeta}';\omega)$ whose support can be confined to any finite-radius ball $B(\zeta_0)$ where $\zeta_0 > 0$:

$$\boldsymbol{\chi}(\boldsymbol{\zeta};\omega) = \int d^3\boldsymbol{\zeta}'\, D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{g}_\perp(\boldsymbol{\zeta}';\omega) = \int\limits_{V(\zeta_0)} d^3\boldsymbol{\zeta}'\, D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{g}_\perp(\boldsymbol{\zeta}';\omega) \tag{4.170}$$

To see this, first note that if we choose

$$\boldsymbol{g}(\boldsymbol{\zeta};\omega) = \boldsymbol{g}_\perp(\boldsymbol{\zeta};\omega) \propto \Theta(\zeta_0 - \zeta) j_\ell(k\zeta) \boldsymbol{X}_{\ell m}(\theta,\varphi), \tag{4.171}$$

(which is clearly divergenceless by the chain rule and the transverse nature of $\boldsymbol{X}_{\ell m}(\hat{\boldsymbol{\zeta}})$), then, by using the explicit decomposition (4.169) for $D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)$ and the orthogonality properties of the spherical harmonics as well as their behavior under the action of the orbital angular momentum operator $\boldsymbol{L}$, we find

$$\boldsymbol{\chi}(\boldsymbol{\zeta};\omega) \propto \left[\int_0^{\zeta_0} d\zeta' \,\left|\zeta' j_\ell(k\zeta')\right|^2\right] j_\ell(k\zeta)\boldsymbol{X}_{\ell m}(\theta,\varphi), \tag{4.172}$$

and thus any term of the magnetic multipole form in (4.155) can be generated in this manner. Using standard vector identities, the symmetry of $D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)$ in its spatial arguments, and integration by parts relying on the rapid fall-off of $\boldsymbol{g}_\perp$, we also have

$$\begin{aligned}\boldsymbol{\partial}\times\int d^3\boldsymbol{\zeta}'\, D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\,\boldsymbol{g}(\boldsymbol{\zeta}';\omega) &= \int d^3\boldsymbol{\zeta}'\,\boldsymbol{\partial}\times\big[D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\,\boldsymbol{g}_\perp(\boldsymbol{\zeta}';\omega)\big]\\ &= \int d^3\boldsymbol{\zeta}'\, D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\,\boldsymbol{\partial}'\times\boldsymbol{g}_\perp(\boldsymbol{\zeta}';\omega),\end{aligned} \tag{4.173}$$

so any required term of the electric multipole form can also be so generated. By linearity, an arbitrary homogeneous solution may then be constructed as a convolution over the fundamental homogeneous solution. We will see explicitly below that the effective source/sink $\boldsymbol{g}(\boldsymbol{\zeta};\omega)$ appearing in the convolutional representation may be taken to be identical to that determining the multipole coefficients in the spherical wave expansion.

4.4.3 Hilbert Space Results

In the non-paraxial case, the wave-equation is no longer equivalent to a time-dependent Schrödinger equation, although the source-free Helmholtz equation is analogous to a time-independent, free-particle Schrödinger equation, albeit typically with boundary conditions leading to non-normalizable, scattering-state solutions. However, when convenient, we can still recast our non-paraxial mathematical results into a Hilbert-space-like formalism. Without a distinguished optic axis, we extend the spatial dependence of spinors and observables from two dimensions to three dimensions, each coordinate now treated on an equal footing, and we allow for more general polarization bases, using some given family (labeled here by n, standing for some set of integer indices or "quantum numbers") of smooth, possibly complex-valued vector fields $\hat{\boldsymbol{\epsilon}}^n_s(\boldsymbol{\zeta};\omega)$ (still indexed, say by $s=-1,0,+1$) which constitute a local orthonormal triad at every spatial position $\boldsymbol{\zeta}$:

$$\boldsymbol{\epsilon}^n_s(\boldsymbol{\zeta};\omega)^*\cdot\boldsymbol{\epsilon}^n_{s'}(\boldsymbol{\zeta};\omega)=\delta_{s\,s'}. \tag{4.174}$$

For example, we could use, as convenient, a fixed Cartesian basis $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$, $\hat{\boldsymbol{z}}$; the spherical-coordinate basis $\hat{\boldsymbol{\zeta}}$, $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\varphi}}$; the helicity basis $\hat{\boldsymbol{e}}_+$, $\hat{\boldsymbol{e}}_-$, $\hat{\boldsymbol{z}}$ (also known as the spherical basis because of its role in defining spherical tensors, not to be confused with the above basis vectors for spherical coordinates); normalized vector spherical harmonic basis vector field for some $\ell\geq 1$ and m, i.e., $\hat{\boldsymbol{X}}_{\ell m}(\hat{\boldsymbol{\zeta}})$, $\hat{\boldsymbol{N}}_{\ell m}(\hat{\boldsymbol{\zeta}})$, $\hat{\boldsymbol{Z}}_{\ell m}(\hat{\boldsymbol{\zeta}})$; or a natural radiation basis field derived from some reference vector potential $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$, namely $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{a}\perp}$, $\hat{\boldsymbol{s}}_{\boldsymbol{a}}$, $\hat{\boldsymbol{b}}_{\boldsymbol{a}}$.

Much of the "quantum"-like formalism developed in the paraxial case can be generalized in the obvious way to a three-dimensional Hilbert space $\mathcal{H}_{3\mathrm{D}}$ consisting of square-integrable

functions $\boldsymbol{g} : \mathbb{R}^3 \to \mathbb{C}^3$, and will not be reproduced here, except to note that we will now denote the (possibly generalized, i.e., non-normalizable) eigenket corresponding to the vector field $\boldsymbol{g}(\boldsymbol{\zeta};\omega)$ by $|\boldsymbol{g}(\omega)\rangle$, and the generalized 3D position/polarization eigenkets by $|\boldsymbol{\zeta};\hat{\boldsymbol{\epsilon}}_s^n(\boldsymbol{\zeta};\omega)\rangle$, such that

$$\langle \boldsymbol{\zeta};\hat{\boldsymbol{\epsilon}}_s^n(\boldsymbol{\zeta};\omega)\,|\boldsymbol{g}(\omega)\rangle = \hat{\boldsymbol{\epsilon}}_s^n(\boldsymbol{\zeta};\omega)^* \cdot \boldsymbol{g}(\boldsymbol{\zeta};\omega), \tag{4.175}$$

where the Hilbert-space inner product now involves integration over all spatial variables:

$$\langle \boldsymbol{g}\,|\,\boldsymbol{g}'\rangle = \int d^3\boldsymbol{\zeta}\; \boldsymbol{g}(\boldsymbol{\zeta})^* \cdot \boldsymbol{g}'(\boldsymbol{\zeta}). \tag{4.176}$$

Now, consider the Hilbert-space operator $K = K(\omega)$ defined via its spatial/polarization kernel:

$$K(\boldsymbol{\zeta},s;\boldsymbol{\zeta}',s';\omega) \equiv \langle \boldsymbol{\zeta};\hat{\boldsymbol{\epsilon}}_s^n(\boldsymbol{\zeta};\omega)\,\big|K(\omega)\big|\,\boldsymbol{\zeta}';\hat{\boldsymbol{\epsilon}}_{s'}^n(\boldsymbol{\zeta};\omega)\rangle \equiv -\tfrac{1}{2}i\omega\,\delta_{s\,s'}\,D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega), \tag{4.177}$$

which is given explicitly by

$$K(\boldsymbol{\zeta},0;\boldsymbol{\zeta}',0;\omega) = \frac{\omega}{4\pi}\,\mathrm{sinc}\left(k\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|\right) = \omega^2 \sum_{\ell,m} j_\ell(k\zeta) j_\ell(k\zeta')^* Y_{\ell\,m}(\hat{\boldsymbol{\zeta}}) Y_{\ell\,m}(\hat{\boldsymbol{\zeta}}')^*. \tag{4.178}$$

The kernel is seen to be real-valued and symmetric in all spin/spatial arguments, so the corresponding operator $K = K^\dagger$ is Hermitian. Because the kernel acts as the identity on polarization degrees of freedom, the operator is independent of spin observables, and may be formally written as a function only of spatial observables. In fact, because the free-space, fundamental source-free solution is translationally and rotationally invariant, depending on the spatial coordinates only through $\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|$, K must be a function solely of the momentum magnitude $\|\boldsymbol{P}_{3\mathrm{D}}\| = \sqrt{-\partial^2}$. Crucially, K is also positive semidefinite, in the sense that

$$\langle \boldsymbol{g}(\omega)\,\big|K(\omega)\big|\,\boldsymbol{g}'(\omega)\rangle = \int d^3\boldsymbol{\zeta} \int d^3\boldsymbol{\zeta}'\; K(\boldsymbol{\zeta},0;\boldsymbol{\zeta}',0;\omega)\,\boldsymbol{g}(\boldsymbol{\zeta};\omega)^* \cdot \boldsymbol{g}(\boldsymbol{\zeta}';\omega) \geq 0, \tag{4.179}$$

for any complex vector field $\boldsymbol{g}(\boldsymbol{\zeta};\omega)$ which is sufficiently well-behaved for the integral to exist. This can be seen most easily from the diagonal expansion in spherical waves:

$$\langle \boldsymbol{g}(\omega)\,\big|K(\omega)\big|\,\boldsymbol{g}'(\omega)\rangle = \omega^2 \sum_{\ell,m} \sum_s \left| \int d^3\boldsymbol{\zeta}\; \hat{\boldsymbol{\epsilon}}_s^n(\boldsymbol{\zeta})^* \cdot \boldsymbol{g}(\boldsymbol{\zeta};\omega) j_\ell(k\zeta)^* Y_{\ell\,m}(\hat{\boldsymbol{\zeta}})^* \right|^2 \geq 0. \tag{4.180}$$

It is also a direct consequence of the well-known Bochner theorem, which says that a continuous function in real-space is nonnegative definite as a spatial kernel if and only if its Fourier transform is proportional to a positive measure over reciprocal space (and positive definite if this measure is everywhere non-zero). Transforming to the scaled momentum

(i.e., wavevector, or reciprocal-space) representation where the operator K is diagonal, we find:

$$K(\omega) = \pi\delta(\boldsymbol{P}_{3\mathrm{D}}\cdot\boldsymbol{P}_{3\mathrm{D}} - \omega^2), \tag{4.181}$$

so that

$$\big\langle \boldsymbol{g}(\omega)\,\big|K(\omega)\big|\,\boldsymbol{g}'(\omega)\big\rangle = \pi\int d^3\boldsymbol{k}\;|\boldsymbol{g}(\boldsymbol{k};\omega)|^2\,\delta(\|\boldsymbol{k}\|^2-\omega^2) \geq 0, \tag{4.182}$$

where $\boldsymbol{k}$ is the scaled wavevector conjugate to $\boldsymbol{\zeta}$, and $\boldsymbol{g}(\boldsymbol{k};\omega)$ is the (scaled) spatial Fourier transform of $\boldsymbol{g}(\boldsymbol{\zeta};\omega)$. Note that $K(\omega)$ acts to restrict the spectral content of any vector field $\boldsymbol{g}(\boldsymbol{\zeta};\omega)$ in its domain to the manifold of wavevectors satisfying the vacuum dispersion relation $\omega^2 = \|\boldsymbol{k}\|^2$. However, $K(\omega)$ is not idempotent, so cannot be considered a true projection.

That $K(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)$ is only positive semidefinite and not strictly positive definite is a consequence of the well-known existence of non-radiating sources for the Helmholtz problem, which invariably complicate uniqueness results and inverse-scattering calculations. For example, suppose specifically we take

$$\boldsymbol{g}_0(\boldsymbol{\zeta};\omega) = \boldsymbol{g}_{0\,\perp}(\boldsymbol{\zeta};\omega) \propto \Theta(\zeta_0-\zeta)g_\ell(k\zeta)\boldsymbol{X}_{\ell m}(\theta,\varphi), \tag{4.183}$$

for any allowed ℓ and m, some $\zeta_0 > 0$, and any well-behaved, non-trivial scalar function $g_\ell(k\zeta)$ which is chosen to satisfy

$$\int_0^{\zeta_0} d\zeta\,\zeta^2 g_\ell(k\zeta)j_\ell(k\zeta) = 0, \tag{4.184a}$$

$$\int_0^{\zeta_0} d\zeta\,\zeta^2 g_\ell(k\zeta)g_\ell(k\zeta) > 0, \tag{4.184b}$$

Then using the spherical wave expansions, it immediately follows for this source that, at any position $\boldsymbol{\zeta}$,

$$\boldsymbol{\chi}_0(\boldsymbol{\zeta};\omega) = \int d^3\boldsymbol{\zeta}'\,D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{g}_0(\boldsymbol{\zeta}';\omega) = \boldsymbol{0}; \tag{4.185}$$

and also, anywhere in the extra-source region, i.e., for $\zeta > \zeta_1$,

$$\boldsymbol{a}_0^{\mathrm{out}}(\boldsymbol{\zeta};\omega) = \int d^3\boldsymbol{\zeta}'\,G^{\mathrm{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{g}_0(\boldsymbol{\zeta}';\omega) = \boldsymbol{0}, \tag{4.186a}$$

$$\boldsymbol{a}_0^{\mathrm{in}}(\boldsymbol{\zeta};\omega) = \int d^3\boldsymbol{\zeta}'\,G^{\mathrm{adv}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{g}_0(\boldsymbol{\zeta}';\omega) = \boldsymbol{0}, \tag{4.186b}$$

and indeed no radiation is produced by $\boldsymbol{g}_0$ acting either as an actual source, time-reversed source (sink), or effective (homogeneous) source/sink, although some near fields must be produced by any non-zero $\boldsymbol{g}_0$ acting as a source for G^{ret} or sink for G^{adv}. Any such vector

field like $\boldsymbol{g}_0(\boldsymbol{\zeta};\omega)$ producing no far-zone fields corresponds to a ket $|\boldsymbol{g}_0(\omega)\rangle$ in the nullspace of the operator K; i.e., $K(\omega)\,|\boldsymbol{g}_0(\omega)\rangle = 0$.

The nullspace of the Hermitian operator $K(\omega)$ is therefore infinite-dimensional (and in fact non-denumerable). Modulo some (admittedly thorny) normalization issues due to the unbounded nature of $K(\omega)$, the basis of homogeneous spherical waves solutions (4.155) are the orthogonal eigenfunctions of K in the position basis corresponding to non-zero eigenvalue. Although it would involve more mathematical detail than justified in this presentation, the properties and behavior of K can be analyzed more rigorously within the theory of Reproducing Kernel Hilbert Spaces (RKHS), where the spherical wave expansion of the kernel emerges as a consequence of Mercer's theorem [127]. In a similar manner, we may also define Green operators $G^{\text{ret}}(\omega)$ and $G^{\text{adv}}(\omega)$ from their spatial representations; these operators are complex-symmetric, but neither Hermitian nor positive semidefinite, and both act as right (but not left) inverses for the Helmholtz wave operator $L(\omega) = \partial^2 + \omega^2$.

However, the problems encountered with normalization cannot be entirely ignored. While this Hilbert space $\mathcal{H}_{\text{3D}}$, equipped with the Euclidean/L_2 inner product, is the natural setting for the source-terms such as $\boldsymbol{j}(\boldsymbol{\zeta};\omega)$, $\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega)$, $\boldsymbol{g}(\boldsymbol{\zeta};\omega)$, and $\boldsymbol{g}_\perp(\boldsymbol{\zeta};\omega)$, it is not always a comfortable home for the vector potentials $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$ or $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega)$ or the corresponding fields, since the presence of far-zone radiation fields, falling off asymptotically like $O(1/\|\boldsymbol{\zeta}\|)$, implies that the vector potentials and fields are not normalizable with respect to this inner product, so do not strictly belong to the Hilbert space $\mathcal{H}_{\text{3D}}$. However, they may be approximated arbitrarily closely (point-wise) by vector fields in the Hilbert-space, and inner products between radiation and sources, e.g., $\langle \boldsymbol{j}(\omega)\,|\,\boldsymbol{a}(\omega)\rangle$ will exist for allowed current densities, which must be localized.

The lack of nomalizability may be traced to the fact that radiation fields can contain, in principle, an infinite amount of energy when integrated over all space, while transmitting only a finite amount of power. If necessary, we could use the familiar regularizing device of introducing a finite "quantization" volume $V(\zeta_Q)$ together with periodic or conducting-wall boundary conditions, and then take the limit $\zeta_Q \to \infty$ at the very end of the calculation. For the currently envisioned application, this is not terribly convenient, as we are interested primarily in the far-field from the start. Alternatively, we can choose a different inner product, which effectively normalizes the radiation fields based on power (or power spectral density) rather than energy. We can define a Hilbert space $\mathcal{H}_{\text{out}}$ spanned by the basis consisting of outgoing solenoidal vector spherical waves, with an inner product defined as a

Euclidean dot product on the multipole expansion coefficients:

$$\big(\boldsymbol{a}(\omega)\,,\,\boldsymbol{a}'(\omega)\big)_{\text{out}} = \sum_{\ell\,m} \big[a^{\text{E}*}_{\ell m}(\omega) a'^{\text{E}}_{\ell m}(\omega) + a^{\text{M}*}_{\ell m}(\omega) a'^{\text{M}}_{\ell m}(\omega)\big]. \tag{4.187}$$

The exact relationship between this inner product and the scaled power spectral density will be seen shortly, but it should be apparent that this quantity may be expressed solely in in terms of far-field data, rather than on the fields throughout all space, and that physical radiation fields, with finite power, will be normalizable in the sense $(\boldsymbol{a}(\omega)\,,\,\boldsymbol{a}(\omega)) < \infty$. An isomorphic Hilbert space $\mathcal{H}_{\text{in}} \equiv \mathcal{H}^*_{\text{out}} \cong \mathcal{H}_{\text{out}}$ may be defined in an exactly analogous manner for incoming-wave solutions, such as $\boldsymbol{a}(\boldsymbol{\zeta};\omega)^*$. Complex conjugation (or equivalently time reversal) provides one natural isometry between them, or instead one may use the mapping which simply swaps the spherical Hankel functions $h^{(1)}(kr)$ and $h^{(2)}(kr)$ in the spherical wave expansions without conjugating or otherwise changing the multipole expansion coefficients or the vector spherical harmonics. Strictly outside the support of the sources (if any), any solenoidal solution $\boldsymbol{\psi}$ to the vector Helmholtz equation (homogeneous or inhomogeneous), of finite power but with otherwise arbitrary boundary conditions, may be expressed as the superposition of incoming and outgoing spherical waves, so it necessarily lies in the Hilbert space $\mathcal{H}_{\text{out}} \oplus \mathcal{H}_{\text{in}}$. Thus, we can uniquely decompose (for $\zeta > \zeta_2$) any such solution $\boldsymbol{\psi}(\boldsymbol{\zeta};\omega)$ as

$$\boldsymbol{\psi}(\boldsymbol{\zeta};\omega) = \boldsymbol{\psi}^{\text{out}}(\boldsymbol{\zeta};\omega) + \boldsymbol{\psi}^{\text{in}}(\boldsymbol{\zeta};\omega), \tag{4.188}$$

where $\boldsymbol{\psi}^{\text{out}}$ and $\boldsymbol{\psi}^{\text{in}}$ are, respectively, the purely outgoing and purely incoming components of the vector field, in the sense that they satisfy the appropriate Sommerfeld boundary conditions as $\zeta \to \infty$. Even within the support of the sources (for $\zeta \leq \zeta_1$), vector potentials may be uniquely decomposed in this fashion, with the understanding that at finite positions, the meaning of incoming versus outgoing fields is determined by the point-multipole convention described above, and therefore in the vicinity of the sources the outgoing component so calculated is an extrapolation, and may differ by rapidly-decaying, non-radiative fields from the actual fields produced by an actual source with some finite extent.

The inherited inner product in this space is given by

$$\big(\boldsymbol{\psi}(\omega)\,,\,\boldsymbol{\psi}'(\omega)\big) = \big(\boldsymbol{\psi}^{\text{out}}(\omega)\,,\,\boldsymbol{\psi}'^{\text{out}}(\omega)\big)_{\text{out}} + \big(\boldsymbol{\psi}^{\text{in}}(\omega)\,,\,\boldsymbol{\psi}'^{\text{in}}(\omega)\big)_{\text{in}}\,. \tag{4.189}$$

Our approximating "trial" vector potentials $\boldsymbol{\chi}$ will lie in the proper Hilbert subspace $\mathcal{H}_{\text{hom}} \subsetneq \mathcal{H}_{\text{out}} \oplus \mathcal{H}_{\text{in}}$, consisting of source-free solutions to the wave equation which are regular everywhere in space. This space of homogeneous solutions is also isomorphic to either the outgoing or incoming wave spaces: $\mathcal{H}_{\text{hom}} \cong \mathcal{H}_{\text{out}} \cong \mathcal{H}_{\text{in}}$, as can be seen by imagining the mapping that replaces, in the spherical wave expansions, each real-valued spherical

Bessel function $j_\ell(kr)$ with the corresponding complex spherical Hankel function $h^{(1)}(kr)$ or $h^{(2)}(kr)$, respectively, without altering the angular dependence or the expansion coefficients.

It is now not difficult to better understand the origin of the non-radiating sources. Homogeneous solutions to the wave equation must consist of superpositions of purely radiative fields, since there are no longer any localized sources around to define the near or intermediate zones. (Physically, all electromagnetic fields must have their origin in charges somewhere, but in the source-free idealization, we in effect imagine that these sources had been located infinitely far away in space and/or turned off infinitely far in the past, so only radiation fields are observed.) These source-free radiation fields, however must carefully combine incoming and outgoing waves (defined with respect to the chosen origin) in order to satisfy the wave equation everywhere in time and space without any sources or singularities.

Now, by definition, outgoing, purely radiative fields have lost information about the exact details of the current configurations that produced them or the resulting near fields associated with these currents. Different sources, say $\boldsymbol{j}_1(\boldsymbol{\zeta};\omega)$ and $\boldsymbol{j}_2(\boldsymbol{\zeta};\omega)$, can emit exactly the same outgoing radiation fields in the far-field limit (i.e., identical spectral radiant intensities, or average power radiated into a given solid angle and frequency bandwidth), differing only in their near-zone and intermediate zone fields. Specifically, suppose

$$\int d^3\boldsymbol{\zeta}'\, G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_{\perp 1}(\boldsymbol{\zeta}';\omega) = \int d^3\boldsymbol{\zeta}'\, G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_{\perp 2}(\boldsymbol{\zeta}';\omega) + O\big(\tfrac{1}{\|\boldsymbol{\zeta}\|^2}\big), \tag{4.190}$$

in the precise sense that all multipole moments in the "extra-source" region, (or, what amounts to the same thing, for the equivalent singular solenoidal source concentrated at the origin) are equal:

$$\int d\zeta'\, j_\ell(k\zeta) Y_{\ell m}(\hat{\boldsymbol{\zeta}}')^* \boldsymbol{j}_{\perp 1}(\boldsymbol{\zeta}';\omega) = \int d\zeta'\, j_\ell(k\zeta) Y_{\ell m}(\hat{\boldsymbol{\zeta}}')^* \boldsymbol{j}_{\perp 1}(\boldsymbol{\zeta}';\omega). \tag{4.191}$$

Then by using the expansion (4.166) for the Green function, it is straightforward to verify that these same sources also lead to (or rather, absorb) the same incoming radiation fields

$$\int d^3\boldsymbol{\zeta}'\, G^{\text{adv}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_{\perp 1}(\boldsymbol{\zeta}';\omega) = \int d^3\boldsymbol{\zeta}'\, G^{\text{adv}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_{\perp 2}(\boldsymbol{\zeta}';\omega) + O\big(\tfrac{1}{\|\boldsymbol{\zeta}\|^2}\big). \tag{4.192}$$

When such current densities are used as effective sources for homogeneous solutions, the non-radiative local fields produced by the advanced and retarded halves of D will cancel, leading to the identical source-free, everywhere radiative fields:

$$\int d^3\boldsymbol{\zeta}'\, D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_{\perp 1}(\boldsymbol{\zeta}';\omega) = \int d^3\boldsymbol{\zeta}'\, D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_{\perp 2}(\boldsymbol{\zeta}';\omega). \tag{4.193}$$

By linearity, the difference between any two such sources, $\boldsymbol{g}_0 = \boldsymbol{j}_1 - \boldsymbol{j}_2$, therefore emits no radiation as a real source (i.e., under the action of G^{ret}), absorbs no radiation as a sink (i.e., under the action of G^{adv}), and produces no net source-free fields as a sink:

$$\int d^3\boldsymbol{\zeta}'\, G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{g}_{\perp 0}(\boldsymbol{\zeta}';\omega) = 0 + O\big(\tfrac{1}{\|\boldsymbol{\zeta}\|^2}\big); \tag{4.194a}$$

$$\int d^3\boldsymbol{\zeta}'\, G^{\text{adv}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{g}_{\perp 0}(\boldsymbol{\zeta}';\omega) = 0 + O\big(\tfrac{1}{\|\boldsymbol{\zeta}\|^2}\big); \tag{4.194b}$$

$$\int d^3\boldsymbol{\zeta}'\, D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{g}_{\perp 0}(\boldsymbol{\zeta}';\omega) = 0. \tag{4.194c}$$

Two implications follow immediately: to any effective source $\boldsymbol{g}_\perp$ for a given source-free field $\boldsymbol{\chi}$ we may add a multiple of any non-radiating source $\boldsymbol{g}_{0\perp}$, so such effective sources are never unique; and, conversely, near-field information cannot be unambiguously extracted from the source-free extension of these fields, without knowing the actual sources. That is, any source-free solution has a unique direct-sum decomposition into what we have defined as the incoming and outgoing-wave Hilbert spaces, but in each of these, the fields by construction exhibit a singularity at a single point (chosen to be the origin), and contain no information about the actual near fields in a neighborhood of the origin.

4.4.4 Energy Balance and Poynting Flux

Using the Helmholtz equation, gauge constraint, and various vector identities, one can derive a generalized impedance relation, or complexified power-balance equation, for the fields in the frequency domain over any closed spatial region R, and valid for any consistent boundary conditions imposed on the radiation at spatial infinity:

$$\int_R d^3\boldsymbol{\zeta}\ [\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega)^*\cdot\boldsymbol{\varepsilon}_{\boldsymbol{a}\perp}(\boldsymbol{\zeta};\omega)] + i\omega\int_R d^3\boldsymbol{\zeta}\ \big[\|\boldsymbol{\varepsilon}_{\boldsymbol{a}\perp}(\boldsymbol{\zeta};\omega)\|^2 - \|\boldsymbol{b}_{\boldsymbol{a}}(\boldsymbol{\zeta};\omega)\|^2\big] + \int_{\partial R} d^2\sigma\ [\hat{\boldsymbol{n}}\cdot\boldsymbol{s}_{\boldsymbol{a}}(\boldsymbol{\zeta};\omega)] = 0, \tag{4.195}$$

where $\hat{\boldsymbol{n}} = \hat{\boldsymbol{n}}(\boldsymbol{\zeta})$ is a outward normal unit vector on the boundary surface at $\boldsymbol{\zeta} \in \partial R$. This relation is quite similar to the usual conservation law derived in textbooks from the time-harmonic formulation of Maxwell's equations [89, 92], except that it involves only the solenoidal components of the sources and the fields – the longitudinal component $\boldsymbol{j}_\parallel$ of the current density and the scalar potential ϕ and its derivatives do not appear.

In the most general, inhomogeneous case, this relation implies:

$$-\operatorname{Re}\int_R d^3\boldsymbol{\zeta}\ [\boldsymbol{j}_\perp^*\cdot\boldsymbol{\varepsilon}_{\boldsymbol{a}\perp}] = \operatorname{Re}\int_{\partial R} d^2\sigma\ [\hat{\boldsymbol{n}}\cdot\boldsymbol{s}_{\boldsymbol{a}}]\,, \tag{4.196}$$

which is an expression of time-averaged energy conservation, relating the (negative) rate of mechanical work done on the solenoidal sources within R to the rate of energy escape in the form of radiation through the boundary ∂R. If we let $R = V(\zeta_2)$ for any $\zeta_2 > \zeta_1$, we can replace $\boldsymbol{j}_\perp$ with the full current density $\boldsymbol{j}$ on the right-hand-side, and this becomes

$$-\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{a}\perp};\boldsymbol{j}](\omega) \equiv -\operatorname{Re}\int d^3\boldsymbol{\zeta}\ \ [\boldsymbol{j}^*\cdot\boldsymbol{\varepsilon}_{\boldsymbol{a}\perp}] = \operatorname{Re}\int\limits_{\partial V(\zeta_2)} d^2\sigma\ \ [\hat{\boldsymbol{n}}\cdot\boldsymbol{s}_{\boldsymbol{a}}] \equiv \mathscr{P}'_{\text{EM}}[\boldsymbol{a}](\zeta_2;\omega). \tag{4.197}$$

The radiated power spectral density is independent of radius ζ_2 as long as $\zeta_2 > \zeta_1$, as assumed, so we may take $\zeta_2 \to \infty$ on the left-hand-side whenever convenient. Using Green's Second Identity and other vector identities, the spectral density of the radiative flux can also be written in the following useful forms within the Coulomb gauge:

$$\mathscr{P}'_{\text{EM}}[\boldsymbol{a}](\zeta_2;\omega) = \frac{i\omega}{2}\int\limits_{V(\zeta_2)} d^3\boldsymbol{\zeta}\ \left[\boldsymbol{a}\cdot\nabla^2\boldsymbol{a}^* - \boldsymbol{a}^*\cdot\nabla^2\boldsymbol{a}\right] = \frac{i\omega}{2}\int\limits_{\partial V(\zeta_2)} d^2\sigma\ \left[\boldsymbol{a}\cdot(\hat{\boldsymbol{n}}\cdot\boldsymbol{\partial})\boldsymbol{a}^* - \boldsymbol{a}^*\cdot(\hat{\boldsymbol{n}}\cdot\boldsymbol{\partial})\boldsymbol{a}\right]. \tag{4.198}$$

Similarly, if $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$ is the exact solenoidal radiation corresponding to $\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega)$, and $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega)$ is any arbitrary solenoidal solution to the source-free Helmholtz equation, then for $\zeta_2 > \zeta_1$,

$$\begin{aligned}-\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp};\boldsymbol{j}](\omega) &= \frac{i\omega}{2}\int\limits_{V(\zeta_2)} d^3\boldsymbol{\zeta}\ \left[\boldsymbol{\chi}\cdot\partial^2\boldsymbol{a}^* - \boldsymbol{a}^*\cdot\partial^2\boldsymbol{\chi} + \boldsymbol{a}\cdot\partial^2\boldsymbol{\chi}^* - \boldsymbol{\chi}^*\cdot\partial^2\boldsymbol{a}\right] \\ &= \frac{i\omega}{2}\int\limits_{\partial V(\zeta_2)} d^2\sigma\ \left[\boldsymbol{\chi}\cdot(\hat{\boldsymbol{n}}\cdot\boldsymbol{\partial})\boldsymbol{a}^* - \boldsymbol{a}^*\cdot(\hat{\boldsymbol{n}}\cdot\boldsymbol{\partial})\boldsymbol{\chi} + \boldsymbol{a}\cdot(\hat{\boldsymbol{n}}\cdot\boldsymbol{\partial})\boldsymbol{\chi}^* - \boldsymbol{\chi}^*\cdot(\hat{\boldsymbol{n}}\cdot\boldsymbol{\partial})\boldsymbol{a}\right].\end{aligned} \tag{4.199}$$

For any such solution $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega)$ to the homogeneous Helmholtz equation (i.e., where $\partial^2\boldsymbol{\chi} = -\omega^2\boldsymbol{\chi}$), we have, from (4.195) or (4.199):

$$\operatorname{Re}\int\limits_{\partial R} d^2\sigma\ \ [\hat{\boldsymbol{n}}\cdot\boldsymbol{s}_{\boldsymbol{\chi}}] = \operatorname{Re}\int\limits_{\partial R} d^3\boldsymbol{\zeta}\ \ [\boldsymbol{\partial}\cdot\boldsymbol{s}_{\boldsymbol{\chi}}] = \frac{i\omega}{2}\int\limits_{R} d^3\boldsymbol{\zeta}\ \ \left[\boldsymbol{\chi}\cdot\partial^2\boldsymbol{\chi}^* - \boldsymbol{\chi}^*\cdot\partial^2\boldsymbol{\chi}\right] = 0, \tag{4.200}$$

i.e., the real Poynting flux vanishes when integrated over any closed surface, reflecting the fact that, as is intuitively evident from the convolution representation (4.170) or from a plane-wave decomposition, any vector potential which satisfies the homogeneous wave equation everywhere must, in fact, result in an equal amount of averaged radiative flux passing into and out from any closed surface. In addition, a plane wave or spherical wave expansion will verify that $\int\limits_{\mathbb{R}^3} d^3\boldsymbol{\zeta}\ \left[\|\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\boldsymbol{\zeta};\omega)\|^2 - \|\boldsymbol{b}_{\boldsymbol{\chi}}(\boldsymbol{\zeta};\omega)\|^2\right] = 0$ for such source-free fields,

so if we consider only the *truly* radiative component of the Poynting flux $\boldsymbol{s}_{\boldsymbol{\chi}}$ by going to the far-field limit, we find

$$\lim_{\zeta_0\to\infty} \mathrm{Im} \int\limits_{\partial V(\zeta_0)} d^2\sigma \ [\hat{\boldsymbol{n}}\cdot\boldsymbol{s}_{\boldsymbol{\chi}}] = 0, \tag{4.201}$$

so neither is there any net reactive storage of energy in the radiative components of the free fields, as would be expected.

Moreover, as we saw above, we can locally decompose arbitrary source-free fields into what represent, or at least eventually become (as $\zeta \to \infty$) , separately outwardly-radiating and inwardly-radiating waves:

$$\boldsymbol{\chi}(\boldsymbol{\zeta};\omega) = \boldsymbol{\chi}^{\mathrm{out}}(\boldsymbol{\zeta};\omega) + \boldsymbol{\chi}^{\mathrm{in}}(\boldsymbol{\zeta};\omega), \tag{4.202}$$

where for all $\boldsymbol{\zeta} \in \partial V(\zeta_2)$, $\mathrm{Re}\left[\hat{\boldsymbol{n}}(\boldsymbol{\zeta})\cdot\boldsymbol{s}^{\mathrm{out}}_{\boldsymbol{\chi}}\right] \geq 0$ and $\mathrm{Re}\left[\hat{\boldsymbol{n}}(\boldsymbol{\zeta})\cdot\boldsymbol{s}^{\mathrm{in}}_{\boldsymbol{\chi}}\right] \leq 0$ as $\zeta_2 \to \infty$. This decomposition can be effected in the spherical-wave basis, but is more easily accomplished simply by separating the fundamental representation into its retarded (outgoing) and advanced (incoming) parts, if an effective source/sink is known:

$$\boldsymbol{\chi}^{\mathrm{out}}(\boldsymbol{\zeta};\omega) = \int d^3\boldsymbol{\zeta}'\, G^{\mathrm{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{g}_\perp(\boldsymbol{\zeta}';\omega); \tag{4.203}$$

and

$$\boldsymbol{\chi}^{\mathrm{in}}(\boldsymbol{\zeta};\omega) = -\int d^3\boldsymbol{\zeta}'\, G^{\mathrm{adv}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{g}_\perp(\boldsymbol{\zeta}';\omega). \tag{4.204}$$

Note that these vector potentials separately satisfy the homogeneous equation only for $\zeta > \zeta_0$, but satisfy the inhomogeneous equation everywhere for the appropriate effective source terms and specified boundary conditions. Although $\boldsymbol{\chi}^{\mathrm{in}}(\boldsymbol{\zeta};\omega)$ involves the time-reversed (advanced) Green function, it is not the time-reversal of $\boldsymbol{\chi}^{\mathrm{out}}(\boldsymbol{\zeta};\omega)$, because it involves the negation, not complex conjugation, of the effective source/sink $g_\perp(\boldsymbol{\zeta};\omega)$. Indeed, the conjugate (i.e., time-reversed) wave $\boldsymbol{\chi}^{\mathrm{in}}(\boldsymbol{\zeta};\omega)^*$ may be interpreted as the usual outgoing (retarded) radiation from the effective source $-\boldsymbol{g}_\perp(\boldsymbol{\zeta}';\omega)^*$, the complex-conjugate (time-reversal) *and* negation of $\boldsymbol{g}_\perp(\boldsymbol{\zeta}';\omega)$. Using the symmetry of the Green functions and a simple change of variables, it is easily shown that for any $\zeta_2 > \zeta_0$,

$$\begin{aligned}\mathscr{P}'_{\mathrm{EM}}[\boldsymbol{\chi}^{\mathrm{out}}](\zeta_2;\omega) = -\mathscr{P}'_{\mathrm{mech}}[\boldsymbol{\varepsilon}^{\mathrm{out}}_{\boldsymbol{\chi}\perp};\boldsymbol{g}](\omega) &= \int d^3\boldsymbol{\zeta}\int d^3\boldsymbol{\zeta}'\ K(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\,\boldsymbol{g}_\perp(\boldsymbol{\zeta};\omega)^*\cdot\boldsymbol{g}_\perp(\boldsymbol{\zeta}';\omega) \\ &= \langle \boldsymbol{g}_\perp(\omega)\,|K(\omega)|\,\boldsymbol{g}_\perp(\omega)\rangle \geq 0,\end{aligned} \tag{4.205}$$

which vanishes if and only if $\boldsymbol{g}_\perp(\boldsymbol{\zeta};\omega)$ is a non-radiating source at frequency ω. In the same

manner, we also find

$$\begin{aligned}\mathscr{P}'_{\mathrm{EM}}[\boldsymbol{\chi}^{\mathrm{in}}](\zeta_2;\omega) &= -\mathscr{P}'_{\mathrm{mech}}[\boldsymbol{\varepsilon}^{\mathrm{in}}_{\boldsymbol{\chi}\perp};-\boldsymbol{g}](\omega) = -\int d^3\boldsymbol{\zeta}\int d^3\boldsymbol{\zeta}'\ K(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\,\boldsymbol{g}_\perp(\boldsymbol{\zeta};\omega)^*\cdot\boldsymbol{g}_\perp(\boldsymbol{\zeta}';\omega)\\ &= -\left\langle \boldsymbol{g}_\perp(\omega)\,|K(\omega)|\,\boldsymbol{g}_\perp(\omega)\right\rangle \le 0,\end{aligned} \tag{4.206}$$

also vanishing if and only if $\boldsymbol{g}_\perp(\boldsymbol{\zeta};\omega)$ is a non-radiating source. Thus, we see that, as anticipated,

$$\mathscr{P}'_{\mathrm{EM}}[\boldsymbol{\chi}^{\mathrm{out}}](\zeta_2;\omega) = -\mathscr{P}'_{\mathrm{EM}}[\boldsymbol{\chi}^{\mathrm{in}}](\zeta_2;\omega) = \mathscr{P}'_{\mathrm{EM}}[\boldsymbol{\chi}^{\mathrm{in}*}](\zeta_2;\omega) \ge 0. \tag{4.207}$$

It then follows that

$$\begin{aligned}-\tfrac{1}{2}\mathscr{P}'_{\mathrm{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp};\boldsymbol{g}](\omega) &= -\tfrac{1}{2}\,\mathrm{Re}\int d^3\boldsymbol{\zeta}\ [\boldsymbol{g}^*\cdot\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}] = \tfrac{1}{2}\int d^3\boldsymbol{\zeta}\ \ [\boldsymbol{g}^*\cdot\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}]\\ &= \tfrac{1}{2}\int d^3\boldsymbol{\zeta}\int d^3\boldsymbol{\zeta}'\ i\omega D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\,\boldsymbol{g}_\perp(\boldsymbol{\zeta};\omega)^*\cdot\boldsymbol{g}_\perp(\boldsymbol{\zeta}';\omega)\\ &= \int d^3\boldsymbol{\zeta}\int d^3\boldsymbol{\zeta}'\ K(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\,\boldsymbol{g}_\perp(\boldsymbol{\zeta};\omega)^*\cdot\boldsymbol{g}_\perp(\boldsymbol{\zeta}';\omega)\\ &= \left\langle \boldsymbol{g}_\perp(\omega\,|K(\omega)|\,\boldsymbol{g}_\perp(\omega)\right\rangle = -\mathscr{P}'_{\mathrm{mech}}[\boldsymbol{\varepsilon}^{\mathrm{out}}_{\boldsymbol{\chi}\perp};\boldsymbol{g}](\omega) \ge 0,\end{aligned} \tag{4.208}$$

which is the generalization of the paraxial result (4.93).

If we take $\boldsymbol{g}(\boldsymbol{\zeta};\omega) = \boldsymbol{j}(\boldsymbol{\zeta};\omega)$ to be the physical current density of interest, then the homogeneous approximation

$$\boldsymbol{\chi}_{\boldsymbol{j}}(\boldsymbol{\zeta};\omega) = \int d^3\boldsymbol{\zeta}'\,D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega) \tag{4.209}$$

agrees exactly in its outgoing components with the actual radiation:

$$\boldsymbol{\chi}^{\mathrm{out}}_{\boldsymbol{j}}(\boldsymbol{\zeta};\omega) = \boldsymbol{a}(\boldsymbol{\zeta};\omega), \tag{4.210}$$

but in order to actually satisfy the homogeneous Helmholtz equation everywhere, it must possess an equal magnitude of net power in the corresponding (i.e., conjugate) incoming modes making up $\boldsymbol{\chi}^{\mathrm{in}}(\boldsymbol{\zeta};\omega)$, power which will flow in the opposite direction through any surface enclosing the sources. This extra power appears as extra work which would be done by the actual source charges on these incoming fields, if they were present. (These advanced fields do not lead to negative work by the sources making up $\boldsymbol{j}$ because, again, they are the time-reversal of the retarded fields which would be produced by $-\boldsymbol{j}^*$, not $\boldsymbol{j}$ itself.)

Using Green's identities and the governing Helmholtz equations, one can show that this homogeneous solution may also be formally expressed in the Kirchoff-integral form, only

with the Green function $G^{\mathrm{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)$ replaced with the fundamental homogeneous solution $D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)$:

$$\boldsymbol{\chi}(\boldsymbol{\zeta};\omega) = \int\limits_{\partial V(\zeta_3)} d\sigma' \left[\boldsymbol{\chi}^{\mathrm{out}}(\boldsymbol{\zeta}';\omega)(\hat{\boldsymbol{n}}'\cdot\boldsymbol{\partial}')D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) - D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)(\hat{\boldsymbol{n}}'\cdot\boldsymbol{\partial}')\boldsymbol{\chi}^{\mathrm{out}}(\boldsymbol{\zeta}';\omega)\right]. \quad (4.211)$$

Unlike the inhomogeneous case, this remains entirely valid for any boundary radius $\zeta_3 > 0$ and any observation position $\boldsymbol{\zeta}$. In particular it holds everywhere for the outgoing vector potential $\boldsymbol{a}$ produced by the actual source of interest $\boldsymbol{j}_\perp$ and the corresponding source-free extension $\boldsymbol{\chi}_{\boldsymbol{j}}$.

Not surprisingly, this homogeneous approximation $\boldsymbol{\chi}_{\boldsymbol{j}}(\boldsymbol{\zeta};\omega)$ to $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$ is just the result of the above-mentioned one-to-one mapping between the isomorphic spaces $\mathcal{H}_{\mathrm{out}}$ and $\mathcal{H}_{\mathrm{hom}}$, involving the replacement of each "outgoing" spherical Hankel function $h_\ell^{(1)}(k\zeta)$ with the corresponding "standing-wave" spherical Bessel function $j_\ell(k\zeta)$ in the spherical wave expansion for $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$, without altering any of the multipole expansion coefficients or basis vector fields for the angular dependence. Schematically, we have:

$$\mathcal{H}_{\mathrm{out}} \to \mathcal{H}_{\mathrm{hom}} \quad (4.212a)$$

$$\boldsymbol{a}(\boldsymbol{\zeta};\omega) \to \boldsymbol{\chi}_{\boldsymbol{j}}(\boldsymbol{\zeta};\omega) \quad (4.212b)$$

$$h_\ell^{(1)}(k\zeta) \to j_\ell(k\zeta) \quad (4.212c)$$

$$\boldsymbol{W}_{\ell m}(\hat{\boldsymbol{\zeta}}) \to \boldsymbol{W}_{\ell m}(\hat{\boldsymbol{\zeta}}) \quad (4.212d)$$

$$a^{\mathrm{E}}_{\ell m}(\omega) \to \chi^{\mathrm{E}}_{\ell m}(\omega) = a^{\mathrm{E}}_{\ell m}(\omega) \quad (4.212e)$$

$$a^{\mathrm{M}}_{\ell m}(\omega) \to \chi^{\mathrm{M}}_{\ell m}(\omega) = a^{\mathrm{M}}_{\ell m}(\omega). \quad (4.212f)$$

We expect that this $\boldsymbol{\chi}_{\boldsymbol{j}}(\boldsymbol{\zeta};\omega)$ is in some sense (or, more specifically, in some metric) the closest possible homogeneous approximation to $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$, which will be confirmed below by the non-paraxial form of the Maximum Power Variational Principle.

We can also express the various radiated power spectral densities directly in terms of the multipole expansion coefficients or associated inner product, by using (4.198) in the limit $\zeta_2 \to \infty$. Asymptotically, i.e., for $kr \gg \ell$, the limiting forms of the spherical Bessel,

Neumann, and Hankel functions are:

$$j_\ell(kr) \to \quad \frac{\sin(kr - \frac{\ell\pi}{2})}{kr}, \tag{4.213a}$$

$$n_\ell(kr) \to -\frac{\cos(kr - \frac{\ell\pi}{2})}{kr}, \tag{4.213b}$$

$$h_\ell^{(1)}(kr) \to (-i)^{\ell+1}\,\frac{e^{+ikr}}{kr}, \tag{4.213c}$$

$$h_\ell^{(2)}(kr) \to (+i)^{\ell+1}\,\frac{e^{-ikr}}{kr}. \tag{4.213d}$$

Examining the form of the Sommerfeld radiation conditions, we see that, asymptotically as $k\zeta \to \infty$, any vector potential $\boldsymbol{\psi}(\boldsymbol{\zeta};\omega)$ (not necessarily a source-free solution) may be locally decomposed into its incoming and outgoing components on the spherical surface $\partial B(\zeta)$ by the projections:

$$\boldsymbol{\psi}^{\text{out}}(\boldsymbol{\zeta};\omega) \to \tfrac{1}{2}\left(1 - \tfrac{i}{k}\tfrac{\partial}{\partial\zeta}\right)\boldsymbol{\psi}(\boldsymbol{\zeta};\omega), \tag{4.214a}$$

$$\boldsymbol{\psi}^{\text{in}}(\boldsymbol{\zeta};\omega) \to \tfrac{1}{2}\left(1 + \tfrac{i}{k}\tfrac{\partial}{\partial\zeta}\right)\boldsymbol{\psi}(\boldsymbol{\zeta};\omega). \tag{4.214b}$$

Equipped with these results, it is straightforward to show that for any solution $\boldsymbol{\psi} \in \mathcal{H}_{\text{out}} \oplus \mathcal{H}_{\text{in}}$, the outgoing power spectral density is given by

$$\begin{aligned}
\mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}^{\text{out}}](\infty;\omega) &= \lim_{\zeta_2\to\infty} \frac{i\omega}{2} \int\limits_{\partial B(\zeta_2)} d^2\sigma \left[\boldsymbol{\psi}^{\text{out}}\cdot(\hat{\boldsymbol{n}}\cdot\boldsymbol{\partial})\boldsymbol{\psi}^{\text{out}*} - \boldsymbol{\psi}^{\text{out}*}\cdot(\hat{\boldsymbol{n}}\cdot\boldsymbol{\partial})\boldsymbol{\psi}^{\text{out}}\right] \\
&= \lim_{\zeta_2\to\infty}\left[\frac{i\omega}{8} \int\limits_{\partial B(\zeta_2)} d^2\sigma \left[(1 - \tfrac{i}{k}\tfrac{\partial}{\partial\zeta})\boldsymbol{\psi}\cdot(\tfrac{\partial}{\partial\zeta} - \tfrac{i}{k}\tfrac{\partial^2}{\partial\zeta^2})\boldsymbol{\psi}^*\right]\right] + c.c \\
&= \lim_{\zeta_2\to\infty}\left[\frac{i\omega}{8}\,\zeta_2^2 \int d\varphi \int d\theta\, \sin\theta \big[(1 - \tfrac{i}{k}\tfrac{\partial}{\partial\zeta})\boldsymbol{\psi}\cdot(\tfrac{\partial}{\partial\zeta} - \tfrac{i}{k}\tfrac{\partial^2}{\partial\zeta^2})\boldsymbol{\psi}^*\big]\right] + c.c,
\end{aligned} \tag{4.215}$$

which, after a little algebra, reduces to the inner product over the outgoing multipole coefficients:

$$\mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}^{\text{out}}](\infty;\omega) = \sum_{\ell m}\left[\left|\psi^{\text{E (out)}}_{\ell m}(\omega)\right|^2 + \left|\psi^{\text{M (out)}}_{\ell m}(\omega)\right|^2\right] = \left(\boldsymbol{\psi}^{\text{out}}(\omega)\,,\,\boldsymbol{\psi}^{\text{out}}(\omega)\right)_{\text{out}} \geq 0, \tag{4.216}$$

with equality (vanishing outgoing power) in the last step if and only if $\boldsymbol{\psi}^{\text{out}}(\boldsymbol{\zeta};\omega)$ vanishes identically. Similarly, we find

$$-\mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}^{\text{in}}](\infty;\omega) = \left(\boldsymbol{\psi}^{\text{in}}(\omega)\,,\,\boldsymbol{\psi}^{\text{in}}(\omega)\right)_{\text{in}} \geq 0, \tag{4.217}$$

so we may write

$$\begin{aligned}(\boldsymbol{\psi}(\omega)\,,\,\boldsymbol{\psi}(\omega)) &\equiv (\boldsymbol{\psi}^{\text{out}}(\omega)\,,\,\boldsymbol{\psi}^{\text{out}}(\omega))_{\text{out}} + (\boldsymbol{\psi}^{\text{in}}(\omega)\,,\,\boldsymbol{\psi}^{\text{in}}(\omega))_{\text{in}} \\ &= \mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}^{\text{out}}](\infty;\omega) - \mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}^{\text{in}}](\infty;\omega) \\ &= \left|\mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}^{\text{out}}](\infty;\omega)\right| + \left|\mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}^{\text{in}}](\infty;\omega)\right|.\end{aligned} \tag{4.218}$$

The inner products previously introduced for $\mathcal{H}_{\text{out}}$ and $\mathcal{H}_{\text{in}}$, and $\mathcal{H}_{\text{out}} \oplus \mathcal{H}_{\text{in}}$ are just equal, respectively, to the (scaled) outgoing power spectral density, the absolute value of the incoming spectral density, and the magnitude of total (not net) in-flowing and out-flowing power spectral density, passing through an arbitrarily remote boundary somewhere outside the support of the (real or effective) sources. Choosing $\boldsymbol{\psi} = \boldsymbol{\chi}_{\boldsymbol{j}}$, the relations (4.216) and (4.217) provide an explicit verification of (4.207). As an aside, we note that these separate incoming and outgoing power spectral densities and their (incoherent) sum, are often of more physical interest than the net power $\mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}](\infty;\omega)$. For UV and lower frequencies, detectors (such as photomultipliers or solid state photoelectric devices) are typically directionally sensitive, in the sense of an acceptance solid angle $\Omega \leq 2\pi$, so measurements of the power in the radiation $\boldsymbol{\psi}$ will typically reveal something closer to either $|\mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}^{\text{out}}](\infty;\omega)|$ or $|\mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}^{\text{in}}](\infty;\omega)|$, depending on the orientation, (or some fraction thereof, with less than full 4π solid-angular coverage of the localized source), rather than $\mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}](\infty;\omega)$. In the X-ray regime, shielding is difficult and detectors are often not directionally sensitive, and the response will likely depend on the total fluence into the detector volume, i.e., on $|\mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}^{\text{out}}](\infty;\omega)| + |\mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}^{\text{in}}](\infty;\omega)|$ rather than $\mathscr{P}'_{\text{EM}}[\boldsymbol{\psi}](\infty;\omega)$.

In a similar fashion, we may formally express the spectral density of mechanical work in terms of the multipole coefficients for the exact radiation $\boldsymbol{a}$ and the homogeneous approximant $\boldsymbol{\chi}$, by using (4.199), choosing a spherical boundary $\partial B(\zeta_2)$ and letting $\zeta_2 \to \infty$. After the algebraic dust settles, we find

$$\begin{aligned}-\tfrac{1}{2}\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp};\boldsymbol{j}](\omega) &= \text{Re}\sum_{\ell m}\left[\chi^{\text{E}}_{\ell m}(\omega)^* a^{\text{E}}_{\ell m}(\omega) + \chi^{\text{M}}_{\ell m}(\omega)^* a^{\text{M}}_{\ell m}(\omega)\right] \\ &= \text{Re}\,(\boldsymbol{\chi}^{\text{out}}(\omega)\,,\,\boldsymbol{a}(\omega))_{\text{out}}\,.\end{aligned} \tag{4.219}$$

In effect, we have managed to re-express the three-dimensional, "volume" inner product $\langle \boldsymbol{j}(\omega)\,|\,\boldsymbol{a}(\omega)\rangle$ in terms of the two-dimensional, "boundary" inner product $(\boldsymbol{\chi}^{\text{out}}(\omega)\,,\,\boldsymbol{a}(\omega))_{\text{out}}$. If we choose $\boldsymbol{\chi} = \boldsymbol{\chi}_{\boldsymbol{j}}$, then by applying (4.216), the power spectral density in the form (4.219) is seen to satisfy (4.208) explicitly.

If we attempt to use a homogeneous analog of the Kirchoff-integral expression (4.165), both $\boldsymbol{\chi}(\boldsymbol{\zeta}';\omega)$ and $D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)$ are regular solutions of the homogeneous Helmholtz equation

everywhere in space, so, by applying Green's second identity to the interior of any closed region $V(\zeta_3)$, we find that

$$\int\limits_{\partial V(\zeta_3)} d\sigma' \left[\boldsymbol{\chi}(\boldsymbol{\zeta}';\omega)(\hat{\boldsymbol{n}}'\cdot\boldsymbol{\partial}')D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) - D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)(\hat{\boldsymbol{n}}'\cdot\boldsymbol{\partial}')\boldsymbol{\chi}(\boldsymbol{\zeta}';\omega)\right] = \mathbf{0}, \tag{4.220}$$

which, as will be seen below, will have important implications for the Poynting flux of a source-free solution, but cannot be used to determine $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega)$ from spatial data on some remote boundary.

However, in the time domain the fundamental homogeneous solution $D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau)$ can be used as a kernel for the Cauchy "initial" value problem for the wave equation. Because of time translation invariance, the time-domain versions of the free-space Green functions are just inverse Fourier transforms of their frequency-domain cousins:

$$G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau') = \frac{\delta(\tau-\tau'-\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|)}{4\pi\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|}, \tag{4.221a}$$

$$G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau') = \frac{\delta(\tau-\tau'+\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|)}{4\pi\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|}, \tag{4.221b}$$

and the fundamental homogeneous solution in the time domain is therefore

$$\begin{aligned} D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau')(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau') &= G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau') - G^{\text{adv}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau') \\ &= \frac{\delta(\tau-\tau'-\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|) - \delta(\tau-\tau'+\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|)}{4\pi\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|}. \end{aligned} \tag{4.222}$$

Using this explicit form, it is easy to verify that

$$\left(\partial^2 - \tfrac{\partial^2}{\partial\tau^2}\right) D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau') = \left[\partial^2 - \tfrac{\partial^2}{\partial\tau'^2}\right] D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau') = 0. \tag{4.223a}$$

$$\lim_{\tau\to\tau'} D(\boldsymbol{\zeta}-\boldsymbol{\zeta}';\tau-\tau') = \lim_{\tau\to\tau'} \tfrac{\partial^2}{\partial\tau^2} D(\boldsymbol{\zeta}-\boldsymbol{\zeta}';\tau-\tau') = 0, \tag{4.223b}$$

$$\lim_{\tau\to\tau'} \tfrac{\partial}{\partial\tau} D(\boldsymbol{\zeta}-\boldsymbol{\zeta}';\tau-\tau') = 2\delta(\boldsymbol{\zeta}-\boldsymbol{\zeta}'). \tag{4.223c}$$

With these results, it is straightforward to verify that any source-free solution $\boldsymbol{\chi}(\boldsymbol{\zeta},\tau)$ to the wave equation may be written in terms of functional and derivative data over all space at any one time τ'' as:

$$\boldsymbol{\chi}(\boldsymbol{\zeta};\tau) = \tfrac{1}{2}\int d\boldsymbol{\zeta}' \left[D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau'')\tfrac{\partial}{\partial\tau''}\boldsymbol{\chi}(\boldsymbol{\zeta}';\tau'') + \tfrac{\partial}{\partial\tau}D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau'')\boldsymbol{\chi}(\boldsymbol{\zeta}';\tau'')\right], \tag{4.224}$$

or equivalently in the convolutional form

$$\boldsymbol{\chi}(\boldsymbol{\zeta};\tau) = \int d\boldsymbol{\zeta}' \int d\tau'\, D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau') \left[\delta(\tau'-\tau'')\tfrac{\partial}{\partial\tau'}\boldsymbol{\chi}(\boldsymbol{\zeta}';\tau'') + \tfrac{1}{2}\delta'(\tau'-\tau'')\boldsymbol{\chi}(\boldsymbol{\zeta}';\tau')\right]. \tag{4.225}$$

This may be thought of as a precise quantitative statement of Huygens' principle for source-free fields. By integrating by parts, it is straightforward to show that if the "initial" vector potential $\boldsymbol{\chi}(\boldsymbol{\zeta}';\tau'')$ satisfies the Coulomb gauge constraint, then $\boldsymbol{\chi}(\boldsymbol{\zeta}';\tau)$ will as well. Performing a Fourier transform in the time τ, we find an implicit effective source formulation:

$$\boldsymbol{\chi}(\boldsymbol{\zeta};\omega) = \int d\boldsymbol{\zeta}'\, D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\tfrac{1}{2}e^{i\omega\tau''}\left[\tfrac{\partial}{\partial\tau''}\boldsymbol{\chi}(\boldsymbol{\zeta}';\tau'') - i\omega\boldsymbol{\chi}(\boldsymbol{\zeta}';\tau'')\right], \tag{4.226}$$

so $\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega;\tau'') = \frac{1}{2}e^{i\omega\tau''}\left[\frac{\partial}{\partial\tau''}\boldsymbol{\chi}(\boldsymbol{\zeta}';\tau'') - i\omega\boldsymbol{\chi}(\boldsymbol{\zeta}';\tau'')\right]$, provides, at least formally, one effective solenoidal source for the solenoidal, homogeneous solution $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega)$. As a function of ω, it is a little ill-behaved (not square-integrable, for example), but it is regularized by spatial integration over the kernel $D(\omega)$. (We will see shortly that effective sources $\boldsymbol{j}_\perp$ for given source-free solutions $\boldsymbol{\chi}$ are not unique.)

4.4.5 Variational Principle

Our general variational principle follows directly from the nonnegative-definiteness of the operator $K(\omega)$, i.e., equation (4.179), and the above energy conservation considerations. Re-tracing the standard proof of the Cauchy-Schwarz inequality, we find that, with a little extra care, it goes through almost unchanged if we replace the positive definite inner product with a positive-semidefinite pseudo-product. In particular, with the bilinear form associated with $K(\omega)$, one can show

$$\left|\left\langle \boldsymbol{g}(\omega)\left|K(\omega)\right|\boldsymbol{g}'(\omega)\right\rangle\right|^2 \leq \left\langle \boldsymbol{g}(\omega)\left|K(\omega)\right|\boldsymbol{g}(\omega)\right\rangle \left\langle \boldsymbol{g}'(\omega)\left|K(\omega)\right|\boldsymbol{g}'(\omega)\right\rangle, \tag{4.227}$$

with strict equality if and only if there exist constants δ, δ', $\delta_0 \in \mathbb{C}$, not all zero, and a non-radiating source $\boldsymbol{g}_0(\boldsymbol{\zeta};\omega)$ which is not identically zero but is otherwise arbitrary, such that

$$\delta\,\boldsymbol{g}(\boldsymbol{\zeta};\omega) + \delta'\,\boldsymbol{g}'(\boldsymbol{\zeta};\omega) + \delta_0\,\boldsymbol{g}_0(\boldsymbol{\zeta};\omega) = \boldsymbol{0}. \tag{4.228}$$

We also have, for any complex number $c \in \mathbb{C}$, the inequalities

$$\mathrm{Re}[c] \leq |\mathrm{Re}[c]| \leq |c|, \tag{4.229}$$

with overall equality if and only if c is real and nonnegative. Choosing $\boldsymbol{g}' = \boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega)$ to be the solenoidal source responsible for the actual radiation $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$ and $\boldsymbol{g} = \boldsymbol{g}_\perp(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha})$ to be an effective source corresponding to a solenoidal, source-free trial vector field $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha})$ depending on certain variational parameters collected in the vector $\boldsymbol{\alpha}$, these inequalities

yield:

$$\begin{aligned} -\tfrac{1}{2}\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp};\boldsymbol{j}](\omega;\boldsymbol{\alpha}) &\leq \tfrac{1}{2}\left|\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp};\boldsymbol{j}](\omega;\boldsymbol{\alpha})\right| \\ &\leq \left[\mathscr{P}'_{\text{EM}}[\boldsymbol{\chi}^{\text{out}}](\zeta_2;\omega;\boldsymbol{\alpha})\right]^{1/2}\left[\mathscr{P}'_{\text{EM}}[\boldsymbol{a}](\zeta_2;\omega)\right]^{1/2}, \end{aligned} \tag{4.230}$$

with overall equality assuming $\mathscr{P}'_{\text{EM}}[\boldsymbol{a}](\zeta_2;\omega) \neq 0$, if and only if

$$\boldsymbol{\chi}^{\text{out}}(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha}) \propto \boldsymbol{a}(\boldsymbol{\zeta};\omega). \tag{4.231}$$

(In the trivial case where $\mathscr{P}'_{\text{EM}}[\boldsymbol{a}](\zeta_2;\omega) = 0$, equality is achieved trivially for $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega) = \boldsymbol{0}$ or indeed for any $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega)$ which is Hilbert space orthogonal to $\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega)$.)

In practice, we can obtain an actual variational approximation with one of two different but ultimately equivalent procedures, each arriving at the same final result when optimizing over equivalent families of trial solutions. In the first approach, we may partition the variational parameters as $\boldsymbol{\alpha} = (\eta_0, \theta_0, \boldsymbol{\alpha}')^{\text{T}}$, where $\boldsymbol{\alpha}' = (\alpha'_1, \alpha'_2, \dots,)^{\text{T}}$, and consider a restricted family of solenoidal, homogeneous trial solutions $\boldsymbol{v}(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha}')$ normalized to unit power:

$$\mathscr{P}'_{\text{EM}}[\boldsymbol{v}^{\text{out}}](\zeta_2;\omega;\boldsymbol{\alpha}') = 1, \tag{4.232}$$

or really to any conveniently-chosen but *constant* value. (Recall that this power is independent of the precise choice of surface $V(\zeta_2)$ as long as $\zeta_2 > \zeta_1$.) We then maximize $|\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{v}\perp};\boldsymbol{j}](\omega;\boldsymbol{\alpha}')|$ for fixed output power, in order to determine the optimal parameter values $\tilde{\boldsymbol{\alpha}}' = \tilde{\boldsymbol{\alpha}}'[\boldsymbol{j}](\omega)$ and therefore the *relative* shape (spatial profile and polarization) of the radiation. Once the relative profile is determined, we then renormalize this trial solution, taking

$$\boldsymbol{\chi}(\boldsymbol{\zeta};\omega;\tilde{\boldsymbol{\alpha}}) = \tilde{\eta}_0 e^{i\tilde{\theta}_0}\,\boldsymbol{v}(\boldsymbol{\zeta};\omega;\tilde{\boldsymbol{\alpha}}'), \tag{4.233}$$

for some real phase $\tilde{\theta}_0$, $0 \leq \tilde{\theta}_0 < 2\pi$, and nonnegative amplitude $\tilde{\eta}_0 \geq 0$. To maximize the mechanical work term, the overlap integral (3D inner product) between (scaled) solenoidal electric field and current density must necessarily be real and negative, so $\tilde{\theta}_0$ is chosen so as to ensure

$$-\tfrac{1}{2}\,\text{Re}\Big[e^{-i\tilde{\theta}_0}\left\langle \boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\omega;\tilde{\boldsymbol{\alpha}}')\,\middle|\,\boldsymbol{j}(\omega)\right\rangle\Big] \geq 0 \tag{4.234a}$$

$$\text{Im}\Big[e^{-i\tilde{\theta}_0}\left\langle \boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\omega);\tilde{\boldsymbol{\alpha}}'\,\middle|\,\boldsymbol{j}(\omega)\right\rangle\Big] = 0. \tag{4.234b}$$

Finally, the positive amplitude $\tilde{\eta}_0 > 0$ is chosen to ensure energy balance between the power (spectral density) in the approximate outgoing fields and the work which would be done on these fields, if present in the region of the sources, by those actual sources:

$$-\tfrac{1}{2}\eta_0 e^{-i\tilde{\theta}_0}\left\langle \boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\omega)\,|\,\boldsymbol{j}(\omega)\right\rangle = \eta_0^2\,\mathscr{P}'_{\text{EM}}[\boldsymbol{v}^{\text{out}}](\zeta_2;\omega;\tilde{\boldsymbol{\alpha}}), \tag{4.235}$$

or

$$\tilde{\eta}_0 = \frac{-\frac{1}{2}\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}; \boldsymbol{j}](\omega; \tilde{\boldsymbol{\alpha}}')}{\mathscr{P}'_{\text{EM}}[\boldsymbol{\chi}^{\text{out}}](\zeta_2; \omega; \tilde{\boldsymbol{\alpha}})}. \tag{4.236}$$

Note that under the assumptions of energy conservation given the sources, the overall amplitude, set by $\tilde{\eta}_0$, is uniquely determined, and the overall phase $\tilde{\theta}_0$ is determined modulo 2π, because the mechanical work done on/by the sources is linear in the fields, but the radiated power is quadratic in the fields. As in the paraxial case, we can also interpret this as a maximization of the small-signal, no-recoil gain which would be experienced by the trial field if it were to be incident on the sources. Upon substituting back into the Cauchy-Schwarz inequality, canceling common terms, and squaring, we deduce

$$-\tfrac{1}{2}\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}; \boldsymbol{j}](\omega; \tilde{\boldsymbol{\alpha}}) \leq \mathscr{P}'_{\text{EM}}[\boldsymbol{\chi}^{\text{out}}](\zeta_2; \omega; \tilde{\boldsymbol{\alpha}}), \tag{4.237}$$

with equality if and only if

$$\boldsymbol{\chi}^{\text{out}}(\boldsymbol{\zeta}; \omega; \tilde{\boldsymbol{\alpha}}) = \boldsymbol{a}(\boldsymbol{\zeta}; \omega), \tag{4.238}$$

or equivalently,

$$\boldsymbol{\chi}(\boldsymbol{\zeta}; \omega; \tilde{\boldsymbol{\alpha}}) = \boldsymbol{\chi}_{\boldsymbol{j}}(\boldsymbol{\zeta}; \omega) = \int d^3\boldsymbol{\zeta}'\, D(\boldsymbol{\zeta}; \boldsymbol{\zeta}'; \omega)\, \boldsymbol{j}_\perp(\boldsymbol{\zeta}; \omega). \tag{4.239}$$

Thus, we see that the variational approximation $\boldsymbol{\chi}^{\text{out}}(\zeta_2; \omega; \tilde{\boldsymbol{\alpha}})$ may also be obtained more directly from a simultaneous optimization, with respect to all parameters in $\boldsymbol{\alpha}$, of the outgoing power, subject to the constraint of energy conservation:

$$\tilde{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \Big[\mathscr{P}'_{\text{EM}}[\boldsymbol{\chi}^{\text{out}}](\zeta_2; \omega; \boldsymbol{\alpha})\Big] \tag{4.240a}$$

$$\text{s.t.} \quad -\tfrac{1}{2}\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}; \boldsymbol{j}](\omega; \boldsymbol{\alpha}) = \mathscr{P}'_{\text{EM}}[\boldsymbol{\chi}^{\text{out}}](\zeta_2; \omega; \boldsymbol{\alpha}). \tag{4.240b}$$

This is our desired result. As anticipated, we have arrived at a variational principle which holds in the general, non-paraxial case, which is exactly analogous to that found previously in the paraxial limit. Unfortunately, there seems to be no simple way to express the constrained optimization problem (4.240) as an unconstrained maximization of a ratio between the mechanical work and radiated power (or some functions thereof).

It is also instructive to also see how this all works out explicitly in the spherical wave basis. Applying the usual Cauchy-Schwarz inequality directly in $\mathcal{H}_{\text{out}}$, and using equations (4.216) and (4.219), we have

$$\begin{aligned} -\tfrac{1}{2}\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}; \boldsymbol{j}](\omega; \boldsymbol{\alpha}) &= \text{Re}\,(\boldsymbol{\chi}^{\text{out}}(\omega; \boldsymbol{\alpha})\,,\, \boldsymbol{a}(\omega))_{\text{out}} \leq |(\boldsymbol{\chi}^{\text{out}}(\omega; \boldsymbol{\alpha})\,,\, \boldsymbol{a}(\omega))_{\text{out}}| \\ &\leq (\boldsymbol{\chi}^{\text{out}}(\omega; \boldsymbol{\alpha})\,,\, \boldsymbol{\chi}^{\text{out}}(\omega; \boldsymbol{\alpha}))_{\text{out}}^{1/2}\, (\boldsymbol{a}(\omega)\,,\, \boldsymbol{a}(\omega))_{\text{out}}^{1/2} \\ &= \mathscr{P}'_{\text{EM}}[\boldsymbol{\chi}^{\text{out}}](\infty; \omega)^{1/2} \mathscr{P}'_{\text{EM}}[\boldsymbol{a}](\infty; \omega)^{1/2}, \end{aligned} \tag{4.241}$$

with overall equality if and only if $\boldsymbol{\chi}^{\text{out}}(\omega;\boldsymbol{\alpha}) = \kappa(\omega)\,\boldsymbol{a}(\omega)$ for some positive real scalar $\kappa(\omega) > 0$, or equivalently when

$$\frac{a^{\text{E}}_{\ell m}(\omega)}{\chi^{\text{E}}_{\ell m}(\omega;\boldsymbol{\alpha})} = \frac{a^{\text{M}}_{\ell m}(\omega)}{\chi^{\text{M}}_{\ell m}(\omega;\boldsymbol{\alpha})} = \frac{1}{\kappa(\omega)} > 0 \tag{4.242}$$

for all multipole coefficients of $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$. Imposing the additional energy conservation relation on $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha})$, this becomes

$$-\tfrac{1}{2}\mathscr{P}'_{\text{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp};\boldsymbol{j}](\omega;\boldsymbol{\alpha}) \leq \mathscr{P}_{\text{EM}}[\boldsymbol{a}](\infty;\omega), \tag{4.243}$$

with equality if and only if $\boldsymbol{\chi}^{\text{out}}(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha}) = \boldsymbol{a}(\boldsymbol{\zeta};\omega)$. Because in practice we will not have knowledge of the actual multipole expansion coefficients for $\boldsymbol{a}$ in situations where we are resorting to a variational approximation, this "angular-representation" or "boundary-value" formulation is not directly applicable for calculation, but it does lead directly to the relevant power inequality without any worries about a semidefinite or "pseudo" inner product. We also see that the MPVP follows immediately as a consequence of the minimum distance property of orthogonal projectors in these multipolar Hilbert spaces: the variational approximation $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega;\tilde{\boldsymbol{\alpha}})$ is simply that which minimizes the distance, in the outgoing space, to the actual solution:

$$\tilde{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}}\Big[\big(\boldsymbol{a}(\omega) - \boldsymbol{\chi}^{\text{out}}(\omega;\boldsymbol{\alpha})\,,\,\boldsymbol{a}(\omega) - \boldsymbol{\chi}^{\text{out}}(\omega;\boldsymbol{\alpha})_{\text{out}}\big)^{1/2}\Big], \tag{4.244}$$

or equivalently minimizes the distance, in the homogenous space, between trial function and the source-free extension to the actual outgoing radiation,

$$\tilde{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}}\Big[\big(\boldsymbol{\chi}_{\boldsymbol{j}}(\omega) - \boldsymbol{\chi}(\omega;\boldsymbol{\alpha})\,,\,\boldsymbol{a}(\omega) - \boldsymbol{\chi}_{\boldsymbol{j}}(\omega;\boldsymbol{\alpha})\big)^{1/2}\Big], \tag{4.245}$$

each within the considered family of solenoidal, source-free, trial wavefields.

For the variational principle to be fully consistent, we should also verify that the source-free approximant $\boldsymbol{\chi}_{\boldsymbol{j}}(\boldsymbol{\zeta};\omega)$ is the closest, or best, source-free extension to the actual outgoing component $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$ in some precise sense. In fact, using the Green function representations, it is easy to see that the source-free solution $\boldsymbol{\chi}_{\boldsymbol{j}}(\boldsymbol{\zeta};\omega)$ is the *unique* natural homogeneous extension to, or extrapolant of, the outgoing radiation $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$. In the (scaled) time domain, for a given localized source $\boldsymbol{j}(\boldsymbol{\zeta};\tau)$, the most general solenoidal solution to the wave-equation can always be written in terms of *either* the advanced or retarded Green function

$$\begin{aligned}\boldsymbol{a}^{\text{gen}}(\boldsymbol{\zeta};\tau) &= \boldsymbol{a}^{\text{init}}(\boldsymbol{\zeta};\tau) + \int d\tau \int d^3\boldsymbol{\zeta}'\, G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau')\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\tau') \\ &= \boldsymbol{a}^{\text{fin}}(\boldsymbol{\zeta};\tau) + \int d\tau \int d^3\boldsymbol{\zeta}'\, G^{\text{adv}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau')\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\tau'),\end{aligned} \tag{4.246}$$

where $\boldsymbol{a}^{\text{init}}(\boldsymbol{\zeta};\tau)$ and $\boldsymbol{a}^{\text{fin}}(\boldsymbol{\zeta};\tau)$ are arbitrary solutions to the source-free wave equation everywhere in space and time (to be determined by boundary conditions), while the time-domain Green functions may be obtained from their respective frequency-domain versions by Fourier transforms:

$$G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau')\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\tau') = \frac{\delta(\tau-\tau'-\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|}{4\pi\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|}, \tag{4.247a}$$

$$G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau')\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\tau') = \frac{\delta(\tau-\tau'+\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|}{4\pi\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|}. \tag{4.247b}$$

In light of the time-translation invariance of the free-space Green functions, this is of course equivalent to the frequency-domain expression

$$\begin{aligned}\boldsymbol{a}^{\text{gen}}(\boldsymbol{\zeta};\omega) &= \boldsymbol{a}^{\text{init}}(\boldsymbol{\zeta};\omega) + \int d^3\boldsymbol{\zeta}'\, G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega) \\ &= \boldsymbol{a}^{\text{fin}}(\boldsymbol{\zeta};\omega) + \int d\tau \int d^3\boldsymbol{\zeta}'\, G^{\text{adv}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega).\end{aligned} \tag{4.248}$$

where $\boldsymbol{a}^{\text{init}}(\boldsymbol{\zeta};\omega)$ and $\boldsymbol{a}^{\text{fin}}(\boldsymbol{\zeta};\omega)$ are the Fourier transforms of the corresponding homogeneous time-domain solutions, so for each frequency component ω are solutions to the source-free Helmholtz equation over all space. Given the known existence and uniqueness results for this equation, the homogeneous contributions $\boldsymbol{a}^{\text{init}}$ and $\boldsymbol{a}^{\text{fin}}$ must be uniquely determined by the source $\boldsymbol{j}_\perp$ and well-posed boundary conditions on $\boldsymbol{a}^{\text{gen}}$. As $\tau \to -\infty$ (i.e., before the localized source $\boldsymbol{j}_\perp(\boldsymbol{\zeta};\tau)$ is turned on, or otherwise becomes appreciable), $\boldsymbol{a}^{\text{gen}}(\boldsymbol{\zeta};\tau) \to \boldsymbol{a}^{\text{init}}(\boldsymbol{\zeta};\tau)$, which represents any *initial* source-free fields, i.e., any radiation effectively produced by some additional, infinitely-remote sources at $\tau \to -\infty$ and/or $\|\boldsymbol{\zeta}\| \to \infty$, and then propagated to later times according to the source-free wave equation. As $\tau \to +\infty$ (i.e., after the localized source $\boldsymbol{j}_\perp(\boldsymbol{\zeta};\tau)$ is turned off, or otherwise becomes negligible), $\boldsymbol{a}^{\text{gen}}(\boldsymbol{\zeta};\tau) \to \boldsymbol{a}^{\text{fin}}(\boldsymbol{\zeta};\tau)$, which represents the *final* source-free fields everywhere in space, extrapolated back to earlier times according to free-space propagation. By construction, $\boldsymbol{a}^{\text{fin}}$ must contain, in its outgoing components, precisely (and only) that outgoing radiation already present in $\boldsymbol{a}^{\text{fin}}$ or that which is actually generated by the source $\boldsymbol{j}_\perp$. Returning to the frequency-domain, the difference between these homogeneous solutions

$$\begin{aligned}\boldsymbol{a}^{\text{fin}}(\boldsymbol{\zeta};\omega) - \boldsymbol{a}^{\text{init}}(\boldsymbol{\zeta};\omega) &= \int d^3\boldsymbol{\zeta}'\, G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega) - \int d^3\boldsymbol{\zeta}'\, G^{\text{adv}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega) \\ &= \int d^3\boldsymbol{\zeta}'\, \left[G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) - G^{\text{adv}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\right]\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega) \\ &= \int d^3\boldsymbol{\zeta}'\, D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)\boldsymbol{j}_\perp(\boldsymbol{\zeta}';\omega) = \boldsymbol{\chi}_{\boldsymbol{j}}(\boldsymbol{\zeta};\omega)\end{aligned} \tag{4.249}$$

is itself a solenoidal solution to the source-free wave equation, and in fact must be the unique solution which agrees exactly in its outgoing components with the retarded radiation $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$ produced by $\boldsymbol{j}(\boldsymbol{\zeta};\omega)$. By assumption, we are ignoring radiation from other, remote sources, so we take $\boldsymbol{a}^{\text{init}}(\boldsymbol{\zeta};\omega) = \boldsymbol{0}$, and $\boldsymbol{a}^{\text{fin}}(\boldsymbol{\zeta};\omega) = \boldsymbol{\chi}_{\boldsymbol{j}}(\boldsymbol{\zeta};\omega)$ is thus the unique homogeneous extrapolant to $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$. Any other vector potential will either include different outgoing radiation, contain some additional irrotational component violating the gauge constraint, or else not satisfy the source-free wave equation everywhere in space and time.

In principle, we can extract the outgoing components from the homogeneous trial solution, at least up to some possible arbitrariness in the near-field terms. In the time-domain, provided the source $\boldsymbol{j}_\perp(\boldsymbol{\zeta};\tau)$ eventually turns off or at at least becomes negligible, the (approximation to the) source-free extrapolant $\boldsymbol{\chi}_{\boldsymbol{j}}(\boldsymbol{\zeta};\tau)$ will actually be equal to the (approximation to the) retarded solution $\boldsymbol{a}(\boldsymbol{\zeta};\tau)$ at sufficiently late times and/or large distances, because the added incoming, or time-reversed, component is non-zero only at advanced times relative to the active duration of the source; that is, $\boldsymbol{\chi}(\boldsymbol{\zeta};\tau) \to \boldsymbol{\chi}_{\text{out}}(\boldsymbol{\zeta};\tau)$ asymptotically for $\|\boldsymbol{\zeta}\| + \tau \gg \zeta_1 + \tau_1$, where ζ_1 is the effective radial size of the source and τ_1 is its effective duration. In the frequency domain and in spherical coordinates, the outgoing components may be extracted asymptotically from the homogeneous solution in the far field, i.e., for $\boldsymbol{\zeta} \gg \zeta_1$ and $k\|\boldsymbol{\zeta}\| \gg 1$, by using the Sommerfeld projections (4.214). Formally, at least, this outgoing solution can be extrapolated back to finite $\boldsymbol{\zeta}$ using the Kirchoff-like integral relation (4.165), resulting in

$$\begin{aligned}\boldsymbol{\chi}_{\text{out}}(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha}) = \lim_{\zeta_3\to\infty} \int_{\partial V(\zeta_3)} d\sigma' \, \Big[&(\hat{\boldsymbol{n}}'\cdot\boldsymbol{\partial}') G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega) \left[\tfrac{1}{2}\left(1 - \tfrac{i}{k}\tfrac{\partial}{\partial\zeta'}\right)\right] \boldsymbol{\chi}(\boldsymbol{\zeta}';\omega;\boldsymbol{\alpha}) \\ &- G^{\text{ret}}(\boldsymbol{\zeta};\boldsymbol{\zeta}';\omega)(\hat{\boldsymbol{n}}'\cdot\boldsymbol{\partial}') \left[\tfrac{1}{2}\left(1 - \tfrac{i}{k}\tfrac{\partial}{\partial\zeta'}\right)\right] \boldsymbol{a}(\boldsymbol{\zeta}';\omega)\Big] .\end{aligned} \tag{4.250}$$

Of course, as an approximation to $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$ this is only expected to agree with the radiative components as the trial function approximation improves and $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha}) \to \boldsymbol{\chi}_{\boldsymbol{j}}(\boldsymbol{\zeta};\omega)$

Turning for a moment to the time domain, formally, the outgoing component $\boldsymbol{a}(\boldsymbol{\zeta};\tau)$ at earlier times may be extracted from the asymptotic late-time and/or remotely-observed $\boldsymbol{\chi}_{\boldsymbol{j}}(\boldsymbol{\zeta};\tau)$ via an an extrapolation procedure, by using the fundamental homogeneous solution $D(\boldsymbol{\zeta} - \boldsymbol{\zeta}';\tau)$ as the kernel for the Cauchy "initial" value problem for the wave equation. Using the explicit form for the Green functions, it is straightforward to verify that

$$\lim_{\tau\to\tau'} D(\boldsymbol{\zeta} - \boldsymbol{\zeta}';\tau - \tau') = \lim_{\tau\to\tau'} \tfrac{\partial^2}{\partial\tau^2} D(\boldsymbol{\zeta} - \boldsymbol{\zeta}';\tau - \tau') = 0, \tag{4.251a}$$

$$\lim_{\tau\to\tau'} \tfrac{\partial}{\partial\tau} D(\boldsymbol{\zeta} - \boldsymbol{\zeta}';\tau - \tau') = 2\delta(\boldsymbol{\zeta} - \boldsymbol{\zeta}'), \tag{4.251b}$$

so that the any source-free solution $\boldsymbol{\chi}(\boldsymbol{\zeta};\tau)$ to the wave function can be written in terms of functional and derivative data at any one time τ' as:

$$\boldsymbol{\chi}(\boldsymbol{\zeta};\tau) = \tfrac{1}{2}\int d\boldsymbol{\zeta}' \left[D(\boldsymbol{\zeta}-\boldsymbol{\zeta}';\tau-\tau')\tfrac{\partial}{\partial\tau'}\boldsymbol{\chi}(\boldsymbol{\zeta}';\tau') + \tfrac{\partial}{\partial\tau}D(\boldsymbol{\zeta}-\boldsymbol{\zeta}';\tau-\tau')\boldsymbol{\chi}(\boldsymbol{\zeta}';\tau')\right]. \quad (4.252)$$

Therefore, from uniqueness it must be the case that

$$\boldsymbol{\chi_j}(\boldsymbol{\zeta};\tau) = \tfrac{1}{2}\int d^3\boldsymbol{\zeta}' \left[D(\boldsymbol{\zeta}-\boldsymbol{\zeta}';\tau-\tau')\tfrac{\partial}{\partial\tau'}\boldsymbol{a}(\boldsymbol{\zeta}';\tau') + \tfrac{\partial}{\partial\tau}D(\boldsymbol{\zeta}-\boldsymbol{\zeta}';\tau-\tau')\boldsymbol{a}(\boldsymbol{\zeta}';\tau)\right], \quad (4.253)$$

choosing any convenient time $\tau' > \tau_1$ where $\boldsymbol{a}(\boldsymbol{\zeta};\tau') = \boldsymbol{\chi_j}(\boldsymbol{\zeta};\tau')$ and its time derivative $\frac{\partial}{\partial\tau'}\boldsymbol{a}(\boldsymbol{\zeta};\tau')$ is assumed known (or approximated) over all space. Alternatively, one could use the time-domain version of the Kirchoff-integral relation (4.211), obtaining:

$$\boldsymbol{\chi_j}(\boldsymbol{\zeta};\tau) = \int\limits_{\partial V(\zeta_3)} d\sigma' \int d\tau' \left[\boldsymbol{a}(\boldsymbol{\zeta}';\tau)(\hat{\boldsymbol{n}}'\cdot\boldsymbol{\partial}')D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau') - D(\boldsymbol{\zeta};\boldsymbol{\zeta}';\tau-\tau')(\hat{\boldsymbol{n}}'\cdot\boldsymbol{\partial}')\boldsymbol{a}(\boldsymbol{\zeta}';\tau)\right] \quad (4.254)$$

given vector potential and normal-derivative data on some closed surface $\partial V(\zeta_3)$.

The output $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega;\tilde{\boldsymbol{\alpha}})$ of the MPVP is therefore actually better considered a variational approximation to the source-free "extension," "embedding," or "extrapolation" $\boldsymbol{\chi_j}(\boldsymbol{\zeta};\omega)$ rather than to the actual outgoing vector potential $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$. If the radiation is sufficiently directional, then the outgoing component $\boldsymbol{\chi}^{\text{out}}(\boldsymbol{\zeta};\omega;\tilde{\boldsymbol{\alpha}})$ of the trial function may be determined simply by restricting attention to some finite solid angle or finite range of wavevectors. In the most general case, where the "spurious" incoming radiation may overlap spatially with the "physical" outgoing radiation, at least in the frequency domain, it is more difficult to disentangle $\boldsymbol{\chi}^{\text{out}}$ and $\boldsymbol{\chi}^{\text{in}}$ at an arbitrary position. If the spherical wave expansion is known, then the outgoing projection may be easily determined everywhere in space, but, generically, direct use of the spherical-waves as the variational basis is not expected to prove convenient. (This would correspond to just using a truncated multipole expansion for the radiation.) In the far-field limit $k\zeta \to \infty$, at least, the outgoing component may be determined easily by using the Sommerfeld projections (4.214). In the time domain, the unwanted advanced component of the radiation will pass any particular remote point before it is "absorbed" by the source (or rather, appears to "bounce" off the source), so if the source could be assumed to be of finite extent in both time and space, then at points sufficiently far from the support of the source, the retarded component of the radiation could be unambiguously separated from the advanced component simply by suitably delaying the onset time of observation. However, this method cannot necessarily distinguish incoming from outgoing radiation at points within or close to the source, and, in practice, will encounter other ambiguities; it is desirable to allow for analytic sources which are only

weakly localized in time (i.e., rapidly decaying outside of some nominal time interval, but non-zero), and, as we have discussed, the transverse current density $\boldsymbol{j}_\perp(\boldsymbol{\zeta},\tau)$ is in general rapidly decaying, but not strictly vanishing, even when the full current density $\boldsymbol{j}(\boldsymbol{\zeta},\tau)$ is of compact support, so the distinction between advanced and retarded times at some finite distance from an extended solenoidal source may not always be clear.

4.5 A Simpler, Self-Contained Derivation

After developing this rather extensive framework, and rather lengthy demonstration, the simplicity of the MPVP itself suggested to us that a simpler proof should be possible, and little searching in fact uncovered one which, although perhaps a bit less rigorous, is far briefer. Because some readers may be turn directly to this section after skipping or skimming the previous sections, we have decided to present a brief but more or less *self-contained* development, entailing some repetition in a slightly different form, and unlike other sections have even retained standard Gaussian units for easier comparison. Readers who have had their fill of mathematical manipulations can skip to the discussion without missing any new results.

4.5.1 Governing Equations in Physical Units

We will still find it to convenient to work in the frequency-domain, and to make use the Helmholtz-Hodge theorem to formally decompose fields and sources into solenoidal (i.e., divergenceless, or functionally transverse) and irrotational (i.e., curl-free, or functionally longitudinal) components. Maxwell's equations can then be written in Gaussian units as

$$\boldsymbol{e}_\parallel = -\boldsymbol{\nabla}\phi = -\tfrac{4\pi i}{\omega}\boldsymbol{j}_\parallel, \tag{4.255a}$$

$$\boldsymbol{b}_\parallel = \boldsymbol{0}, \tag{4.255b}$$

$$\boldsymbol{\nabla}\times\boldsymbol{e}_\perp = ik\boldsymbol{\nabla}\times\boldsymbol{a}_\perp = ik\boldsymbol{b}_\perp = ik\boldsymbol{b}, \tag{4.255c}$$

$$\boldsymbol{\nabla}\times\boldsymbol{b}_\perp = -\nabla^2\boldsymbol{a}_\perp = \tfrac{4\pi}{c}\boldsymbol{j}_\perp - ik\boldsymbol{e}_\perp = \tfrac{4\pi}{c}\boldsymbol{j}_\perp + k^2\boldsymbol{a}_\perp. \tag{4.255d}$$

We will mostly work in coordinate-independent language, but for convenience we will choose the spatial origin to lie somewhere in the vicinity of the support of the localized source $\boldsymbol{j}$, and will sometimes express the position vector $\boldsymbol{x}$ in either Cartesian coordinates $(x,\, y,\, z)$ or spherical coordinates (r,θ,ϕ) with respect to this origin and some orientation of axes.

In the frequency-domain, and in free space (apart from the prescribed sources), the

Coulomb-gauge vector potential satisfies the inhomogeneous vector Helmholtz equation:

$$\left(\nabla^2 + k^2\right) \boldsymbol{a}_\perp(\boldsymbol{x};\omega) = -\tfrac{4\pi}{c}\boldsymbol{j}_\perp(\boldsymbol{x};\omega), \tag{4.256}$$

together with the transverse gauge constraint

$$\boldsymbol{\nabla}\cdot\boldsymbol{a}_\perp(\boldsymbol{x};\omega) = 0, \tag{4.257}$$

everywhere in space, and for any reasonably well-behaved source current density has a unique solution satisfying the Sommerfeld outgoing radiation asymptotic boundary conditions as $r \to \infty$:

$$\lim_{r\to\infty} r\left(\tfrac{\partial}{\partial r} - ik\right)\boldsymbol{a}_\perp = \boldsymbol{0} \qquad \text{(uniformly in } \theta,\, \phi), \tag{4.258}$$

ensuring that the sources act as sources and not sinks.

All physical quantities are purely real in the time-domain, so sources and fields satisfy symmetries such as $\boldsymbol{a}_\perp(\boldsymbol{x};-\omega) = \boldsymbol{a}_\perp(\boldsymbol{x};\omega)^*$, and zero-frequency solutions must be electrostatic or magnetostatic in nature and will contain no actual radiation fields, so only positive-frequency ($\omega > 0$) components (analytic signals) need be considered here. Note that while the full current density $\boldsymbol{j}(\boldsymbol{x};\omega)$ is assumed to vanish outside of some finite region, recall that the transverse component $\boldsymbol{j}_\perp(\boldsymbol{x};\omega)$ may have tails decaying asymptotically as $O(\frac{1}{r^2})$ outside the compact support of the physical sources.

Green functions

The outgoing solution may be written formally in terms of the retarded scalar Green operator $\mathcal{G}_{\text{ret}}(\omega)$ or its spatial kernel, the Green function $G_{\text{ret}}(\boldsymbol{x};\boldsymbol{x}';\omega) = \langle \boldsymbol{x}\,|\,\mathcal{G}_{\text{ret}}(\omega)\boldsymbol{x}'\rangle$:

$$\boldsymbol{a}_\perp(\boldsymbol{x};\omega) \equiv \langle \boldsymbol{x}\,|\,\boldsymbol{a}_\perp(\omega)\rangle = \tfrac{1}{c}\,\langle \boldsymbol{x}\,|\mathcal{G}_{\text{ret}}(\omega)|\,\boldsymbol{j}_\perp\rangle \equiv \tfrac{1}{c}\int d^3\boldsymbol{x}'\, G_{\text{ret}}(\boldsymbol{x};\boldsymbol{x}';\omega)\boldsymbol{j}_\perp(\boldsymbol{x}';\omega), \tag{4.259}$$

in which we make use of quantum-like Dirac notation for operators and vector fields, wherein

$$\langle \boldsymbol{f}\,|\,\boldsymbol{g}\rangle \equiv \int_{\mathbb{R}^3} d^3\boldsymbol{x}\, \boldsymbol{f}(\boldsymbol{x})^*\cdot\boldsymbol{g}(\boldsymbol{x}) \tag{4.260}$$

is a volume-based inner product for 3-dimensional complex vector fields. Note that any physically reasonable, spatially-localized current densities will be normalizable with respect to this volume-based inner product, as will their solenoidal or irrotational components, but harmonic electromagnetic fields can extend over all space and contain infinite energy, so may not be so normalizable.

We have chosen a scaling such that in free space, the retarded Green function is given by

$$G_{\text{ret}}(\boldsymbol{x};\boldsymbol{x}';\omega) = \tfrac{e^{+ik\|\boldsymbol{x}-\boldsymbol{x}'\|}}{\|\boldsymbol{x}-\boldsymbol{x}'\|}, \tag{4.261}$$

and satisfies

$$\left(\nabla^2 + k^2\right) G_{\text{ret}}(\boldsymbol{x};\boldsymbol{x}';\omega) = -4\pi\delta(\boldsymbol{x}-\boldsymbol{x}'), \tag{4.262}$$

subject to outgoing radiation boundary conditions analogous to (4.258), and where $\delta(\boldsymbol{x})$ is the 3-dimensional Dirac delta function. Note that in free space this Green function is both symmetric and translation invariant. Also note that in equation (4.259), we include only transverse fields generated directly by the actual transverse source $\boldsymbol{j}_\perp$, and ignore any external or background fields, including any fields needed to induce or maintain these currents, or any incoming radiation from other remote sources. Explicit use of the transverse current density as source allows us the convenience of using a scalar rather than dyadic green function.

This same outgoing, transverse solution $\boldsymbol{a}_\perp(\boldsymbol{x};\omega) \equiv \boldsymbol{a}_\perp^{\text{out}}(\boldsymbol{x};\omega)$ can instead be written in terms of the time-reversed, or advanced Green function $G_{\text{adv}}(\boldsymbol{x};\boldsymbol{x}';\omega) = \langle \boldsymbol{x} \left| \mathcal{G}_{\text{adv}}(\omega) \right| \boldsymbol{x}' \rangle$ and a solenoidal homogeneous, or source-free solution:

$$\boldsymbol{a}_\perp(\boldsymbol{x};\omega) = \boldsymbol{a}_\perp^{\text{hom}}(\boldsymbol{x};\omega) + \tfrac{1}{c}\int d^3\boldsymbol{x}'\, G_{\text{ret}}(\boldsymbol{x};\boldsymbol{x}';\omega)\boldsymbol{j}_\perp(\boldsymbol{x}';\omega), \tag{4.263}$$

where $\boldsymbol{a}_\perp^{\text{hom}}(\boldsymbol{x};\omega)$ is an appropriately-chosen solution of (4.256 and (4.257). In free space, the advanced Green operator is just the Hermitian adjoint of of the retarded Green operator, i.e., $\mathcal{G}_{\text{adv}}(\omega) = \mathcal{G}_{\text{adv}}(\omega)^\dagger$, or

$$G_{\text{adv}}(\boldsymbol{x};\boldsymbol{x}';\omega) \equiv \left\langle \boldsymbol{x} \left| \mathcal{G}_{\text{adv}}(\omega) \right| \boldsymbol{x}' \right\rangle = G_{\text{ret}}(\boldsymbol{x}';\boldsymbol{x};\omega)^* = \tfrac{e^{-ik\|\boldsymbol{x}-\boldsymbol{x}'\|}}{\|\boldsymbol{x}-\boldsymbol{x}'\|}, \tag{4.264}$$

and also satisfies (4.262) but with incoming radiation conditions at infinity:

$$\lim_{r\to\infty} r\left(\tfrac{\partial}{\partial r} + ik\right) G_{\text{adv}}(\boldsymbol{x};\boldsymbol{x}';\omega) = 0. \tag{4.265}$$

Subtracting equation (4.263) from (4.259) and rearranging, we find

$$\boldsymbol{a}_\perp^{\text{hom}}(\boldsymbol{x};\omega) = \tfrac{1}{c}\int d^3\boldsymbol{x}'\, D(\boldsymbol{x};\boldsymbol{x}';\omega)\boldsymbol{j}_\perp(\boldsymbol{x}';\omega), \tag{4.266}$$

where we have defined an anti-Hermitian operator $\mathcal{D}(\omega) = -\mathcal{D}(\omega)^\dagger = \mathcal{G}_{\text{ret}}(\omega) - \mathcal{G}_{\text{adv}}(\omega)$ with spatial kernel

$$\left\langle \boldsymbol{x} \left| \mathcal{D}(\omega) \right| \boldsymbol{x}' \right\rangle \equiv D(\boldsymbol{x},\boldsymbol{x}';\omega) \equiv 2i\tfrac{\sin(k\|\boldsymbol{x}-\boldsymbol{x}'\|)}{\|\boldsymbol{x}-\boldsymbol{x}'\|} = 2ik\,\text{sinc}(k\|\boldsymbol{x}-\boldsymbol{x}'\|), \tag{4.267}$$

which is free of any singularities and which generates source-free solutions from auxiliary "sources" in the same way the retarded and advanced Green operators determine outgoing or incoming solutions (with respect to Sommerfeld boundary conditions) from actual sources.

The outgoing part $\boldsymbol{a}_\perp^{\text{out}}(\boldsymbol{x};\omega)$, of $\boldsymbol{a}_\perp^{\text{hom}}(\boldsymbol{x};\omega)$ coincides exactly with the causal potential $\boldsymbol{a}_\perp(\boldsymbol{x};\omega)$, but in addition contains, as we will see, an equal amount of incoming power in the advanced components $\boldsymbol{a}_\perp^{\text{in}}(\boldsymbol{x};\omega)$ needed to cancel any singularities and satisfy the free-space Maxwell equations everywhere, so $\boldsymbol{a}_\perp^{\text{out}}(\boldsymbol{x};\omega)$, will be called the homogeneous or source-free extension or extrapolant of the actual outgoing radiation $\boldsymbol{a}_\perp(\boldsymbol{x};\omega)$.

Back in the time domain, the homogeneous $\boldsymbol{a}_\perp^{\text{hom}}(\boldsymbol{x};t)$, corresponds to the source-free solution describing the transverse fields at sufficiently late times, after the currents have turned off, so the source-free extension is in fact sometimes called the radiation associated with emission from the source $\boldsymbol{j}(\boldsymbol{x},t)$. However, It is important to remember that, if examined over all space and time, the source-free extension contains incoming fields never actually present, and in fact ultimately attributes the physical fields emitted by the source by fields which came in from infinity, or that in effect were emitted by arbitrarily distant sources at arbitrarily remote times in the past. For example, if the far zone fields consist of an expanding spherical wave, then the source-free extension will also contain a time-reversed spherical wave which will converge during the past toward the locations of the actual sources and appear to be reflected into outgoing waves at the precise times when and locations where the actual current sources are turned on. A collimated beam will appear not to be produced by a source but to have come in from infinity on the opposite side of the source.

Although radiation-zone fields are uniquely determined by the sources, such fields do not uniquely determine the sources, because of the well-known existence of so-called non-radiating sources at particular frequencies, of either the strong type (producing strictly vanishing outside the support of the sources producing them) or of the weak variety (yielding fields falling off like $O(\frac{1}{r^2})$ or faster outside the support of the sources). If $\boldsymbol{a}_\perp^{\text{nr}}(\boldsymbol{x};\omega)$ is any such non-radiative vector potential, then $\boldsymbol{j}_\perp^{\text{nr}} = -\frac{c}{4\pi}\left(\nabla^2 + k^2\right)\boldsymbol{a}_\perp^{\text{nr}}$ is the solenoidal part of the non-radiating source producing it.

Non-invertibility arises because, the operator $\mathcal{D}(\omega)$ has a highly nontrivial kernel consisting precisely of these non-radiating sources. Mathematically,this arises because the Sommerfeld boundary conditions differentiate only the radiation-zone fields, both $\mathcal{G}_{\text{ret}}$ and $\mathcal{G}_{\text{ret}}$ produce the same non-radiation fields from a given source $\boldsymbol{j}_\perp$, which then cancels from the difference fields. Mathematically, any homogeneous extension contains no non-radiative

near-zone or intermediate-zone fields anywhere, for there are no actual sources with respect to which the field at any point can be near or far, so sources producing only near-zone and induction-zone fields do not contribute.

These source-free extensions naturally arise as trial functions in variational principle, but in the following, we will assume that any potentials and fields are the physical, outgoing solutions unless explicitly noted otherwise.

4.5.2 Poynting Theorem and Power Relations

As $kr \to \infty$, the asymptotic far-fields become:

$$\boldsymbol{e}_\perp = \quad ik\boldsymbol{a}_\perp \quad \sim \tfrac{ik}{c}\frac{e^{ikr}}{r}\boldsymbol{N}_\perp(\hat{\boldsymbol{r}};\omega), \tag{4.268a}$$

$$\boldsymbol{b}_\perp = -\tfrac{i}{k}\boldsymbol{\nabla}\times\boldsymbol{e}_\perp \sim \hat{\boldsymbol{r}}\times\boldsymbol{e}_\perp, \tag{4.268b}$$

where $\boldsymbol{N}_\perp(\hat{\boldsymbol{r}})$ is the transverse part of the so-called radiation vector, or far-field radiation pattern, which is just the propagative part of the angular spectrum for real wavevectors of magnitude k:

$$\boldsymbol{N}_\perp(\hat{\boldsymbol{r}};\omega) = \int d^3\boldsymbol{x}\ \boldsymbol{j}_\perp(\boldsymbol{x};\omega)e^{-ik\hat{\boldsymbol{r}}} = (1-\hat{\boldsymbol{r}}\hat{\boldsymbol{r}}^{\mathrm{T}})\int d^3\boldsymbol{x}\ \boldsymbol{j}(\boldsymbol{x};\omega)e^{-ik\hat{\boldsymbol{r}}}, \tag{4.269}$$

where in the last step we exploited the fact that in reciprocal (i.e., wavevector, or spatial-Fourier) space, functional transversality and geometric transversality coincide.

Consider $\boldsymbol{e}_{1\perp}(\boldsymbol{x};\omega) = \frac{ik}{c}\langle\boldsymbol{x}\,|\mathcal{G}(\omega)|\,\boldsymbol{j}_{1\perp}\rangle$ and $\boldsymbol{e}_{1\perp}(\boldsymbol{x};\omega) = \frac{ik}{c}\langle\boldsymbol{x}\,|\mathcal{D}(\omega)|\,\boldsymbol{j}_{2\perp}\rangle$, the transverse fields produced by transverse sources $\boldsymbol{j}_{1\perp}(\boldsymbol{x};\omega)$ and $\boldsymbol{j}_{2\perp}(\boldsymbol{x};\omega)$, respectively. Let $B(R)$ denote a ball in $\mathbb{R}^3$, of radius $R>0$ and centered on the origin, and let $V(R)$ denote any connected volume containing $B(R)$ with continuous, closed, piecewise-smooth boundary. Then using Maxwell's equations, standard vector identities, and Gauss's theorem, one finds

$$\tfrac{c}{4\pi}\int\limits_{\partial V(R)} d^2S\,\hat{\boldsymbol{n}}\cdot(\boldsymbol{e}_{1\perp}^*\times\boldsymbol{b}_2) = \tfrac{i\omega}{4\pi}\int\limits_{V(R)} d^3\boldsymbol{x}\ (\boldsymbol{e}_{1\perp}^*\cdot\boldsymbol{e}_{2\perp} - \boldsymbol{b}_1^*\cdot\boldsymbol{b}_1) - \int\limits_{V(R)} d^3\boldsymbol{x}\,\boldsymbol{e}_{1\perp}^*\cdot\boldsymbol{j}_{2\perp}, \tag{4.270}$$

where $\hat{\boldsymbol{n}}$ is an outward normal unit vector field on the surface $\partial V(R)$. This is a version of Poynting's theorem for solenoidal harmonic fields. In particular, if we take $\boldsymbol{j}_1 = \boldsymbol{j}_2 = \boldsymbol{j}$, $V(R) = B(R)$, and consider the real part in the $R\to\infty$ limit, we find

$$-\operatorname{Re}\langle\boldsymbol{e}_\perp\,|\,\boldsymbol{j}_\perp\rangle = \tfrac{c}{4\pi}(\boldsymbol{e}_\perp\,,\,\boldsymbol{b}) = \tfrac{k^2}{4\pi c}\int d^2\Omega(\hat{\boldsymbol{r}})\boldsymbol{N}_\perp(\hat{\boldsymbol{r}};\omega)^*\cdot\boldsymbol{N}_\perp(\hat{\boldsymbol{r}};\omega) \geq 0, \tag{4.271}$$

where $d^2\Omega(\hat{\boldsymbol{r}}) = \sin(\theta)\, d\theta\, d\phi$ is the differential element of solid angle, and where we have introduced the semi-definite far-field "surface" sesquilinear product:

$$(\boldsymbol{e}_{1\perp}\,,\, \boldsymbol{b}_2) = \lim_{R\to\infty} \int_{r=R} d^2S\; \hat{\boldsymbol{r}}\cdot(\boldsymbol{e}^*_{1\perp} \times \mathbf{b_2}) = \tfrac{k^2}{4\pi c}\int d^2\Omega(\hat{\boldsymbol{r}})\boldsymbol{N}_{1\perp}(\hat{\boldsymbol{r}};\omega)^*\cdot\boldsymbol{N}_{2\perp}(\hat{\boldsymbol{r}};\omega), \qquad (4.272)$$

which is positive semi-definite for all fields satisfying outgoing Sommerfeld conditions, but is not strictly positive definite and therefore not a true inner product with respect to physical fields because of the possibility of localized non-radiative fields which do not contribute to the far-field radiated power. However, from the final form it is easy to see that is its a true inner product with respect to the radiation vectors, which can be uniquely determined from either the far-fields or the actual sources. Mathematically, Poynting's theorem relates the volume and surface inner products; physically, it expresses energy conservation, whereby the real part attributes the spectral density of outgoing radiated power to the spectral density of mechanical work done against the source charges by the electric fields in their vicinity, and the imaginary part is associated with "reactive" transfer of energy between the local electric and magnetic fields.

Using (4.256), (4.262), and Green's theorem, this natural correspondence between the outgoing vector potential or fields and their source-free extensions may be established using data consisting only of the value of the outgoing vector potential and its normal derivative on some closed surface:

$$\boldsymbol{a}^{\rm hom}_\perp(\boldsymbol{x}';\omega) = \tfrac{1}{4\pi}\int_{\partial V(R)} d^2S\, \boldsymbol{a}^{\rm out}_\perp(\boldsymbol{x};\omega)\hat{\boldsymbol{n}}\cdot\boldsymbol{\nabla}D(\boldsymbol{x},\boldsymbol{x}';\omega) - D(\boldsymbol{x},\boldsymbol{x}';\omega)\hat{\boldsymbol{n}}\cdot\boldsymbol{\nabla}\boldsymbol{a}^{\rm out}_\perp(\boldsymbol{x};\omega), \qquad (4.273)$$

which is a modification of Kirchoff's diffraction integral, essentially a quantitative expression of Huygen's principle. In particular, considering outgoing data on the spherical boundary surface $\partial B(R)$ in the far-field limit $kR \to \infty$. and after replacing the vector potential and the Green functions in the integrand by their asymptotic far-field forms, (and swapping primed and unprimed coordinates) we find

$$\boldsymbol{a}^{\rm hom}_\perp(\boldsymbol{x};\omega) = \tfrac{ik}{c}\tfrac{1}{2\pi}\int d^2\Omega(\hat{\boldsymbol{r}}')\, e^{ik\hat{\boldsymbol{r}}'\cdot\boldsymbol{x}}\boldsymbol{N}_\perp(\hat{\boldsymbol{r}}';\omega), \qquad (4.274)$$

where, again, $\boldsymbol{N}_\perp$ is the transverse radiation vector associated with the outgoing component of $\boldsymbol{a}^{\rm hom}_\perp$. Assuming only that the radiation vector is sufficiently well-behaved so that angular integration and spatial differentiation may be commuted, it is straightforward to verify that this indeed is a solenoidal solution to the homogeneous Helmholtz equation.

To unambiguously decompose an arbitrary source-free solution into outgoing and incoming components at finite spatial locations, some convention for the behavior of the

non-radiative local fields must be adopted. Obviously, purely outgoing or purely incoming solutions cannot satisfy the source-free Helmholtz equation everywhere in space, but, if desired, and if sufficiently generalized effective currents (involving delta functions and derivatives of arbitrarily high order) are allowed, then the effective sources/sinks where the homogenous equation is not satisfied can be confined to a closed surface, or in the limiting case even a single point singularity, say at the origin. Such a "minimal" decomposition can be obtained by the well-known multipole expansion of plane waves as

$$e^{ik\hat{\boldsymbol{r}}'\cdot\boldsymbol{x}} = \sum_{\ell=0}^{\infty} i^\ell (2\ell+1) j_\ell(kr) P_\ell(\hat{\boldsymbol{r}}'\cdot\hat{\boldsymbol{r}}), \tag{4.275}$$

where $j_\ell(s)$ is the spherical Bessel function of order ℓ, and $P_\ell(\gamma)$ is the Legendre polynomial of degree ℓ. Writing $j_\ell(s) = \frac{1}{2}h_\ell^{(1)}(s) + \frac{1}{2}h_\ell^{(2)}(s)$, where $h_\ell^{(1)}(s)$ is a spherical Hankel function of the first kind, satisfying outgoing radiation boundary conditions, and $h^{(2)}(s) = h^{(1)}(s)^*$ is the spherical Hankel function of the second kind, satisfying incoming radiation conditions, then one obtains for the minimal decomposition:

$$\boldsymbol{a}_\perp^{\text{out}}(\boldsymbol{x};\omega) = \tfrac{ik}{c}\tfrac{1}{4\pi} \sum_{\ell=0}^{\infty} i^\ell \, (2\ell+1) \, h_\ell^{(1)}(kr') \int d^2\Omega(\hat{\boldsymbol{r}}') \, P_\ell(\hat{\boldsymbol{r}}'\cdot\hat{\boldsymbol{r}}), \tag{4.276a}$$

$$\boldsymbol{a}_\perp^{\text{in}}(\boldsymbol{x};\omega) = \tfrac{ik}{c}\tfrac{1}{4\pi} \sum_{\ell=0}^{\infty} i^\ell \, (2\ell+1) \, h_\ell^{(2)}(kr') \int d^2\Omega(\hat{\boldsymbol{r}}') \, P_\ell(\hat{\boldsymbol{r}}'\cdot\hat{\boldsymbol{r}}). \tag{4.276b}$$

We will actually only make use of the incoming and outgoing components of source-free solutions in the far-field, which can be unambiguously determined. In the $kr \to \infty$ limit, (4.275) can be expanded using the asymptotic forms for the spherical Bessel functions and the spherical harmonics, with the result that we recover the outgoing field exactly, as well as an incoming field of equal power spectral density but with opposite angular pattern:

$$\boldsymbol{e}_\perp^{\text{out}}(\boldsymbol{x};\omega) \xrightarrow[kr\to\infty]{} \quad \tfrac{ik}{c}\frac{e^{+ikr}}{r}\boldsymbol{N}_\perp(\hat{\boldsymbol{r}};\omega), \tag{4.277a}$$

$$\boldsymbol{b}_\perp^{\text{out}}(\boldsymbol{x};\omega) \xrightarrow[kr\to\infty]{} \quad \tfrac{ik}{c}\frac{e^{+ikr}}{r}\hat{\boldsymbol{r}} \times \boldsymbol{N}_\perp(\hat{\boldsymbol{r}};\omega), \tag{4.277b}$$

$$\boldsymbol{e}_\perp^{\text{in}}(\boldsymbol{x};\omega) \xrightarrow[kr\to\infty]{} -\tfrac{ik}{c}\frac{e^{-ikr}}{r}\boldsymbol{N}_\perp(-\hat{\boldsymbol{r}};\omega), \tag{4.277c}$$

$$\boldsymbol{b}_\perp^{\text{in}}(\boldsymbol{x};\omega) \xrightarrow[kr\to\infty]{} \quad \tfrac{ik}{c}\frac{e^{-ikr}}{r}\hat{\boldsymbol{r}} \times \boldsymbol{N}_\perp(-\hat{\boldsymbol{r}};\omega). \tag{4.277d}$$

In practice, the radiation vector $\boldsymbol{N}_\perp(\hat{\boldsymbol{r}};\omega)$ may not be available (if it were known exactly, then we might not need the variational principle in the first place), but in the far field, the outgoing and incoming components can be extracted more simply from the source-free extension by asymptotic projection into the subspaces satisfying the corresponding

Sommerfeld boundary conditions. That is, as $kr \to \infty$, we may take

$$\boldsymbol{e}_\perp^{\text{out}} \sim \tfrac{1}{2}(ik + \tfrac{\partial}{\partial r})\boldsymbol{a}_\perp^{\text{hom}}, \tag{4.278a}$$

$$\boldsymbol{b}_\perp^{\text{out}} \sim \hat{\boldsymbol{r}} \times \boldsymbol{e}_\perp^{\text{out}}, \tag{4.278b}$$

$$\boldsymbol{e}_\perp^{\text{in}} \sim \tfrac{1}{2}(ik - \tfrac{\partial}{\partial r})\boldsymbol{a}_\perp^{\text{hom}}, \tag{4.278c}$$

$$\boldsymbol{b}_\perp^{\text{out}} \sim \hat{\boldsymbol{r}} \times \boldsymbol{e}_\perp^{\text{in}}. \tag{4.278d}$$

These are accurate to $O(\frac{1}{r^2})$, which is sufficient to calculate the radiated outgoing or incoming power.

Using (4.277), it is easily shown that in any source-free solution the incoming and outgoing spectral densities of radiated power are equal in magnitude but of course opposite in sign:

$$\tfrac{c}{4\pi}\left(\boldsymbol{e}_{1\perp}^{\text{out}}\,,\,\boldsymbol{b}_2^{\text{out}}\right) = -\tfrac{c}{4\pi}\left(\boldsymbol{e}_{1\perp}^{\text{in}}\,,\,\boldsymbol{b}_2^{\text{in}}\right) = \tfrac{k^2}{4\pi c}\int d^2\Omega(\hat{\boldsymbol{r}})\,\boldsymbol{N}_{1\perp}(\hat{\boldsymbol{r}};\omega)^*\cdot\boldsymbol{N}_{2\perp}(\hat{\boldsymbol{r}};\omega) \geq 0, \tag{4.279}$$

while the net directed power in a source-free mode cancels:

$$\tfrac{c}{4\pi}\,\text{Re}\left(\boldsymbol{e}_\perp^{\text{hom}}\,,\,\boldsymbol{b}^{\text{hom}}\right) = \tfrac{c}{4\pi}\left[\left(\boldsymbol{e}_\perp^{\text{out}}\,,\,\boldsymbol{b}^{\text{out}}\right) + \left(\boldsymbol{e}_\perp^{\text{in}}\,,\,\boldsymbol{b}^{\text{in}}\right)\right] + \tfrac{c}{4\pi}\,\text{Re}\left[\left(\boldsymbol{e}_\perp^{\text{out}}\,,\,\boldsymbol{b}^{\text{in}}\right) + \left(\boldsymbol{e}_\perp^{\text{in}}\,,\,\boldsymbol{b}^{\text{out}}\right)\right] = 0 \tag{4.280}$$

Now, using the representation (4.274) and the definition (4.269) for the radiation vector, one easily obtains

$$-\left\langle \boldsymbol{e}_{1\perp}^{\text{hom}} \,\middle|\, \boldsymbol{j}_{2\perp}\right\rangle = \tfrac{k^2}{2\pi c}\int d^2\Omega(\hat{\boldsymbol{r}})\,\boldsymbol{N}_{1\perp}(\hat{\boldsymbol{r}};\omega)^*\cdot\boldsymbol{N}_{2\perp}(\hat{\boldsymbol{r}};\omega) = 2\left(\boldsymbol{e}_{1\perp}^{\text{out}}\,,\,\boldsymbol{b}_2^{\text{out}}\right). \tag{4.281}$$

In fact, using equation (4.281), the orthogonality of solenoidal and irrotational fields, and the fact that $\left\langle \boldsymbol{e}_\parallel \,\middle|\, \boldsymbol{j}_\parallel\right\rangle = -\frac{4\pi i}{\omega}\left\langle \boldsymbol{j}_\parallel \,\middle|\, \boldsymbol{j}_\parallel\right\rangle$ is purely imaginary, we obtain the equivalence between all the following expressions for the spectral density of mechanical power:

$$-\,\text{Re}\left\langle \boldsymbol{e} \,|\, \boldsymbol{j}\right\rangle = -\,\text{Re}\left\langle \boldsymbol{e}_\perp \,|\, \boldsymbol{j}\right\rangle = -\,\text{Re}\left\langle \boldsymbol{e} \,|\, \boldsymbol{j}_\perp\right\rangle = -\,\text{Re}\left\langle \boldsymbol{e}_\perp \,|\, \boldsymbol{j}_\perp\right\rangle = -\tfrac{1}{2}\left\langle \boldsymbol{e}_\perp^{\text{hom}} \,|\, \boldsymbol{j}_\perp\right\rangle, \tag{4.282}$$

where the factor of $\frac{1}{2}$ appearing in the last expression arises to avoid over-counting in the energetics: homogeneous fields must contain equal amounts of outgoing and incoming power (spectral density) in complementary angular patterns, so the "virtual" energy exchange between the actual sources and homogenous trial fields turns out to be exactly twice that between that the same sources and the corresponding outgoing component alone:

$$-\,\text{Re}\left\langle \boldsymbol{e}_\perp \,|\, \boldsymbol{j}_\perp\right\rangle = \left(\boldsymbol{e}_\perp^{\text{out}}\,,\,\boldsymbol{b}_\perp^{\text{out}}\right) = -\tfrac{1}{2}\left\langle \boldsymbol{e}_\perp^{\text{hom}} \,|\, \boldsymbol{j}_\perp\right\rangle = \tfrac{1}{2}\left(\boldsymbol{e}_\perp^{\text{out}}\,,\,\boldsymbol{b}_\perp^{\text{out}}\right) - \tfrac{1}{2}\left(\boldsymbol{e}_\perp^{\text{in}}\,,\,\boldsymbol{b}_\perp^{\text{in}}\right) \geq 0 \tag{4.283}$$

This factor is a radiative analog of the well-known factor of $\frac{1}{2}$ that appears in the expression for the potential self-energy of a charge distribution in electrostatics. This may also be

written as

$$-\tfrac{1}{2}\left\langle \boldsymbol{e}_\perp^{\mathrm{hom}} \,|\, \boldsymbol{j}_\perp \right\rangle = (\boldsymbol{e}_\perp^{\mathrm{out}}\,,\, \boldsymbol{b}_\perp^{\mathrm{out}}) = \left(\tfrac{1}{2}(1-\tfrac{i}{k}\tfrac{\partial}{\partial r})\boldsymbol{e}_\perp^{\mathrm{hom}}\,,\, \tfrac{1}{2}(1-\tfrac{i}{k}\tfrac{\partial}{\partial r})\boldsymbol{b}_\perp^{\mathrm{hom}}\right), \tag{4.284}$$

which will be the most useful form for calculation.

4.5.3 Maximum Power Variational Principle

With all the preliminary mathematical machinery in place, derivation of the actual variational principle is straightforward. By linearity, we may consider the radiation emitted by the difference $(\boldsymbol{j}_\perp - \boldsymbol{s}_\perp)$ between the actual solenoidal source of interest and any auxiliary solenoidal source. Were it present, the outgoing power (spectral density) that would be radiated by this difference source would necessarily be non-negative:

$$-\operatorname{Re}\left\langle (\boldsymbol{e}_\perp - \boldsymbol{\chi}_\perp^{\mathrm{out}}) \,|\, (\boldsymbol{j}_\perp - \boldsymbol{s}_\perp) \right\rangle \geq 0 \tag{4.285}$$

where the outgoing component of the trial electric field is given by $\boldsymbol{\chi}_\perp^{\mathrm{out}}(\boldsymbol{x};\omega) = \left\langle \boldsymbol{x} \,\middle|\, \frac{ik}{c}\mathcal{G}_{\mathrm{ret}}(\omega)\boldsymbol{s}_\perp(\omega) \right\rangle$. Equality will hold in (4.285) if and only if the true transverse source $\boldsymbol{j}_\perp$ radiating the actual fields and the auxiliary transverse source $\boldsymbol{s}_\perp$ generating the trial fields differ by some non-radiating source, i.e., if and only if the difference $\boldsymbol{j}_\perp - \boldsymbol{s}_\perp$ lies in the kernel of $\mathcal{D}(\omega)$ and so has a vanishing transverse radiation vector, or, equivalently, if and only if $\boldsymbol{\chi}_\perp^{\mathrm{out}}(\boldsymbol{x};\omega) = \boldsymbol{e}_\perp(\boldsymbol{x};\omega)$. Using linearity and re-arranging, after certain substitutions based on the Green function and power relations developed above, this inequality becomes:

$$\tfrac{c}{4\pi}\left(\boldsymbol{e}_\perp\,,\,\boldsymbol{b}\right) \geq \left\langle \tfrac{ik}{c}\mathcal{G}_{\mathrm{ret}}(\omega)\boldsymbol{s}_\perp \,\middle|\, \boldsymbol{s}_\perp \right\rangle - \operatorname{Re}\left[\left\langle \tfrac{ik}{c}\mathcal{G}_{\mathrm{ret}}(\omega)\boldsymbol{s}_\perp \,\middle|\, \boldsymbol{j}_\perp \right\rangle + \left\langle \tfrac{ik}{c}\mathcal{G}_{\mathrm{ret}}(\omega)\boldsymbol{j}_\perp \,\middle|\, \boldsymbol{s}_\perp \right\rangle\right]. \tag{4.286}$$

Because the trial fields rather than trial source will be of interest, we seek to re-express this inequality entirely in terms of the former. Using the relations (4.283 -4.284), the first term on the right-hand side can be written in terms of the source-free trial field as

$$\begin{aligned}\left\langle \tfrac{ik}{c}\mathcal{G}_{\mathrm{ret}}(\omega)\boldsymbol{s}_\perp \,\middle|\, \boldsymbol{s}_\perp \right\rangle = \left\langle \boldsymbol{\chi}_\perp^{\mathrm{out}} \,|\, \boldsymbol{s}_\perp \right\rangle &= \tfrac{1}{2}\left\langle \boldsymbol{\chi}_\perp^{\mathrm{hom}} \,|\, \boldsymbol{s}_\perp \right\rangle = -\tfrac{c}{4\pi}\left(\boldsymbol{\chi}_\perp^{\mathrm{out}}\,,\, -\tfrac{i}{k}\boldsymbol{\nabla}\times\boldsymbol{\chi}_\perp^{\mathrm{out}}\right) \\ &= \tfrac{c}{16\pi}\tfrac{1}{k^2}\left((1-\tfrac{i}{k}\tfrac{\partial}{\partial r})\boldsymbol{\chi}_\perp^{\mathrm{hom}}\,,\, ik(1-\tfrac{i}{k}\tfrac{\partial}{\partial r})\boldsymbol{\nabla}\times\boldsymbol{\chi}_\perp^{\mathrm{hom}}\right) \leq 0.\end{aligned} \tag{4.287}$$

Writing the real part of the term in brackets as the average of itself and its complex conjugate, and using the symmetry of the Green function, we find after a little algebra that it too can be expressed in terms of the source-free field:

$$\begin{aligned}\operatorname{Re}\left[\left\langle \tfrac{ik}{c}\mathcal{G}_{\mathrm{ret}}(\omega)\boldsymbol{s}_\perp \,\middle|\, \boldsymbol{j}_\perp \right\rangle + \left\langle \tfrac{ik}{c}\mathcal{G}_{\mathrm{ret}}(\omega)\boldsymbol{j}_\perp \,\middle|\, \boldsymbol{s}_\perp \right\rangle\right] &= \operatorname{Re}\left\langle \boldsymbol{s}_\perp \,\middle|\, \tfrac{ik}{c}\mathcal{D}(\omega)\middle|\, \boldsymbol{j}_\perp \right\rangle \\ &= \operatorname{Re}\left\langle \boldsymbol{\chi}_\perp^{\mathrm{hom}} \,|\, \boldsymbol{j}_\perp \right\rangle,\end{aligned} \tag{4.288}$$

and the inequality (4.286) may be re-written as

$$\tfrac{c}{4\pi}\,(\boldsymbol{e}_\perp\,,\,\boldsymbol{b}) \geq -\tfrac{c}{4\pi}\left(\boldsymbol{\chi}_\perp^{\text{out}}\,,\,-\tfrac{i}{k}\boldsymbol{\nabla}\times\boldsymbol{\chi}_\perp^{\text{out}}\right) - \operatorname{Re}\langle\boldsymbol{\chi}_\perp^{\text{hom}}\,|\,\boldsymbol{j}_\perp\rangle\,, \tag{4.289}$$

where the first term on the right-hand side, being the additive inverse of the outgoing power spectral density in the trial fields, is in fact non-positive, and the second term, representing the "virtual" work that would be extracted from the actual sources by the source-free field, if present, can be of either sign.

Now suppose the homogeneous trial electric field actually depends on some vector $\boldsymbol{\alpha}$ of parameters determining the field phase, amplitude, shape, and polarization. We do not require that the parameterized family $\left\{\boldsymbol{\chi}_\perp^{\text{hom}}(\boldsymbol{x};\omega;\boldsymbol{\alpha})\right\}$ of trial solutions constitutes a linear subspace (i.e., the set of possibilities is not necessarily closed under all linear combinations), but we do impose the essentially trivial requirement that it allows for arbitrary complex re-scalings (i.e., the set of possibilities is closed under multiplication by any scalar $\xi \in \mathbb{C}$.)

Since $k > 0$, it is easy to verify that a trial solution $\boldsymbol{\chi}_\perp^{\text{hom}}$ is feasible (satisfying the source-free Helmholtz equation and the solenoidal gauge constraint) if and only if it is an eigenvector of the double-curl operator with eigenvalue k^2, i.e., $\boldsymbol{\nabla}\times\boldsymbol{\nabla}\times\boldsymbol{\chi}_\perp^{\text{hom}} = k^2\boldsymbol{\chi}_\perp^{\text{hom}}$. Because the inequality (4.286) must hold true for all possible parameterized trial fields, it may then be written as

$$\tfrac{c}{4\pi}\,(\boldsymbol{e}_\perp\,,\,\boldsymbol{b}) \geq \max_{\boldsymbol{\alpha}}\left[-\tfrac{c}{4\pi}\left(\boldsymbol{\chi}_\perp^{\text{out}}\,,\,-\tfrac{i}{k}\boldsymbol{\nabla}\times\boldsymbol{\chi}_\perp^{\text{out}}\right) - \operatorname{Re}\langle\boldsymbol{\chi}_\perp^{\text{hom}}\,|\,\boldsymbol{j}_\perp\rangle\,,\right] \tag{4.290a}$$

$$\text{such that: } \boldsymbol{\nabla}\times\boldsymbol{\nabla}\times\boldsymbol{\chi}_\perp^{\text{hom}}(\boldsymbol{x};\omega;\boldsymbol{\alpha}) = k^2\boldsymbol{\chi}_\perp^{\text{hom}}(\boldsymbol{x};\omega;\boldsymbol{\alpha}), \tag{4.290b}$$

where equality is achieved if and only if $\boldsymbol{\chi}_\perp^{\text{out}}(\boldsymbol{x};\omega;\boldsymbol{\alpha}) = \boldsymbol{e}_\perp(\boldsymbol{x};\omega)$, that is, if and only if the far-field outgoing part of the source-free field coincides with the true far-field pattern. (Recall that the trial homogenous fields can be unambiguously and deterministically decomposed into incoming and outgoing components in the far-field limit, where they are needed to determine the radiated power; the outgoing power is written here in terms $\boldsymbol{\chi}_\perp^{\text{out}}$ for compactness and clarity.)

This is actually our desired result, although still in somewhat cryptic form. In fact, under our assumptions it is equivalent to the more physically-transparent constrained maximization:

$$\tfrac{c}{4\pi}\,(\boldsymbol{e}_\perp\,,\,\boldsymbol{b}) \geq \max_{\boldsymbol{\alpha}}\left[\tfrac{c}{4\pi}\left(\boldsymbol{\chi}_\perp^{\text{out}}\,,\,-\tfrac{i}{k}\boldsymbol{\nabla}\times\boldsymbol{\chi}_\perp^{\text{out}}\right)\right] \tag{4.291a}$$

$$\text{such that: } \tfrac{c}{4\pi}\left(\boldsymbol{\chi}_\perp^{\text{out}}\,,\,-\tfrac{i}{k}\boldsymbol{\nabla}\times\boldsymbol{\chi}_\perp^{\text{out}}\right) = -\tfrac{1}{2}\operatorname{Re}\langle\boldsymbol{\chi}_\perp^{\text{hom}}\,|\,\boldsymbol{j}_\perp\rangle\,, \tag{4.291b}$$

$$\text{and: } \boldsymbol{\nabla}\times\boldsymbol{\nabla}\times\boldsymbol{\chi}_\perp^{\text{hom}} = k^2\boldsymbol{\chi}_\perp^{\text{hom}}. \tag{4.291c}$$

That is, the actual fields radiate more outgoing power, and therefore extract more energy from the actual sources, than could any homogenous trial field, if it were actually present in the vicinity of the sources.

To see the equivalence under our assumptions of these seemingly different optimization problems, we suppose the variational parameters are $\boldsymbol{\alpha} = (\mu, \delta, \boldsymbol{\alpha}')$, and write

$$\boldsymbol{\chi}_{\perp}^{\mathrm{hom}}(\boldsymbol{x};\omega;\boldsymbol{\alpha}) = \mu e^{i\delta}\boldsymbol{u}_{\perp}^{\mathrm{hom}}(\boldsymbol{x};\omega;\boldsymbol{\alpha}'), \tag{4.292}$$

where μ is an overall (dimensionless) positive scaling, satisfying $0 < \mu < \infty$; δ is an overall real phase, such that $0 \leq \delta < 2\pi$; and $\boldsymbol{u}_{\perp}^{\mathrm{hom}}$ represents the overall spatial shape and polarization of the trial vector field, parameterized by the remaining set of parameters $\boldsymbol{a}'$, and which of course is still assumed to be a solenoidal solution to the source-free Helmholtz equation.

Then after a some manipulation, one finds that the formal solutions to the variational optimization problems defined either by (4.290) or by (4.291) are equal, and both given by:

$$\tilde{\boldsymbol{\alpha}}' = \arg\max_{\boldsymbol{\alpha}'} \left[\frac{k}{c} \frac{|\langle \boldsymbol{u}_{\perp}^{\mathrm{hom}} \,|\, \boldsymbol{j}_{\perp} \rangle|^2}{\left| \left(\boldsymbol{u}_{\perp}^{\mathrm{out}} \,,\, \boldsymbol{\nabla} \times \boldsymbol{u}_{\perp}^{\mathrm{out}} \right) \right|} \right], \tag{4.293a}$$

$$\tilde{\delta} = \arg\left[- \langle \boldsymbol{u}_{\perp}^{\mathrm{hom}} \,|\, \boldsymbol{j}_{\perp} \rangle|_{\boldsymbol{\alpha}'=\tilde{\boldsymbol{\alpha}}'} \right], \tag{4.293b}$$

$$\tilde{\mu} = \tfrac{2\pi}{c} \left| \frac{\langle \boldsymbol{u}_{\perp}^{\mathrm{hom}} \,|\, \boldsymbol{j}_{\perp} \rangle}{\left(\boldsymbol{u}_{\perp}^{\mathrm{out}} \,,\, \boldsymbol{\nabla} \times \boldsymbol{u}_{\perp}^{\mathrm{out}} \right)} \right|_{\boldsymbol{\alpha}'=\tilde{\boldsymbol{\alpha}}'}, \tag{4.293c}$$

so in fact they describe the same maximum power variational principle.

4.6 Discussion and Interpretations of the MPVP

Here we assess advantages and limitations of the MPVP, briefly consider possible variations and extensions, and compare the method to other variational techniques commonly encountered in electromagnetic theory, in particular those which have been used for paraxial radiation problems, laser-plasma interactions, and FEL-type devices.

4.6.1 Summary

The derived MPVP (4.240), or equivalently (4.291) in physical units, is valid in the paraxial or non-paraxial regimes, says, in effect, that classical (accelerating) charges spontaneously radiate "as much as possible," consistent with energy conservation. By spontaneous

we mean that charges are assumed to follow *prescribed* classical trajectories determined by the external fields and possibly (averaged) quasi-static self-fields, but are not affected by recoil/radiation-reaction, multiple scattering, absorption, or other feedback from the emitted radiation (although arbitrary pre-bunching due to previously-imposed or radiated fields may be included, if it can be characterized). The sources are assumed to be localized in space, so that the far-field (i.e., radiation zone , or wave zone) is defined, and at least weakly localized in time, so that Fourier transforms (i.e., frequency-domain representations) of the current density exist. In turn, the radiation, once emitted by any part of the source, propagates as in free space, without further scattering or absorption.(In the paraxial case, "outgoing" implies propagation nearly along the optic axis away from the localized sources, defining the "downstream" or "post-source" direction, while "incoming" means propagating towards the sources from the upstream direction. In this case, say, a right-going pulse has both ingoing and outgoing components. In the general non-paraxial case, "outgoing" implies wavefronts which, at least asymptotically, diverge in time and propagate away from the localized sources, while "incoming" denotes the reverse, i.e., asymptotically converging wavefronts propagating towards the sources.)

The optimized trial mode-shape (or more accurately, the outgoing component of this source-free mode) becomes the best guess for the actual radiation fields in the extra-source region, within the manifold of possibilities defined by the parameterized family of trial solutions. The approximation enjoys the usual benefits and suffers the usual drawbacks of other variational principles. The optimized outgoing Poynting flux provides a true *lower bound* for the actual spontaneously-radiated power spectral density at the frequency under consideration (apart from any numerical/roundoff errors in the computation), and the accuracy of the estimated power, as well as the field profile or any other physical observables derived from them, should improve monotonically as additional functionally-independent parameters are included in the variational fit to allow for more general envelope shapes. No upper bound on radiated power spectral density is presently known.

As with other variational/trial function approaches, the relative error in the value of the variational functional (here, the power) at the stationary point is generally smaller than the error in the parameterized trial function (here, the radiation profile), i.e., second order in "shape" errors, but conversely the field values or field profile and polarization are then approximated with comparatively less accuracy than is the power.

Note that although the variational principle arises in a linear space setting , the adjustable parameters may actually appear linearly (e.g., as expansion coefficients in some

basis-set decomposition, such as Gauss-Hermite modes in the paraxial case) or appear non-linearly (e.g., as a spot size or waist location in an adjustable Gaussian mode) in the trial-functions. If they are taken to be the linear expansion coefficients in a sum over orthogonal modes, then the MPVP is equivalent to a truncated basis-set approach.

This variational principle can be variously interpreted according to one's tastes or application. As we have seen, the best variational approximation maximizes the radiated power consistent with the constraint that this energy could have arisen from work extracted from the actual sources.

The variational approximation also minimizes a Hilbert-space distance between the actual fields and the parameterized family of solenoidal, free-space radiation fields, and in many cases the optimal solution may actually be regarded as an orthogonal projection of the true radiation field into this manifold of trial solutions.

It also maximizes, for each frequency component, the spatial overlap, or physical resemblance, between the actual current density and the trial fields, when the latter are extrapolated back into the region of the sources assuming source-free propagation. Because the rate of energy exchange between the electric field $\boldsymbol{\varepsilon}_\perp$ and the charges associated with the current distribution $\boldsymbol{j}$ takes the form of an overlap integral $\int d^3\boldsymbol{\zeta}\, \boldsymbol{\varepsilon}_\perp{}^* \cdot \boldsymbol{j}$ between these quantities, an obvious "folk theorem" has suggested itself to many authors, wherein currents should in some sense "look like" the fields they produce. The MPVP is in fact a precise and quantitative version of this often vaguely formulated folk theorem.

Equivalently, one can say the optimal radiation field profile is that which, if it actually were to be incident on the sources, would maximally couple to the given sources and would experience maximal small-signal gain (i.e., neglecting saturation effects or indeed any appreciable back-action on the sources) due to energy absorbed from those sources, and furthermore the "virtual" gain delivered would be equal to the estimated power spontaneously radiated.

Whatever the shape of the trial function, and whether the variational parameters appear linearly or non-linearly, after optimization this profile may be regarded as one member of a complete set of solenoidal basis functions solving the source-free wave equation. After introducing an inner product associated with the radiated power (not energy), these basis functions can be orthonormalized via a Gram-Schmidt procedure, leaving the original trial function unchanged. Bessel's inequality then confirms that the power radiated in (the

outgoing component of) this one mode can be no greater than the total outgoing power radiated in all the modes.

All this is obvious, and makes the MPVP sound almost too trivial to be useful. What is perhaps more surprising (if we do not think too hard about the Fundamental Theorem of Calculus or its multidimensional generalizations) is that we can exactly relate the outgoing power radiated in the far-field to the "virtual" work which would be performed by the localized sources on the "source-free extrapolant" of this radiation field, as if it and it alone were present in the region of the source, i.e., relate the radiated power to the overlap integral between the actual source and the extrapolated electric field, even though the source-free extrapolant may not closely resemble the actual emitted fields and near fields in that region, as those must necessarily be described by the inhomogeneous Maxwell equations.

The factor of $\frac{1}{2}$ in the resulting conservation constraint emerges in order to avoid double-counting in the energetics. The radiated power in the outgoing component of the variational approximation is being related to the mechanical power which would be delivered by the sources to the full source-free fields, if they were actually present in the region of the sources, rather than the actual, inhomogeneous fields; and, any wave-field satisfying the source-free wave equation everywhere in space must have the power in each outgoing mode exactly balanced by that in a corresponding incoming mode, or else singularities would arise somewhere in space. (Without any sources, in the paraxial case, radiation observed to emerge from, say, the right side of any longitudinal interval must have previously traveled into that interval from the left. In the free-space non-paraxial case, analogous conclusions can be drawn for plane waves traveling along any chosen direction, or, thinking in terms of multipole radiation, it is necessary that any diverging spherical wave must be matched to a converging spherical wave in the past to avoid a singularity at their common center of wavefront curvature.)

4.6.2 Assessment

Again, the present approach involves approximation of the actual radiation fields by trial modes which are solenoidal and satisfy the homogeneous (source-free) wave equation. As we are primarily interested in the net radiation from the given sources observed in the free-space region beyond the sources, it is appropriate to use solutions of the source-free equation there, but then it is natural, efficient, and even almost inevitable, in the context of an approximation scheme, to effectively extrapolate the fields back into the support of the

sources by free-space propagation. (Otherwise, if we could somehow propagate according to the actual dynamics, we would not need to resort to any approximation – but we would also be carrying around unneeded information about the near-fields.) In the paraxial limit, these free-space solutions are uniquely specified by the carrier frequency and the (complex) profile in any transverse plane, which can be decomposed into a convenient, countable set of modes consisting of the eigenmodes of some quantum-like operator(s). In the general case the solutions are more complicated, but from the spherical wave expansion, we know at least that they also form a separable Hilbert space, parameterized by the multipole expansion coefficients. Because there is no known closed-form, analytic solution for a general focused radiation beam outside the paraxial approximation, the formalism and variational principle are probably of most use in the paraxial case, although may also be applied, with little added complication, to cases where the radiation consists of a superposition of multiple beams, each paraxial individually, but with sufficiently divergent directions to violate overall paraxiality, or to cases where higher-order terms in the generalized paraxial (i.e., diffraction-angle) expansion are included to describe a moderately-collimated beam.

The output $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega;\tilde{\boldsymbol{\alpha}})$ of the MPVP is therefore actually better considered a variational approximation to the source-free "extension," "embedding" or "extrapolation" $\boldsymbol{\chi_j}(\boldsymbol{\zeta};\omega)$ rather than to the actual outgoing vector potential $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$. If the radiation is sufficiently directional, then the outgoing component $\boldsymbol{\chi}^{\text{out}}(\boldsymbol{\zeta};\omega;\tilde{\boldsymbol{\alpha}})$ of the trial function may be determined simply by restricting attention to some finite solid angle or finite range of wavevectors. In the most general case, where the "spurious" incoming radiation may overlap spatially with the "physical" outgoing radiation, at least in the frequency domain, it is more difficult to disentangle $\boldsymbol{\chi}^{\text{out}}$ and $\boldsymbol{\chi}^{\text{in}}$ at an arbitrary position. If the spherical wave expansion is known, then the outgoing projection may be easily determined everywhere in space, but, generically, direct use of the spherical-waves as the variational basis is not expected to prove convenient. (This would correspond to just using a truncated multipole expansion for the radiation.) In the far-field limit $k\zeta \to \infty$, at least, the outgoing component may be determined easily by using the Sommerfeld projections (4.214). In the time domain, the unwanted advanced component of the radiation will pass any particular remote point before it is "absorbed" by the source (or rather, appears to "bounce" off the source), so if the source could be assumed to be of finite extent in both time and space, then at points sufficiently far from the support of the source, the retarded component of the radiation could be unambiguously separated from the advanced component simply by suitably delaying the onset time of observation. However, this method cannot necessarily distinguish

incoming from outgoing radiation at points within or close to the source, and, in practice, will encounter other ambiguities; it is desirable to allow for analytic sources which are only weakly localized in time (i.e., rapidly decaying outside of some nominal time interval, but non-zero), and, as we have discussed, the transverse current density $\boldsymbol{j}_\perp(\boldsymbol{\zeta},\tau)$ is in general rapidly decaying, but not strictly vanishing, even when the full current density $\boldsymbol{j}(\boldsymbol{\zeta},\tau)$ is of compact support, so the distinction between advanced and retarded times at some finite distance from an extended solenoidal source may not always be clear.

Because of these difficulties, and remaining mindful of the proportionality between the calculated power delivered from the source $\boldsymbol{j}_\perp(\boldsymbol{\zeta};\omega)$ to the radiation fields corresponding to the vector potential $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$ or to the fields of its source-free approximant $\boldsymbol{\chi}(\boldsymbol{\zeta},\omega)$, one might be tempted to apply the variational method directly (just without the extra factor of $\frac{1}{2}$) to outgoing, solenoidal solutions of some convenient inhomogeneous or homogeneous Helmholtz equation. However, purely outgoing-wave, free-space solutions cannot exist everywhere in space, and will inevitably possess one or more singularities in the vicinity of the sources, preventing calculation of the required mechanical work integrals. Furthermore, any solenoidal vector field whatsoever, say $\boldsymbol{\psi}(\boldsymbol{\zeta};\omega)$, is trivially the solution to some inhomogeneous Helmholtz equation for some boundary conditions and some solenoidal source, namely, for those boundary conditions satisfied by $\boldsymbol{\psi}$ itself, and for a solenoidal source given by $-(\partial^2+\omega^2)\boldsymbol{\psi}$. Without further constraints, this class of trial solutions is simply too general to be useful. If we naively apply the MPVP using homogeneous trial functions, the method will converge not to $\boldsymbol{a}$ but to something proportional to $\boldsymbol{j}$, since, for given norm, no other vector field can have greater overlap with $\boldsymbol{j}$ than a multiple of $\boldsymbol{j}$ itself. By using the source-free solutions, we effectively regularize the problem, so that we can converge to something which maximizes the overlap with the source under the assumption that the trial look like a radiation field, and not a current density.

In any case, the method requires not only the parameterization of trial modes but the calculation of inner products between these trial modes and the current density, or else some other, problem-specific, means to determine the absolute level of the power, either analytically, in terms of the free parameters, or numerically and indeed repeatedly, as the parameters are varied in search of the maximum. One can easily imagine counter-examples where it is simply easier to work directly with the Liénard-Wichart potentials or corresponding expressions for the fields, than use the variational principle.

We also have some options (or rather, trade-offs) in handling stochastic sources. So far, we have assumed a specific, prescribed, deterministic current density $\boldsymbol{j}(\boldsymbol{\zeta};\tau)$, but in practice

we must contend with statistical uncertainty in the exact electron trajectories in any one realization, meaning $\boldsymbol{\jmath}(\boldsymbol{\zeta};\omega)$ is in principle stochastic, with statistical properties determined by the phase space distribution of the particle beam (in the reduced six-dimensional phase space, if we may neglecting direct two-body or higher-order interactions, but in an even higher-dimensional phase space if particle-particle correlations are to be included). In certain cases, we can get by with using an averaged or course-grained current directly as the source, without need for further probabilistic considerations. In the Vlasov equation, for example, the fields that appear are self-consistent mean fields, i.e., solutions to Maxwell's equations for the *averaged* sources. The exact fields are linear in the sources, so it makes no difference in the end result whether we average the currents first and then calculate the resulting fields, or calculate the fields for a general instance of the currents and then average the result – the averaging procedure will commute with the convolution over the Green function. (Pragmatically, the former approach will almost always be more efficient if the average is all we need, since the random variables determining the stochastic shape of the current need not carried through the convolution, but the latter approach allows calculation of higher-order correlations or moments.) When using the MPVP to approximate the radiation, the averaged trial fields will equal the trial fields from the averaged sources if the variational parameters appear as linear expansion coefficients in a basis-set expansion, but not, in general, if they occur non-linearly in the trial wavefields. Convolution is a linear process, but optimization is in general a nonlinear one.

Unfortunately, it is often the case that we desire to estimate the fields from given sources, then calculate moments, rather than to determine the fields from averaged sources. This is particularly true of spontaneous wiggler radiation, where the radiation is almost entirely "fluctuational," both in the sense that the squared-magnitude of the average field is small compared to the average of the squared-magnitude, and in that the relative variance in the spectral intensity approaches 100% in each sufficiently narrow frequency-band for radiation from sufficiently long, unbunched electron beams. At least second-order moments are needed to estimate the power spectrum and transverse coherence properties, and fourth-order moments are necessary if we wish to estimate the uncertainty in the estimates for these spectra or coherence functions. Statistical estimates for the power spectral density, based on the moments of the MPVP trial fields, cannot in general be guaranteed to be lower bounds, as in the deterministic case.

Before even calculating moments, we may either approximate the radiation fields from the total current density using the MPVP, or else apply the MPVP separately to each

particle trajectory or family of trajectories, then superimpose the resulting variational fields. If identical, linear basis-sets are used in the expansions for each particle's radiation, identical results will be achieved (although the amount of calculation or computational effort may differ), but if nonlinear variational parameters are used or even if different truncated linear basis sets are used for different particles, the two approaches will be inequivalent.

A simple example should clarify the issues. Suppose we seek to approximate the wiggler radiation from some electron beam with negligible normalized energy spread $\frac{\delta}{\gamma} \ll 1$ but some non-zero transverse emittance $\epsilon > \lambda_0$, where $\lambda_0 \sim \frac{1}{2\gamma^2}\lambda_u(1+\frac{1}{2}a_u^2)$ is the central wavelength of the wiggler radiation in the lab frame, λ_u is the wiggler wavelength, and a_u is the normalized wiggler strength parameter. For radiation from a single on-axis electron, confined within the relative bandwidth $\frac{\delta\lambda}{\lambda_0} \sim \frac{1}{N_u}$, where N_u is the number of wiggler periods, we know the radiation is approximately diffraction-limited, with characteristic angle $\delta\theta_{\rm D} \sim \frac{1}{\sqrt{N_u}\gamma}$ and a waist located at the midpoint of the wiggler. However, the radiation for the beam as a whole will not be diffraction-limited, but will have a transverse degree of coherence $\Gamma_{{\rm coh}\,\perp} \sim \frac{\lambda_0}{4\pi\epsilon} < 1$, due to averaging, or convolution, over the transverse particle distribution.

Suppose first we use the total current density as source, and use paraxial Gauss-Hermite or Gauss-Laguerre modes as trial solutions, with optic axis coinciding with the wiggler axis. If we attempt to include, for $\lambda \approx \lambda_0$, only the fundamental Gaussian mode with undetermined waist size w_0 and location $\xi_{\rm w}$, then, regardless of the exact values of the optimized parameters, the approximation cannot simultaneously predict both the transverse size and angular divergence with accuracy, or equivalently, the partial transverse incoherence cannot be adequately captured, since we are putting all the power in a single, diffraction-limited transverse mode. In fact, we expect that we must include at least $n_\perp \sim \frac{4\pi\epsilon}{\lambda_0}$ diffraction-limited modes to resolve the transverse coherence, even if the average transverse *intensity* profile continues to look Gaussian. If the particle phase space distribution function is also taken to be Gaussian, then the needed overlap integrals in the MPVP may be calculated analytically.

Alternatively, we could decompose the current density into contributions from each electron, and employ a single Gaussian mode for each, only with the optic axis aligned along the average single-particle trajectory determined by its spatial and angular displacement from the reference orbit. The needed overlap integrals in this case are easily performed, at least for typical values of λ_0, γ, and a_u, where the transverse spatial extent of an electron orbit is small compared to w_0, but after determining the single-particle MPVP fields, each with a different propagation axis, they all must be superimposed to determine the full

field. This can be done easily in the far-field so as to determine the angular spectrum, or at arbitrary positions, to a good approximation, provided the angular deviations for the electrons remain moderately small, i.e., $\Delta\theta \lesssim \frac{1}{\gamma}$. The result will agree approximately, but not exactly, with the previous approach.

4.6.3 Possible Extensions

The results derived here applicable to localized sources which emit radiation into otherwise free space. This includes the cases of charged particle beams in bending magnets or undulators, but it would be desirable to extend applicability to more general radiation problems – for example to certain Čerenkov radiation problems, or to radiation confined within waveguides.

For fully three-dimensional propagation in vacuum, the MPVP is somewhat limited in its usefulness by the constraint that trial solutions must be solenoidal solutions to the source-free Helmholtz equation (or equivalently, eigenvectors of the double-curl operator). In three dimensions, convenient closed-form analytic solutions to focused beams or other source-free solutions with more structure than simple plane waves are generally lacking. If the trial solutions do not exactly satisfy these constraints, the results may still be of approximate validity, but the strict lower bound on the radiated power may be lost.

It appears that the proof for vacuum presented here can be more or less generalized using the macroscopic Maxwell's equations to the case of certain structured media with localized inhomogeneities, i.e., to a lossless, possibly frequency-dispersive, homogeneous (but not necessarily isotropic) linear medium characterized by spatially-invariant magnetic and electric susceptibilities everywhere in space, except possibly for a bounded region filled with a linear medium with lossless, inhomogeneous response and/or perfectly conducting boundaries or inserts. By lossless, we mean that both the dielectric tensor and permeability tensors are Hermitian at the frequencies of interest: $\epsilon(\omega)^\dagger = \epsilon(\omega)$ and $\mu(\omega)^\dagger = \mu(\omega)$. Of course, if the medium is dispersive around these frequencies, then from the Kramers-Kronig relations it must be absorptive at some other frequencies, but they may be far from the bandwidth of interest. However, filling in the details remains an open problem, and even if the results can be generalized, finding source-free solutions in such complicated geometries may be daunting.

One important class of geometries for which this formalism should be almost immediately applicable is that of conducting-wall and/or dielectric waveguides with translational

symmetry along the optic axis. In such circumstances, it is well known [88, 128, 92, 129] that such systems share much of the mathematical Hilbert-space structure as that elaborated above for paraxial radiation – in particular, solenoidal modes which are uniquely specified by their cross-section in any transverse plane, and which are orthonormalizable with respect to the natural $\mathscr{L}_2$/Euclidean inner product in that plane. (However, these modes are not complete for the near-fields, as discussed, for example, in [130].) In addition, it is easily shown that, in any such waveguides, the power in any source-free, propagating mode, measured sufficiently far downstream from a *localized* source inserted inside the waveguide, can be related to the overlap integral between the actual current density of the source and the electric field profile of this free-space mode extrapolated back into the support region of the sources as if no sources were present. Modifications due to finite conductivity effects might be then treated perturbatively. Again, this is all familiar, and is exactly analogous to our basis-set expansion technique. What is perhaps not widely appreciated is that the Cauchy-Schwarz inequality then allows one to transform this basis-set approach into a maximum principle, where the exact form of the normal modes are not needed. However, a distinct proof for this case must be supplied due to the differences in the fall-off between free-space and guided radiative modes. Radiation in the interior of closed conducting cavities does not actually propagate but rather consists of standing waves, and it is unclear whether an analogous result can hold.

Attempts to use a more general complex-valued dielectric tensor or index of refraction in order to incorporate linear loss or gain will encounter fundamental difficulties. Our derivations have exploited the fact that the free-space retarded and advanced Green functions are just related by complex conjugation in the frequency domain, but with any dissipation or gain, this time-reversal symmetry will be lost. Ideally, one would like to fully generalize the treatment to allow for dielectric tensors $\epsilon_{\mu\nu}(\boldsymbol{x}, t; \omega)$ with any frequency dispersion consistent with the Kramers-Kroning relations and/or piecewise-slowly-varying spacetime dependence (inhomogeneity), and possibly anisotropy, in order to model the effects of essentially arbitrary conducting or dielectric boundaries, obstacles, or passive optical elements, such as lenses, prisms, polarizers, waveplates, apertures, waveguides, cavities, transmission lines, fibers, gratings, etc., either ideal or lossy, but it not even clear whether this is possible, or if so, practical. Many variational techniques have been used for waveguides, cavities, and other structures, but these are usually of a stationary, rather than extremal, character, an important distinction which is examined below in Section 4.6.5. All of these directions are left as open problems.

4.6.4 Connections to Stimulated Emission

One is also naturally led to speculations as to whether similar ideas may be applied to stimulated emission. Naively, one might imagine some iterative procedure or perturbation expansion, where the approximated radiation, assuming given sources, is somehow used to deduce corresponding energy changes that must have taken place in those sources, which leads to revised estimates for the radiation, and so on. It is not clear whether this is applicable to any real problems, or even if it can converge to a self-consistent solution describing an active medium.

However, results for the case of spontaneous emission from a classical particle beam may be of direct relevance to the problem of stimulated emission, namely in the so-called small-signal regime where saturation and depletion effects are ignored. In his celebrated 1917 quantum analysis of radiation [131], Einstein first classified the radiative processes involving photons (or really any bosons) into spontaneous emission of, stimulated emission by, and stimulated absorption of the photons by atoms or other material sources. By exploiting the fact that the matter and radiation can be in thermodynamic equilibrium, Einstein used the properties of blackbody radiation and the principle of detailed balance to establish the proportionality between the intrinsic rates, within each mode, for each of these processes, as summarized in the famous A and B coefficients, results which must hold for arbitrary initial conditions, not just those consistent with thermal equilibrium. Less widely appreciated is that these relations persist in a purely classical limit of the matter and fields where Planck's constant h never appears – but clearly discussed, for example, by Beckefi in Chapter 2 of [132], and derived using somewhat different arguments in [133] or in [134]

As a simple example, consider the case of radiation transport in an electron plasma, where direct many-body effects are neglected, and with an isotropic, single-particle momentum distribution function $f(p)$, where $p = \|\boldsymbol{p}\|$ is the magnitude of kinetic momentum, and uniform density n_0 over some region. (This might also describe a long electron beam in its average rest frame.) Then, for each type of radiative process (or the aggregate of all processes), the spontaneous emission rate ℓ_ω defined as the power emitted in polarization $\hat{\boldsymbol{e}}$ per unit volume of medium per unit bandwidth per solid angle in the direction of $\hat{\boldsymbol{s}}$, may be written as

$$\ell_\omega[\hat{\boldsymbol{e}};\hat{\boldsymbol{s}}] = n_0 \int d^3\boldsymbol{p}\, \eta_\omega(\boldsymbol{p};\hat{\boldsymbol{e}};\hat{\boldsymbol{s}}) f(p), \tag{4.294}$$

where $\eta_\omega(\boldsymbol{p};\hat{\boldsymbol{e}};\hat{\boldsymbol{s}})$ is the single-particle, intrinsic emission coefficient, which is determined by the details of the microscopic physics, but is *independent* of the particle distribution function

or the incident radiation state. The *net* absorption rate, i.e., bare stimulated absorption less stimulated emission, is given by $\alpha_\omega I_\omega$, where the radiant brightness I_ω is the incident power per unit area per unit solid angle per unit bandwidth, and the absorption coefficient is given by

$$\alpha_\omega[\hat{\boldsymbol{e}};\hat{\boldsymbol{s}}] = -\frac{8\pi c^2}{\mathsf{n}_{\rm ir}^2\omega^2} n_0 \int d^3\boldsymbol{p}\, \eta_\omega(\boldsymbol{p};\hat{\boldsymbol{e}};\hat{\boldsymbol{s}}) \tfrac{\partial}{\partial\varepsilon} f(p), \tag{4.295}$$

where $\varepsilon = \varepsilon(p) = \sqrt{c^2p^2 + m^2c^4}$ is the particle energy, $\mathsf{n}_{\rm ir}$ is the background index of refraction, and $\eta_\omega(\boldsymbol{p};\hat{\boldsymbol{e}};\hat{\boldsymbol{s}})$ is the same quantity that appears in the spontaneous emission. These results are completely classical, as no factors of $\hbar$ anywhere appear. The dependence on the derivative of the distribution function is the only reminder that this expression represents the net difference between stimulated absorption and emission. These connections between spontaneous and stimulated emission/absorption are essentially a generalized manifestation of the well-known fluctuation-dissipation theorem, which relates the response of a system when perturbed to the spontaneous fluctuations which occur in the absence of external perturbation [132, 135, 133].

In situations involving instabilities, amplifiers, or any appreciable amount of stimulated emission, we naturally expect to observe in the presence of gain that mode which grows the fastest, but a similar principle is also applicable in the spontaneous-emission regime, because of these definite connections between spontaneous emission, stimulated emission, and stimulated absorption[136], even when the radiation is completely classical, and even when back-action on the charges is neglected.

In FEL physics, first described quantum-mechanically but now understood classically, such generalizations of Einstein's arguments are the basis for Madey's theorem in one-dimensional theory [137] and its generalizations to higher dimensions (see, for example, [135, 138, 133], or else [136] and references therein), where the gain curve (specifically, the relative change in intensity versus de-tuning) in the small-signal regime is proportional to the *derivative* of the spontaneous emission spectrum. Given these connections, certain properties of the spontaneous wiggler radiation, or approximations thereof, can yield information about the stimulated emission, or gain, in the small-signal regime. Perhaps more satisfying is the obverse relation, whereby our intuitions about stimulated emission provide the clearest justification and interpretation for the appearance of a maximum-power variational principle for classical spontaneous emission.

We might think to approximate the radiation profile by that which maximizes the extraction of energy from the electrons to the laser. For example, this is precisely the

heuristic justification for the variational principle suggested in [139] for optically-guided modes in FELs, where the fundamental mode is estimated by maximizing the imaginary part of the effective wavenumber. If the radiation is to be estimated by one mode, the mode should be chosen to be whatever shape is expected to have the fastest growth (or slowest loss).

We saw above that the MPVP can be interpreted precisely in this manner – as finding the mode shape which, if actually incident, would maximally couple to the given sources, and furthermore the "virtual" gain delivered would be equal to the estimated power spontaneously radiated. In fact, the only essential difference between our case, and Madey's theorem familiar from FEL physics, is that by assuming completely prescribed sources, we ignore radiation reaction, multiple scattering, or any other feedback, so once emitted from part of the source, radiation can then propagate to the far-field, but cannot induce recoil on its source or be subsequently scattered or absorbed by other parts of the source downstream. As a result, we find a relationship between the spontaneous emission spectrum and that of the"bare" stimulated emission, not the "net" response given by the difference between stimulated emission and absorption as in Madey's theorem (proportional to the derivative of the spontaneous emission spectrum in the small-gain limit.)

4.6.5 Relation to Lagrangian Formulation and Other Variational Approaches

A survey of variational techniques in electrostatics, electrodynamics, and optics reveals that, at some fundamental level, they seem to derive from only a very few general concepts: Hamilton's principle for systems derivable from a Lagrangian/action, energy conservation/optimization, electromagnetic reciprocity, and entropy maximization or equivalently free energy minimization (for problems of a thermodynamic nature.) Of course these principles are interrelated – both energy conservation and reciprocity follow from the structure of the governing Lagrangian, while entropy maximization usually relies on energy conservation or other dynamical invariants as constraints and automatically satisfies certain reciprocity relations – but nevertheless they remain useful and conceptually distinct organizational categories. Each variational principle may be further classified by whether it demands (generally constrained) optimality, or mere stationarity, of the relevant function or functional.

We examine in turn the relevance or relation of each of these principles to the MPVP. Obviously the MPVP is most closely related to energy principles, although in the case of

classical spontaneous radiation, we maximize the rate of energy transfer (power) in each spectral band, rather than, say, minimize potential energy as in electrostatics, or free energy in thermodynamics.

Entropy-Based Principles

Entropy maximization might be a useful technique to determine or approximate certain source distributions given only macroscopic thermodynamic constraints or other incomplete information, but is not directly relevant for the problem of finding the radiation from given deterministic sources, which is more a problem of dynamics than thermodynamics. However, as we saw above, new connections and rationales are revealed by generalization of the arguments leading to Einstein's A and B coefficients to classical beams or plasmas, and these arguments rely on the equilibrium properties of black-body radiation, even when the actual system of interest may be far from thermodynamic equilibrium.

Reciprocity-Based Principles

Reciprocity-based variational principles are commonly used to approximate resonant frequencies in cavities or waveguides, impedances of cavities, transmission lines, apertures, or other structures, and scattering cross sections by conductors or dielectrics [140, 89, 93]. They can be suitably generalized to many situations where the background medium is not necessarily homogeneous or isotropic. Despite superficial similarities between the MPVP and reciprocity-based techniques, a closer look reveals them to be quite distinct. Consider the solenoidal electric field $\boldsymbol{\varepsilon}_{\boldsymbol{\psi}\perp}(\boldsymbol{\zeta};\omega)$ derived from some Coulomb-gauge vector potential $\boldsymbol{\psi}(\boldsymbol{\zeta};\omega)$, and any source $\boldsymbol{j}'(\boldsymbol{\zeta};\omega)$, not necessarily that associated with $\boldsymbol{\psi}$. For our purposes, we may define the complex *reaction* of the field $\boldsymbol{\varepsilon}_{\boldsymbol{a}\perp}$ on the source $\boldsymbol{j}'$ as

$$\mathcal{R}[\boldsymbol{\varepsilon}_{\boldsymbol{\psi}\perp},\boldsymbol{j}'] \equiv \int d^3\boldsymbol{\zeta}\, \boldsymbol{\varepsilon}_{\boldsymbol{\psi}\perp}(\boldsymbol{\zeta};\omega)\cdot\boldsymbol{j}'(\boldsymbol{\zeta};\omega) = \left\langle \boldsymbol{\varepsilon}_{\boldsymbol{\psi}\perp}(\omega)^* \,\middle|\, \boldsymbol{j}'(\omega)\right\rangle, \tag{4.296}$$

Which is similar to the original definition in [140], but involves only the solenoidal electric fields, and drops the analogous terms involving the magnetic field and magnetization density, because we are working with the microscopic fields from classical sources. Given any source $\boldsymbol{j}(\boldsymbol{\zeta};\omega)$, let $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$ denote the corresponding Coulomb-gauge, retarded vector potential, and let $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega) = \boldsymbol{\chi}_{\boldsymbol{j}}(\boldsymbol{\zeta};\omega)$ denote the closest homogeneous approximant (i.e., that vector potential obtained by using $\boldsymbol{j}_\perp$ as an effective source in convolution with the fundamental solution, rather than an actual source in convolution with the retarded Green function); and define $\boldsymbol{a}'$ and $\boldsymbol{\chi}'$ in an analogous fashion for the source $\boldsymbol{j}'$.

By using the symmetry properties of the Green functions, it is then straightforward to establish the Raleigh-Carson reciprocity relation for the inhomogeneous case:

$$\mathcal{R}[\boldsymbol{\varepsilon}_{\boldsymbol{a}\perp}, \boldsymbol{j}'](\omega) = \left\langle \boldsymbol{\varepsilon}_{\boldsymbol{a}\perp}(\omega)^* \,\middle|\, \boldsymbol{j}'(\omega) \right\rangle = \left\langle \boldsymbol{\varepsilon}_{\boldsymbol{a}'\perp}(\omega)^* \,\middle|\, \boldsymbol{j}(\omega) \right\rangle = \mathcal{R}[\boldsymbol{\varepsilon}_{\boldsymbol{a}'\perp}, \boldsymbol{j}](\omega), \tag{4.297}$$

and similarly for the homogeneous case:

$$\mathcal{R}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}, \boldsymbol{j}'](\omega) = \left\langle \boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\omega)^* \,\middle|\, \boldsymbol{j}'(\omega) \right\rangle = \left\langle \boldsymbol{\varepsilon}_{\boldsymbol{\chi}'\perp}(\omega)^* \,\middle|\, \boldsymbol{j}(\omega) \right\rangle = \mathcal{R}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}'\perp}, \boldsymbol{j}](\omega). \tag{4.298}$$

Such a quantity was introduced in [140] as a measure of "reaction," or coupling, between the sources $\boldsymbol{j}$ and $\boldsymbol{j}'$, and in many contexts a variety of physical observables such as fields or forces, impedances, etc., may be proportional to or otherwise related to reactions. Despite resemblances, the complex self-reaction $\langle \boldsymbol{\varepsilon}_{\boldsymbol{\psi}\perp}(\omega)^* \,|\, \boldsymbol{j}(\omega) \rangle$ is distinct from the complex power $\langle \boldsymbol{\varepsilon}_{\boldsymbol{\psi}\perp}(\omega) \,|\, \boldsymbol{j}(\omega) \rangle$, because no complex conjugates appear in the overlap integral in the former case. (Adding to the potential for confusion, the imaginary part of the complex power is referred to as the "reactive power," following the conventions of circuit theory.) Said another way, the complex power is associated with a standard complex inner product, which is conjugate-symmetric and positive-definite, while the reaction is associated with a symmetric, bilinear form which is not a true inner product. Physically, the concepts are also distinct: energy conservation arises from the invariance properties of the full electromagnetic Lagrangian (including self-consistent particle dynamics) under time-translations, while reciprocity arises from invariance under time reversals. The real part of the complex power is of course proportional to the time-average (over an optical period) of the power exchanged between source and field. The real part of the reaction is actually proportional to the fluctuation about this average.

The concept of reaction is of interest precisely because of reciprocity properties like (4.297) or (4.298), and the possibilities for variational approximation that emerge, by demanding that all trial sources "look" the same to themselves as to the correct sources. For example, suppose we seek a trial-function approximation $\tilde{\boldsymbol{\psi}}$ to the unknown vector potential $\boldsymbol{\psi}$, given the actual source $\boldsymbol{j}$. Let $\tilde{\boldsymbol{j}}$ correspond to a source associated with $\tilde{\boldsymbol{a}}$ (the existence of non-radiating sources implies non-uniqueness, which is of no consequence here.) From a reaction point-of-view, we would like an approximation such that

$$\mathcal{R}[\boldsymbol{\varepsilon}_{\tilde{\boldsymbol{\psi}}\perp}, \tilde{\boldsymbol{j}}](\omega) \approx \mathcal{R}[\boldsymbol{\varepsilon}_{\boldsymbol{\psi}\perp}, \boldsymbol{j}](\omega), \tag{4.299}$$

but by assumption we have no tractable way of calculating the right-hand side, so instead that approximation is then "best" if we can impose the constraint

$$\mathcal{R}[\boldsymbol{\varepsilon}_{\tilde{\boldsymbol{\psi}}\perp}, \tilde{\boldsymbol{j}}](\omega) \approx \mathcal{R}[\boldsymbol{\varepsilon}_{\tilde{\boldsymbol{\psi}}\perp}, \boldsymbol{j}](\omega). \tag{4.300}$$

In fact, one can show that under this constraint, the reaction $\mathcal{R}[\boldsymbol{\varepsilon}_{\tilde{\boldsymbol{\psi}}\perp}, \boldsymbol{j}](\omega)$ is then stationary for first-order variations of $\tilde{\boldsymbol{\psi}}$ about the actual solution $\boldsymbol{\psi}$.

Note that this leads to a stationarity condition, not an optimization condition. Because reaction and power are not equivalent, in general they lead to different variational principles. In the Method of Moments, which encompasses both the Finite-Element and the Ritz-Galerkin basis-expansion techniques, one can obtain algebraic (typically linear) stationarity conditions using either a symmetric form between trial functions and weight functions, and thus automatically satisfy reciprocity relations, or using a conjugate-symmetric inner product, and satisfy energy constraints, but typically not both simultaneously. Perhaps the reaction approach may also lead to some useful variational principle for our problem, but we do not pursue this question further here. One immediate difficulty is that, if we use inhomogeneous solutions $\tilde{\psi}$ as trial vector fields, the corresponding sources are easily determined by substitution into the Helmholtz equation, but the fields themselves are difficult to parameterize in any economical manner, while if we use source-free solutions for $\tilde{\boldsymbol{\psi}}$, variational families may be more easily characterized, but the effective sources appearing in the reactions are then difficult to determine.

Action-Based Principles

Now we turn to action principles. Because the fundamental dynamics of charge-carrying matter and electromagnetic fields are ultimately derivable from a Lagrangian, we had anticipated that our maximum-power variational principle would be traced back to Hamilton's Principle, but this appears not to be the case. (Perhaps this should not have been so surprising; after all, the Lagrangian involves differences between kinetic and potential energies, while an an energy-based principle should involve the sum.) To understand the issues encountered, let us proceed by re-framing the governing dynamical equations in terms of a standard Lagrangian formulation. For prescribed current and charge densities, the dynamics of the fields are derivable from the action functional, which in terms of the Coulomb-gauge potentials and scaled coordinates may be written as a space-time integral over the Lagrangian densities for the free electromagnetic fields and the source-field interaction:

$$\mathcal{S} = \int d\tau \int d^3\boldsymbol{\zeta} \; [\mathcal{L}_0 + \mathcal{L}_{\text{int}}] \,, \tag{4.301}$$

where

$$\mathcal{L}_0 = \|\boldsymbol{\varepsilon}(\boldsymbol{\zeta};\tau)\|^2 - \|\boldsymbol{b}(\boldsymbol{\zeta};\tau)\|^2 = \|-\tfrac{\partial}{\partial\tau}\boldsymbol{a}(\boldsymbol{\zeta};\tau) - \boldsymbol{\partial}\phi(\boldsymbol{\zeta};\tau)\|^2 - \|\boldsymbol{\partial}\times\boldsymbol{a}(\boldsymbol{\zeta};\tau)\|^2, \tag{4.302}$$

and

$$\mathcal{L}_{\text{int}} = \boldsymbol{j}(\boldsymbol{\zeta};\tau)\cdot\boldsymbol{a}(\boldsymbol{\zeta};\tau) - \mu(\boldsymbol{\zeta};\tau)\phi(\boldsymbol{\zeta};\tau), \tag{4.303}$$

subject as always to the solenoidal gauge constraint

$$\boldsymbol{\partial}\cdot\boldsymbol{a}(\boldsymbol{\zeta};\tau) = 0. \tag{4.304}$$

Using the reality of the physical potentials and sources and the (Hilbert-space) orthogonality between the transverse vector field $\boldsymbol{a}(\boldsymbol{\zeta};\tau)$ and longitudinal electric field $-\boldsymbol{\partial}\phi(\boldsymbol{\zeta};\tau)$ when integrated over all space, we can re-write this as

$$\mathcal{S} = \int d\tau \int d^3\boldsymbol{\zeta}\left[(\tfrac{\partial}{\partial\tau}\boldsymbol{a}^*\cdot\tfrac{\partial}{\partial\tau}\boldsymbol{a}) + (\boldsymbol{\partial}\phi^*\cdot\boldsymbol{\partial}\phi) - (\boldsymbol{\partial}\times\boldsymbol{a}^*)\cdot(\boldsymbol{\partial}\times\boldsymbol{a}) + (\boldsymbol{a}^*\cdot\boldsymbol{j}) - (\phi^*\mu)\right], \tag{4.305}$$

where we may use $\boldsymbol{j}$ or $\boldsymbol{j}_\perp$ interchangeably within the action functional because the current density only appears in an inner product with the solenoidal vector potential $\boldsymbol{a}$. The equations of motion (Euler-Lagrange equations) for the vector and scalar potentials (and their complex conjugates) are obtained as the conditions for this action remaining stationary under independent, infinitesimal variations $\delta\boldsymbol{a}(\boldsymbol{\zeta};\tau)$, $\delta\boldsymbol{a}(\boldsymbol{\zeta};\tau)^*$, $\delta\phi(\boldsymbol{\zeta};\tau)$, and $\delta\phi(\boldsymbol{\zeta};\tau)^*$, which are fixed at the space-time boundaries (here taken to be at infinity), are consistent with the gauge constraint, and which do not destroy integrability of the Lagrangian density, but are otherwise arbitrary.

By interchanging the order of temporal and spatial integrations and using the Parseval-Plancherel Identity, we may express this action in the (scaled) frequency domain:

$$\mathcal{S} = \int d\omega \int d^3\boldsymbol{\zeta}\left[(i\omega\boldsymbol{a})^*\cdot(i\omega\boldsymbol{a}) + (\boldsymbol{\partial}\phi^*\cdot\boldsymbol{\partial}\phi) - (\boldsymbol{\partial}\times\boldsymbol{a}^*)\cdot(\boldsymbol{\partial}\times\boldsymbol{a}) + (\boldsymbol{a}^*\cdot\boldsymbol{j}) - (\phi^*\mu)\right], \tag{4.306}$$

with additional gauge constraint

$$\boldsymbol{\partial}\cdot\boldsymbol{a}(\boldsymbol{\zeta};\omega) = 0, \tag{4.307}$$

where we have used integration by parts on the time-derivatives, and where each potential and source term appearing in the action integral is now interpreted as the Fourier transform in scaled time (assumed to exist) of the function of the same name appearing in the time-domain version above, and where coordinate dependence has been suppressed for the sake of brevity and readability in the equations. Using standard vector identities, this may be written in the more transparent form:

$$\mathcal{S} = \int d\omega \int d^3\boldsymbol{\zeta}\left[\boldsymbol{a}^*\cdot(\partial^2+\omega^2)\boldsymbol{a} + \boldsymbol{a}^*\cdot\boldsymbol{j} - \phi^*(\partial^2\phi) - \phi^*\mu + \boldsymbol{\partial}\cdot(\phi^*\boldsymbol{\partial}\phi - \boldsymbol{a}^*\times\boldsymbol{\partial}\times\boldsymbol{a})\right] \tag{4.308}$$

Demanding stationarity under independent, infinitesimal variations $\delta\boldsymbol{a}(\boldsymbol{\zeta};\omega)$, $\delta\boldsymbol{a}(\boldsymbol{\zeta};\omega)^*$, $\delta\phi(\boldsymbol{\zeta};\omega)$, and $\delta\phi(\boldsymbol{\zeta}'\omega)$ satisfying the gauge constraint and fixed boundary conditions (as

$\|\boldsymbol{\zeta}\| \to \infty$ and $|\omega| \to \infty$) leads to the frequency-domain version of Poisson's equation and the wave equation in the Coulomb gauge. Now, by Gauss's theorem, the final term involving the pure divergence can be expressed as a surface integral over the spatial boundary (at infinity), which will not affect the stationarity conditions (Euler-Lagrange equations) under the assumption of infinitesimal variations with fixed boundary conditions as dictated by Hamilton's principle, so it may be dropped without any change in the resulting dynamical equations. Likewise, we may then introduce the divergence $\boldsymbol{\partial} \cdot \boldsymbol{q}$ of some other, arbitrary differentiable vector field $\boldsymbol{q}$ which may be a function of $\boldsymbol{\zeta}$ and ω and a functional of $\boldsymbol{a}$ and $\boldsymbol{a}^*$ and their spatial derivatives. The significance of this auxiliary field will emerge shortly. Since we are ultimately only interested in the radiative component of the fields, the terms in the action involving the scalar potential ϕ may also be suppressed, as they are constant with respect to variations in the vector potential. Recalling that the time-domain potentials and sources are all real, we can then use as the relevant action the modified functional

$$\tilde{\mathcal{S}} = \int_0^\infty d\omega \int d^3\boldsymbol{\zeta} \, \left[\boldsymbol{a}^* \cdot (\partial^2 + \omega^2)\boldsymbol{a} + \boldsymbol{a}^* \cdot \boldsymbol{j} + \boldsymbol{\partial} \cdot \boldsymbol{q}\right] + c.c., \tag{4.309}$$

and, noting the additivity with respect to contributions from the various frequencies ω, we may use as the variational, or "objective" or "cost" function, for each distinct (positive) frequency component the spectral density associated with this expression:

$$\tilde{\mathcal{S}}'[\boldsymbol{a}, \boldsymbol{a}^*, \boldsymbol{j}, \boldsymbol{j}^*](\omega) \equiv \int d^3\boldsymbol{\zeta} \, \left[\boldsymbol{a}^* \cdot (\partial^2 + \omega^2)\boldsymbol{a} + \boldsymbol{a}^* \cdot \boldsymbol{j} + \boldsymbol{\partial} \cdot \boldsymbol{q}\right] + c.c., \tag{4.310}$$

whose variations, subject to the solenoidal gauge constraint and given spatial boundary conditions, lead to Euler-Lagrange equations consisting of the inhomogeneous Helmholtz equation for the vector potential and its complex conjugate at the frequency $\omega > 0$. Choosing

$$\boldsymbol{q} = \tfrac{1}{2}\lambda \left[\tfrac{1}{4}\left(1 - \tfrac{i}{k}\tfrac{\partial}{\partial\zeta}\right)\boldsymbol{a} \times \left(1 + \tfrac{i}{k}\tfrac{\partial}{\partial\zeta}\right)\boldsymbol{a}^* + \tfrac{\mathcal{P}_0'}{4\pi}\,\boldsymbol{\partial}\,\zeta^{-1}\right], \tag{4.311}$$

or equivalently (at least asymptotically, as $\zeta \to \infty$),

$$\boldsymbol{q} = \tfrac{1}{2}\lambda \left[\boldsymbol{s}_{\boldsymbol{a}}^{\text{out}} - \tfrac{\mathcal{P}_0'}{4\pi}\tfrac{\boldsymbol{\zeta}}{\zeta^3}\right], \tag{4.312}$$

for some undetermined real parameters $\lambda = \lambda(\omega)$ and $\mathcal{P}_0' = \mathcal{P}_0'(\omega)$, equation (4.310) becomes:

$$\tilde{\mathcal{S}}'[\boldsymbol{a}, \boldsymbol{a}^*, \boldsymbol{j}, \boldsymbol{j}^*](\omega) \equiv 2\int d^3\boldsymbol{\zeta} \, \mathrm{Re}\big[\boldsymbol{a}^* \cdot (\partial^2 + \omega^2)\boldsymbol{a} + \boldsymbol{a}^* \cdot \boldsymbol{j}\big] + \lambda(\omega)\big[\mathscr{P}_{\text{EM}}'[\boldsymbol{a}^{\text{out}}](\infty; \omega) - \mathcal{P}_0'(\omega)\big]. \tag{4.313}$$

The chosen boundary term amounts to the addition of a Lagrange multiplier enforcing a specified outgoing power spectral density in the radiation fields at the stationary points of $\frac{\partial}{\partial\omega}\tilde{\mathcal{S}}$.

Now, a natural approach to approximation, associated variously with the names of Raleigh, Ritz, and Galerkin in slightly different contexts, consists in simply replacing this infinite-dimensional variational problem with a finite-dimensional restriction or projection of it onto some more easily characterized space of possibilities.

That is, we restrict the space of possible vector potentials to some parameterized family $\boldsymbol{a}(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha})$ for some finite-dimensional vector $\boldsymbol{\alpha}$ of continuous parameters (over some specified domain), and thereby replace a variational problem involving functional derivatives of the action functional with respect to the vector potential with one involving ordinary derivatives of an action function with respect to the variational shape parameters determining the form of the radiation and the Lagrange multiplier enforcing a power normalization. That is, assuming the family of variational trial functions is implicitly constrained to satisfy the gauge constraint, the inhomogeneous Helmholtz PDE is replaced for each $\omega > 0$ with a system of algebraic equations,

$$\tfrac{\partial}{\partial\boldsymbol{\alpha}}\tilde{\mathcal{S}}'(\boldsymbol{\alpha}) = \mathbf{0}, \tag{4.314a}$$

$$\tfrac{\partial}{\partial\lambda}\tilde{\mathcal{S}}'(\boldsymbol{\alpha}) = 0, \tag{4.314b}$$

whose solution $\tilde{\alpha} = \tilde{\alpha}[\boldsymbol{j}](\omega;\mathcal{P}'_0)$, $\tilde{\lambda} = \tilde{\lambda}[\boldsymbol{j}](\omega;\mathcal{P}'_0)$ determines the variational approximation $\tilde{\boldsymbol{a}}(\boldsymbol{\zeta};\omega;\tilde{\boldsymbol{\alpha}})$ to $\boldsymbol{a}(\boldsymbol{\zeta};\omega)$.

Now, following our earlier development, suppose we consider a family $\boldsymbol{\chi}(\boldsymbol{\zeta};\omega;\boldsymbol{\alpha})$ of trial vector fields for the general inhomogeneous problem which are explicitly constrained to be solenoidal and to satisfy the source-free Helmholtz equation, so the first term in the action density vanishes identically. Then the variational action spectral density becomes

$$\tilde{\mathcal{S}}'(\omega;\boldsymbol{\alpha};\lambda) = \tfrac{2}{\omega}\,\mathrm{Im}\,\langle\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\omega;\boldsymbol{\alpha})\,|\,\boldsymbol{j}(\omega)\rangle + \lambda(\omega)\big[\mathscr{P}'_{\mathrm{EM}}[\boldsymbol{\chi}^{\mathrm{out}}](\infty;\omega) - \mathcal{P}'_0(\omega)\big]. \tag{4.315}$$

Despite certain similarities, the Ritz-Galerkin-type variational principle associated with this action is distinct form the MPVP. The action principle involves finding constrained stationary points of $\mathrm{Im}\,\langle\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\omega)\,|\,\boldsymbol{j}(\omega)\rangle$, while the MPVP involves finding the constrained maxima of $\mathscr{P}_{\mathrm{mech}}[\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp},\boldsymbol{j}](\omega;\boldsymbol{\alpha}) = \mathrm{Re}\,\langle\boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\omega;\boldsymbol{\alpha})\,|\,\boldsymbol{j}(\omega)\rangle$.

Although generically the "true" Euler-Lagrange solutions correspond to saddle-points, when the trial functions are restricted to a parameterized source-free family and when the power-normalization constraint is imposed via a Lagrange multiplier, the overall problem

inherits convexity from the constraint alone, and the critical points turn out to be power-constrained local maxima or minima of $\operatorname{Im}\langle \boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\omega;\boldsymbol{\alpha})\,|\,\boldsymbol{j}(\omega)\rangle$, which implies that the overall phase should be chosen such that $\operatorname{Re}\langle \boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\omega;\boldsymbol{\alpha})\,|\,\boldsymbol{j}(\omega)\rangle = 0$. In the MPVP case, the variational solution corresponds to constrained local maxima of $-\operatorname{Re}\langle \boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\omega;\boldsymbol{\alpha})\,|\,\boldsymbol{j}(\omega)\rangle$, so that $\operatorname{Im}\langle \boldsymbol{\varepsilon}_{\boldsymbol{\chi}\perp}(\omega;\boldsymbol{\alpha})\,|\,\boldsymbol{j}(\omega)\rangle = 0$. Any attempt to employ directly the action-based variational principle with the source-free trial solutions thus encounters two related problems: the ostensible method converges to a solution with a global phase error of $\pi/2$; and as an immediate result, the calculated work performed on/by the sources vanishes, so the problem becomes degenerate, and we are left with no means to consistently choose the *absolute* power level.

It is not difficult to trace the origin of this phase error. Naive application of Fourier transforms to the Green function yields the reciprocal-space, frequency-domain kernel

$$G^{\text{out}}(\boldsymbol{k},\boldsymbol{k};\omega) = \delta(\boldsymbol{k}-\boldsymbol{k}')\frac{1}{\omega^2 - \|\boldsymbol{k}\|^2}, \tag{4.316}$$

but this neglects the requirement of causality demanded of the retarded Green function, wherein the response can appear only after the source is applied, implying that the Green function must be analytic for $\operatorname{Im}\omega \geq 0$. We must be slightly more careful, and here treat the (temporal) Fourier transform as the limiting case of a Laplace transform, where the initial conditions are pushed back into the arbitrarily remote past rather than future, before being forgotten altogether. The Green function (4.316) must be understood in the sense

$$G^{\text{out}}(\boldsymbol{k},\boldsymbol{k};\omega) = \lim_{\epsilon\to 0^+} \delta(\boldsymbol{k}-\boldsymbol{k}')\frac{1}{\|\boldsymbol{k}\|^2 - (\omega+i\epsilon)^2} \tag{4.317}$$

where the limit is taken only *after* G^{out} is convolved with the Fourier transform of the source. In the full Fourier representation, the vector potential and transverse source are then related by

$$\boldsymbol{a}(\boldsymbol{k};\omega) = \lim_{\epsilon\to 0^+} \frac{\boldsymbol{j}_\perp(\boldsymbol{k};\omega)}{\|\boldsymbol{k}\|^2 - (\omega+i\epsilon)^2}, \tag{4.318}$$

where

$$\boldsymbol{j}_\perp(\boldsymbol{k};\omega) = (1-\hat{\boldsymbol{k}}\hat{\boldsymbol{k}}^{\mathrm{T}})\boldsymbol{j}(\boldsymbol{k};\omega), \tag{4.319}$$

and

$$\boldsymbol{a}(\boldsymbol{k};\omega) = \tfrac{1}{(2\pi)^{3/2}} \int d^3\boldsymbol{\zeta}\, \boldsymbol{a}(\boldsymbol{\zeta};\omega) e^{-i\boldsymbol{k}\cdot\boldsymbol{\zeta}}, \tag{4.320a}$$

$$\boldsymbol{j}(\boldsymbol{k};\omega) = \tfrac{1}{(2\pi)^{3/2}} \int d^3\boldsymbol{\zeta}\, \boldsymbol{j}(\boldsymbol{\zeta};\omega) e^{-i\boldsymbol{k}\cdot\boldsymbol{\zeta}} \tag{4.320b}$$

$$\tag{4.320c}$$

are the (scaled) Fourier transforms. For real ω and $\|\boldsymbol{k}\| > |\omega|$, we see that

$$\arg\left[\frac{\hat{\boldsymbol{u}}\cdot\boldsymbol{a}(\boldsymbol{k};\omega)}{\hat{\boldsymbol{u}}\cdot\boldsymbol{j}_\perp(\boldsymbol{k};\omega)}\right] = 0, \tag{4.321}$$

while for $\|\boldsymbol{k}\| < |\omega|$,

$$\arg\left[\frac{\hat{\boldsymbol{u}}\cdot\boldsymbol{a}(\boldsymbol{k};\omega)}{\hat{\boldsymbol{u}}\cdot\boldsymbol{j}_\perp(\boldsymbol{k};\omega)}\right] = \pi, \tag{4.322}$$

for any unit vector $\hat{\boldsymbol{u}}$. In either case, the associated electric field for these spectral components is then in quadrature with the current, and no time-averaged work is performed or energy exchanged. However, on the singular resonant manifold where $\|\boldsymbol{k}\| = |\omega|$ exactly, the magnitude of the Green function diverges in the limit as $\epsilon \to 0^+$, but the phase is such that

$$\arg\left[\frac{\hat{\boldsymbol{u}}\cdot\boldsymbol{a}(\boldsymbol{k};\omega = \|\boldsymbol{k}\|)}{\hat{\boldsymbol{u}}\cdot\boldsymbol{j}_\perp(\boldsymbol{k};\omega = \|\boldsymbol{k}\|)}\right] = \frac{\pi}{2}. \tag{4.323}$$

The relative phase between field and source then jumps discontinuously, to where the currents deliver time-averaged energy to the fields. Our purported action principle fails to capture this, while the MPVP, although operating entirely within this singular manifold of source-free solutions, relies on energy conservation, which then enforces the correct phase relation. This is not to imply that action-based variational principles are not also useful for radiation problems, only that they will not naturally work with source-free solutions, and generically will involve finding saddle-points, not extrema.

Extremal Principles Versus Stationary Principles

In fact, most of the variational principles previously employed for FEL analysis, paraxial wave propagation, or laser-plasma problems are based on precisely this type of generalized Ritz-Galerkin approximations to the underlying Lagrangian dynamics, although some of the authors have erroneously claimed extremal instead of mere stationary character for their techniques. A Lagrangian-based variational principle is developed in [141] for paraxial optical propagation in an inhomogeneous gain medium, which is generalized in [142] to include nonlinear self-focusing effects, but the authors erroneously claim extremal, not just stationary, properties. This approach was further extended to include more general laser-plasma interaction terms in [143], where the authors correctly state that their trial solutions derive from a stationary principle. In [139] the authors state without proof an extremal principle for the fundamental FEL mode in the small-signal regime, which in fact appears to be perfectly correct, but higher modes will merely be stationary. These ideas are further generalized in [144, 145, 145, 146], where the variational techniques are erroneously

described as extremal principles. Similar trial-function-based variational principles are also used in [147, 148, 149] for which, correctly, only stationarity is claimed.

A number of other authors have made extensive use of (correctly-stated) *stationary* action principles both for general wave-wave and wave-particle interactions in kinetic and fluid treatments of plasmas and particle beams [150, 151, 152, 153, 154] and more specifically for analysis of FELs [155]. However, rather than involving parameter-laden trial functions, these approaches employ the variational principles in order to provide economical descriptions and derivations of dynamical equations and conservation laws in the form of conventional ODEs or PDEs, and to effect in an efficient and transparent manner certain eikonal expansions or approximations, or oscillation-center/ponderomotive or other averaging procedures [156, 157] while preserving the Hamiltonian nature of the dynamics.

Such confusion between stationary and extremal principles seems deeply embedded in physics literature and folklore. Linguistically, "optimizing," "maximizing," or "minimizing" simply sound better than "criticalizing" or "making stationary." Philosophically, optimality enjoys a certain teleological appeal which mere stationarity lacks. Historically, the buzzwords "least action" have been invoked so often in the discussion of dynamics that Hamilton's Principle and the Principle of Least Action have been mistakenly conflated, despite being quite distinct concepts. The Principle of Least Action was first articulated by Maupertuis and formalized by Euler, not Hamilton; is restricted to a smaller class of Lagrangians (no explicit time dependence); uses a different action (the abbreviated action integral) and a different set of constrained variations (the energy is fixed but temporal endpoints are not); and in fact it is itself misnamed, requiring only that the abbreviated action be stationary, not necessarily minimal, for the physical trajectory. Practically, both optimization principles and stationary principles share an insensitivity of the variational quantity to the trial functions, but the extremal case additionally provides an upper or lower bound.

Numerically, more efficient and/or reliable computational techniques may be employed for optimization problems. Finding (local) maxima or minima of functions is much easier than finding roots of systems of equations. An "Alpine" analogy is often made: lost in a foggy terrain, the mountain-climber can reliably find a nearby peak (local maximum) by always moving uphill, a nearby valley (local minimum) by wandering downhill, but has no guaranteed strategy for finding a mountain pass (saddle-point.) If the original action density is convex (at least in some sufficiently large region of function space), so that the true solution corresponds to an extremum of the action, then in the Raleigh-Ritz approximation

we have a natural metric to measure both "closeness" to the true solution and "progress" in iteratively determining the approximate one – namely, a metric induced by the action itself. Without convexity, the true and approximate solutions may correspond to saddle-points, and we must introduce some arbitrary external metric to measure similarity or improvement. Even in the classic formulations which employ only linear expansions in basis functions, the extremal case leads to the solution of equations corresponding to a symmetric (or Hermitian), positive definite matrix, while in the merely stationary problem the matrix is in general Hermitian but not positive definite.

In the case of electromagnetic radiation, the stationary points are generically saddles, because of the hyperbolic nature of the wave equation. Alternatively, one can see this by the well-known equivalence of electromagnetic mode dynamics to a collection of harmonic oscillators, for which the action functional is known to be minimal only for sufficiently short times intervals away from turning points along an orbit.

Similar subtleties occur in the usual Raleigh-Ritz approximation for the stationary states in quantum mechanics. The energy-expectation value is always a local minimum for the ground state, and is always stationary for any excited state, but is only minimal for excited states if the variations are constrained to be orthogonal to all lower states.

Comparisons and Contrasts: The Bottom Line

It appears that the MPVP is somewhat reminiscent of, but actually quite distinct from, action-principles in Lagrangian formulations of electrodynamics, as well as various widely-known reaction-based principles used in waveguide and cavity analysis, and minimum energy or free-energy principles, Ritz-Galerkin methods commonly used in numerical simulation, the Rayleigh-Ritz variational principle familiar from textbook quantum mechanics, or various specialized variational principles developed for FEL analysis[139, 144, 146, 145, 158, 147, 148, 149, 143] or laser propagation[141, 142].

Fundamentally, in vacuum (or possibly certain other media), use of the MPVP involves finding constrained *extrema* of quantities of the form:

$$\mathscr{P}'[\omega;\boldsymbol{\alpha}] = -\,\mathrm{Re}\int d^3\boldsymbol{x}\,\boldsymbol{\varepsilon}(\boldsymbol{x};\omega;\boldsymbol{\alpha})^*\cdot\boldsymbol{j}(\boldsymbol{x};\omega); \tag{4.324}$$

while in *reciprocal* media (i.e., with *symmetric* susceptibility tensors), Rumsey reaction-based variational principles[140] would involve finding constrained *stationary* points of

quantities of the form:

$$\mathcal{R}[\omega;\boldsymbol{\alpha}] = \mathrm{Re}\int d^3\boldsymbol{x}\,\boldsymbol{\varepsilon}(\boldsymbol{x};\omega;\boldsymbol{\alpha})\cdot\boldsymbol{j}(\boldsymbol{x};\omega); \tag{4.325}$$

and in *lossless* media (characterized by *Hermitian* susceptibility tensors), Lagrangian action-based principles[92] would involve finding constrained *stationary* points of quantities of the form:

$$\mathcal{S}[\omega;\boldsymbol{\alpha}] = \mathrm{Im}\int d^3\boldsymbol{x}\,\boldsymbol{\varepsilon}(\boldsymbol{x};\omega;\boldsymbol{\alpha})^*\cdot\boldsymbol{j}(\boldsymbol{x};\omega), \tag{4.326}$$

and because of the hyperbolic nature of the wave equation, the stationary points are generically saddle-points, rather than maxima or minima, so no bound on the radiated power can be obtained. So indeed they all appear to be distinct variational principles, although in some vague sense they might be interpreted almost like analytic continuations or Hilbert transforms of each other.

Mathematically, the MPVP is most closely related to the family of extremal variational principles described by the Lax-Milgram theorem [159, 160], involving bilinear (or sesquilinear) forms which are both bounded (or, equivalently, continuous) and coercive (or, synonymously, elliptic), in some norm within a relevant Banach space, which, roughly speaking, means that all elements of the spectrum of the associated linear operator are finite but non-zero. However, for the MPVP, the relevant form, namely $-\langle\boldsymbol{\chi}_\perp\,|ik\mathcal{D}(\omega)|\,\boldsymbol{j}_\perp\rangle$, is not strictly positive definite and is only bounded and coercive in a semi-norm, again because of the existence of non-radiating sources constituting the kernel of $\mathcal{D}(\omega)$. In order to obtain a well-defined extremal variational principle with unique solution, the trial solutions must be restricted to transverse free-space solutions to the Helmholtz equation, which are uniquely related to the far-field radiation pattern but contain no information about the actual non-radiative fields in the vicinity of the sources.

4.7 Application to Harmonic Cascade FEL Radiation

The MPVP actually arose out of an analysis of coherent X-ray generation via harmonic cascade in a radiating electron beam traveling through a series of undulators[116, 117, 118, 119, 120]. Growing interest is focusing on X-ra sources that use seeded electron beams to drive a Free Electron Laser (FEL), rather than relying on amplification of shot noise or Langmuir-Nyquist noise. A simplified schematic of one such idea is shown in Fig. 4.1. Energy modulation induced by overlap of a seed laser in one undulator is converted into spatial modulation (micro-bunching) at the fundamental and higher harmonics (due

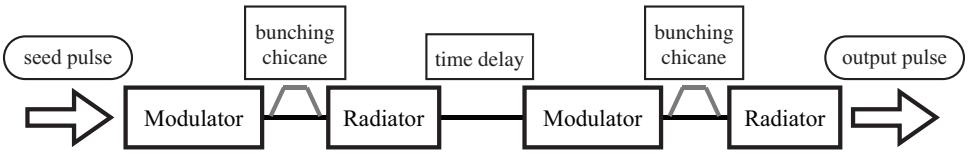

Figure 4.1. Schematic of a harmonic cascade device. The seed laser enters from the left, and the output pulse at some high integral harmonic of the original carrier frequency exits from the right. Two complete stages are shown. Each chicane rotates phase space to convert energy modulations into (spatial) micro-bunching. The time delay between stages allows a fresh slice of the electron beam to be used in the second modulator.

to nonlinearities) in a specialized dispersive beam-line (chicane), and then the beam is induced to radiate at a chosen harmonic in a suitably-tuned second radiator-undulator. The whole process can be cascaded, in which the output radiation from one stage is used as an optical seed for the next stage, overlapping with a fresh part of the beam in the subsequent modulator, to produce still-higher harmonics.

Typically the actual gain is sufficiently low in each radiator-undulator, such that prior bunching from the modulator/chicane sections dominates over the FEL self-bunching instability, so the MPVP may be used[1] to estimate the profile and power of the output radiation for any stage, neglecting the direct feedback effects of the synchrotron radiation emitted in a particular undulator on the electrons themselves during that stage. The radiation from the initial seed or previous radiator is essential to the action of the modulator, but this is modeled simply by assuming that energy modulation is sinusoidal and depends only on the ponderomotive phase Ψ of the electrons and a strength characterized by the a real parameter γ_M, while phase-space rotation in the chicane is governed by a parameter labeled R_{56} due to conventions adopted in the theory of symplectic maps. In the radiator of period $lambda_u$, length $N_u\lambda_u$ and normalized strength a_u, the electrons are assumed to emit "spontaneously" (but partially coherently) due to the effects of their pre-punching, rathe rather than due to the nonlinear FEL instability.

Instead of resorting to detailed but time-consuming numerical simulations or summation over single-particle fields, the modal structure of the radiation may be approximated in terms of a paraxial mode described by certain: the carrier wavenumber k and frequency ω, the Raleigh range $Z_{\rm R}$ (or equivalently the spot size), the waist location s_0, the overall

[1]In the interest of intellectual honesty and full disclosure, we point out explicitly that we had no hand in the actual numerical work, which was performed by G. Penn comparing results of GENESIS, a 3D FEL code developed by S. Reiche, to trial function approximations calculated by the CAMPANILE Mathematica©script written by M. Reinsch. We summarize the work of our colleagues[120] in order to offer a simple example of how the simple result of all of the previous formalism and pedantry actually finds useful application.

amplitude E_0, the carrier phase offset Φ_0, and normalized complex expansion coefficients determining the relative coherent proportions of higher-oder mode contributions to the output mode. In these preliminary studies, the trial solution was chosen for simplicity to be the fundamental Gaussian paraxial mode.

The electron beam is characterized by a full $6D$ phase space distribution function $f(\boldsymbol{Z})$ normalized such that $\int d^6\boldsymbol{Z} f(Z) = 1$, and a mean current I_e. The initial distribution itself (before modulations or bunching are induced) was taken to be either Gaussian or flat-top in various phase space coordinates, characterized longitudinally by a mean normalized energy γ_e and an energy spread parameter Δ_γ (either RMS or total, depending on the shape of the distribution chosen), and transversely by *normalized* emittances ϵ_x and ϵ_y. In addition, the possibility of "beam-conditioning" is accommodated, whereby intentional correlations are introduced between beam energy and the transverse actions to possibly improve performance, with the relative degree of correlations specified by the parameters κ_x and κ_y which have the dimensions of wavenumbers. In all the examples shown here, electron beam parameters were chosen as: $I_e = 500$ A, $\gamma_e = 6067$ corresponding to a beam energy of $\gamma_e m_e c^2 \approx 3.1$ GeV, $0 \lesssim \Delta_\gamma \lesssim 2$, $\epsilon_x = \epsilon_y = 2$ μm, and -0.8 $\mu\text{m}^{-1} \lesssim \kappa_x = \kappa_y \lesssim +0.8$ μm^{-1}. All sections are assumed to use planar wigglers, but electrons are taken to experience equal focusing in both transverse directions.

Two output stages were separately considered: the first stage of a cascade that converts $\lambda_0 = 200$ nm seed radiation into $\lambda = 50_{\text{nm}}$ at the 4th harmonic , and a final output stage which converts 3.13 nm to 1.04 nm radiation at the 3rd harmonic.

As mentioned above, evolution of the electron distribution function in the modulator and chicane sections is effected by parameterized advanced maps. In the radiation undula-

Case	Results	Analytic Theory	GENESIS: $M^2 \equiv 1$	GENESIS: fit M^2
50 nm	P_L (MW)	130.3	134.2	134.2
	Z_R (m)	1.12	0.94	0.97
	s_0 (m)	1.20	1.19	1.21
	M^2	$\equiv 1$	$\equiv 1$	1.04
1.04 nm	P_L (MW)	35.1	39.0	39.0
	Z_R (m)	52.7	49.0	33.0
	s_0 (m)	-10.4	-14.6	0.73
	M^2	$\equiv 1$	$\equiv 1$	1.72

Table 4.1. Comparison between the trial function approximations and GENESIS simulations for two case studies. Initial beam parameters are as given in the text, while beam-line parameterized are optimized. From [120].

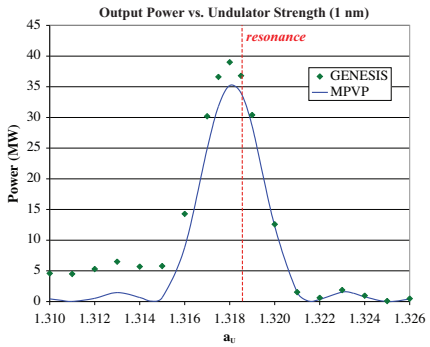

Figure 4.2. Comparison of single-stage GENESIS FEL simulation with variational approximation based on a Gaussian trial mode showing predicted power for harmonic generation at $\lambda = 1.04$ nm, as a function of the normalized undulator strength $a_{\rm u}$. Reproduced from [120].

tor, FEL equations for particle energy and phase and radiation amplitude were developed accounting for pre-bunching, transverse emittance, the quiver and focusing effects of the undulator, but no external focusing.

In this setting, the MPVP is particularly straightforward. Instead of explicitly calculating the beam current density, and then the overlap integral between it and the trial radiation mode, $\frac{d}{dz}\gamma_e$ can be related directly to the $\frac{d}{dz}a_{\rm L}$ of the output field directly by energy conservation, and the output radiation envelope $a_{\rm L}(x, y; z)$ can be written as an integral over the generalized bunching factor, which will be a functional of the beam distribution function and the assumed trial mode profile. The result will be the same, modulo that that factor of two.

At this point, many parameter such as the carrier frequency and wavelength and the

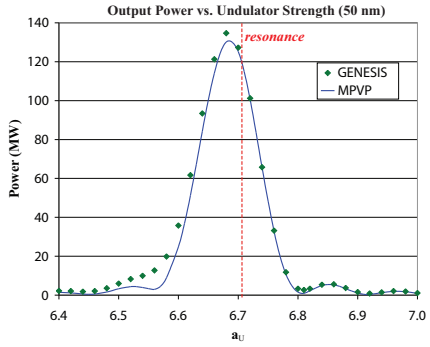

Figure 4.3. Comparison of single-stage GENESIS FEL simulation with variational approximation based on a Gaussian trial mode showing predicted power for harmonic generation at $\lambda = 50$ nm, as a function of the normalized undulator strength $a_{\rm u}$. Reproduced from [120].

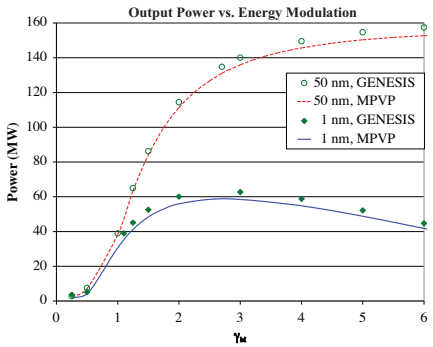

Figure 4.4. Comparison of single-stage GENESIS FEL simulation with variational approximation based on a Gaussian trial mode, for harmonic generation at $\lambda = 50$ nm and $\lambda = 1.04$ nm, showing predicted power as the beam energy modulation parameter γ_{M} is varied and the chicane slippage factor R_{56} is re-optimized. Reproduced from [120].

beam energy, energy spread, and emittances, are assumed given as input or determined directly through subsequent dynamical considerations, but the Raleigh range Z_{R} (or equivalently spot size) and waist location of the output mode, remain free parameters, while the output power can, *ceteris paribus*, be written as a scalar function of these remaining adjustable parameters.

For a given spot size and waist location, there exists any number of complete, orthonormalized, paraxial basis sets, such as the Gauss-Hermite or Gauss-Laguerre modes, which include the fundamental Gaussian mode, and into which any paraxial radiation field can be expanded in the $\mathcal{L}_2$ sense. Because the exact paraxial result may include power in all these modes, the approximation to the radiated power including only the fundamental mode is expected to systematically underestimate the power, suggesting the best procedure is choose Z_{R} and s_0 so as to maximize the corresponding power, yielding a greatest lower bound within the allowed form of trial solutions.

This is precisely a re-statement of the MPVP, which in this context can be seen to be just a specialized form of the Bessel Inequality. All previous mathematics provides formal justification for a simple but effective procedure which seems intuitively plausible, physically appealing, numerically efficient, and sufficiently accurate for the present purposes.

The MPVP approach is particularly well suited to this application, because, for given choice of wavelength, the output power is really the primary figure-of-merit of interest.[2] As mentioned, with the variational approach, the closer the trial function is to the true answer, the more accurate the estimate will be, but because at the maximum the predicted

[2]Transverse coherence is an important secondary performance criterion, but cannot be estimated with the simple framework discussed here.

power radiated is second-order accurate compared to the shape errors in the optimized trial function, even a somewhat poor approximation to the laser field can lead to a reasonably good estimate for the laser power. The single diffraction-limited Gaussian mode can be expected to provide a reasonable approximation to the FEL output for moderate values of the emittance. For small values of the electron emittance (compared to the minimal optical emittance) i.e., $\frac{\epsilon}{\gamma_e} \ll \frac{\lambda}{4\pi}$, the output radiation is expected to be transversely coherent, dominated by the fastest-growing self-guided mode, but the transverse field profile of this single mode is not guaranteed to be Gaussian, although it is often approximated a such. In contrast, in the emittance-dominated regime, $\frac{\epsilon}{\gamma_e} > \frac{\lambda}{4\pi}$ convolutions over the broad electron distribution will lead to fields that, while quite possibly Gaussian in transverse intensity profile, will not be transversely coherent, and cannot be well approximated by a single diffraction-limited mode, but rather require a suitable statistical mixture.

In analyzing, assessing, and comparing such devices, typically iterative parameter searches are employed to improve or even optimize design, and in such efforts the the MPVP proved far simpler, and orders-of-magnitude faster, than detailed numerical simulations using multidimensional, time-dependent FEL codes, while still providing reasonable accuracy. The power-maximization over adjustable parameters in the trial radiation mode can be incorporated quite naturally into, and this case can in fact be performed simultaneously with, the iterative search for other optimal design parameters, such as energy spread and introduced energy modulation, undulator strengths, chicane slippage factors, conditioning parameters, etc., by performing a simultaneous direct multi-dimensional search by Nelder-Mead (simplex), Powell (direction-set), or other standard non-gradient-based numerical methods.

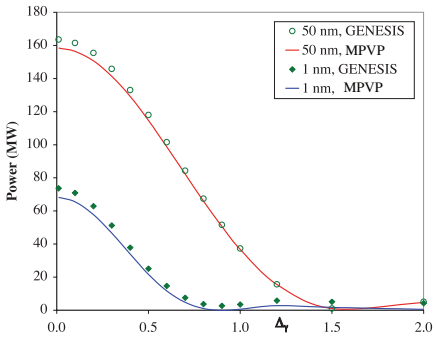

Figure 4.5. Comparison of single-stage GENESIS FEL simulation with variational approximation based on a Gaussian trial mode, for harmonic generation at $\lambda = 50$ nm and $\lambda = 1.04$ nm, showing predicted power as the energy spread is varied. Reproduced from [120].

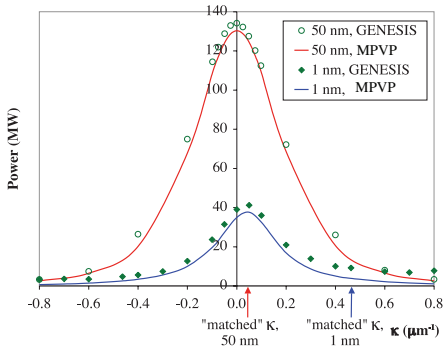

Figure 4.6. Comparison of single-stage GENESIS FEL simulation with variational approximation based on a Gaussian trial mode, showing predicted power with respect to the beam conditioning parameter κ. The matched values refer to predictions from a different geometry and regime, where the FEL output grows from noise and no chicane is used. Reproduced from [120].

In contrast, with a full FEL code such as GENESIS, which involves a three-dimensional, time-dependent representation of the FEL equations for fields and particles in the paraxial approximation, each adjustment in the parameters, however small, requires redoing an intensive computation from scratch, or at least for all points downstream from where any changes to the seed radiation, beam distribution function, or beam-line are introduced.

Typical results for the output radiation parameters are shown in Table 4.1. For the Genesis, the parameters were determined by power-weighted fits, with and without constraints on M^2, which in paraxial optics is frequently taken to characterize the transverse mode structure,[3] and is conventionally defined as the ratio of the RMS optical emittance of the laser to the minimum possible emittance allowed by the Fourier-Heisenberg uncertainty principle, $\frac{\lambda}{4\pi}$, or equivalently as the ratio of the ideal Rayleigh length for the given waist diameter to the actual Raleigh length. As can be seen, agreement on predicted power is reasonably good in both cases (10% or less), even though in the short wavelength case $\lambda_0 = 1.04$ nm) the variational predictions for the effective spot size and waist location deviate significantly from the GENESIS simulations apparently because of some higher-order mode structure. Both the spot size and location can differ significantly from what is predicted for the a spontaneous spectrum from an unbunched beam, using simple Geometric (ray-tracing) arguments.

[3] As a supposed measure of partial transverse coherence, M^2 does not always adequately distinguish between coherent and incoherent admixtures of higher-order modes. In the analogous quantum mechanical language, we would ask in a particular context whether we care about distinguishing mixed and pure states, or minimal uncertainty pure states from other pure states. Low-order moments of course distinguish the latter, but not necessarily the former, but this is a story for another day....

Further comparisons showing the variation in predicted power with changes in various parameters are shown in Figs. 4.2 - 4.6. Because the GENSIS simulations include feedback on the beam, and are subject to their own numerical errors, while the variational approach made certain other approximations to the electron trajectories in addition to the assumption of a simple trial function, there is no strict guarantee that the variational estimates should rigorously be lower found for the GENESIS results, but it is reassuring to see that tin fact they are, systematically underestimating what we might regard as the essentially exact numerical solutions, by an average of about 3% for the 50 nm case and about 10% for the 1 nm case. This suggests that the effects of the FEL instability, trapping or other mechanisms that could violate our presumption of prescribed sources independent of the radiation fields are small. Accuracy could be further improved by including additional higher-order modes, or otherwise adding additional adjustable parameters to allow for ellipticity, annularity, kurtosis, or skew/misalignment, etc. in the radiation profile, but this simple Gaussian trial mode has proven sufficiently accurate for the present purposes.

As an aside, we also point out that the MPVP speaks much more generally to the problem of how to "best" approximate spontaneous wiggler radiation by a Gaussian mode. Because the temptations of a simple mathematical form with analytically tractable or numerically efficient integrals are so great, undulator radiation is conveniently and frequently modeled as a paraxial Gaussian mode, even when the exact solutions are known to exhibit significant deviations form this shape. To choose the "best" fit, authors have advocated various forms of moment-matching, power-matching, least-squares fitting, or other approaches. This effects what is taken to be the spot size or Raleigh range of the radiation, as well as other properties. The MPVP yields an $\mathcal{L}_2$, or-power-based projection of the true solution onto the Gaussian, independent of what paraxial basis is chosen for the orthogonal complement, so should be equivalent to a least-squares approach, but does not need an independent adjustment of the overall amplitude, having determined the power through energy-conservation, and does not require knowledge of the exact solution which is to be approximated.

4.8 Conclusions

We have reviewed in some detail a Hilbert-space and operator-based approach to electromagnetic radiation, and have used this formalism to derive with some rigor a maximum-power variational principle (MPVP) for spontaneous radiation from prescribed classical

sources in what is otherwise vacuum, first in the paraxial limit and then independently in a more general three-dimensional geometry.

The MPVP is most likely to find further application in the paraxial regime, because source-free solenoidal trial functions can be more readily characterized and parameterized. Except for plane-wave or multipole expansions, or far-field limits, few solutions can be found analytically in the general free-space case, while in the paraxial limit, these solutions are uniquely specified just by the carrier frequency and the (complex) profile in any transverse plane, which can be decomposed into a convenient, countable set of expansion modes consisting of the eigenmodes of some quantum-like operators determining the radial structure and the spin and "orbital" angular momentum. We shall conjecture, but have not proven, that a self-consistent form for the MPVP may hold exactly at each order in the generalized paraxial expansion in powers of the diffraction angle.

Although similar to well known variational principles widely used in electromagnetic theory, the MPVP appears to be an independent result, and thus adds to the large family of variational techniques available for electromagnetic problems in general, and undulator radiation in particular.

The techniques have been developed within the context of undulator radiation from relativistic electron beams, for which an example involving high harmonic generation was discussed, but are more broadly applicable to other synchrotron, antenna, or perhaps other radiation problems. A major limitation is the restriction to propagation in free space. We expect that this trial function approach might be generalized to certain structured media, but this remains unproven.

A second limitation is the handling of partially coherent or incoherent radiation fields. The processes of averaging over any statistical uncertainty in the particle trajectories constituting the source and of performing the variational optimization do not in general commute if any variational parameters appear nonlinearly in the trial fields, so some care must be taken to adequately estimate quantities such as transverse or longitudinal optical coherence, or partial polarization. If the mean sources are used directly in the variational fit, then by linearity the expectation values of the fields are approximated, but often the covariances or other higher-order moments are what is desired or relevant in cases of partial coherence. We are currently studying whether a similar variational principle can be established directly for the second-order coherence tensors, based on the van Cittert-Zernicke theorem stating that a coherence tensor will satisfy a wave equation similar to that governing the underlying fields themselves.

Mathematical details aside, at its most essential, the MPVP is really just a straightforward and rather obvious consequence of two simple and rather obvious constraints, together with another fundamental mathematical fact: the power in any one source-free mode of the electromagnetic field may not exceed the total power in all the modes (i.e., Bessel inequality); and the power radiated must be attributable to power delivered by the sources, even in the regime where we ignore back-action on the sources (i.e., conservation of energy); while information about the fields on some boundary surface, needed to determine this radiated power, can be converted into information about the derivatives of the fields in the interior (i.e., Gauss's law, one of the multidimensional generalizations of the Fundamental Theorem of Calculus).

This approximation, or ones similar to it, has been frequently used, almost without comment or perceived need for further justification, in classical or quantum stimulated emission situations, where in the presence of gain we naturally expect to observe that mode which grows the fastest, but it is equally applicable in the spontaneous regime, because arguments along the lines of Einstein's derivation of the A and B coefficients lead to definite connections between spontaneous emission, stimulated emission, and stimulated absorption, even when the radiation is completely classical.

However simple, even trivial, these observations are not without practical content or application to undulator and possibly other radiation problems.

Acknowledgements

The inspiration for this work was the original insight of G. Penn, and our mathematical machinations just serve to more rigorously justify his intuitions. We also acknowledge useful discussions with W. Fawley, R. Lindberg, J. Morehead, A. Zholents, and J. Zimba. R. Lindberg also provided valuable assistance with some finicky eps and pdf graphics files.

Chapter 5

Quantum Mechanical Treatment of Transit-Time Optical Stochastic Cooling of Muons

A. A violent order is disorder; and
B. A great disorder is order. These
Two things are one.

Wallace Stevens
"Connoisseur of Chaos"

In all chaos there is a cosmos, in all disorder a secret order.

Carl Jung

No phenomenon is a phenomenon until it is an observed phenomenon.

John Archibald Wheeler

5.1 Introduction and Overview

Developed by Simon van der Meer and collaborators (see [161, 162, 163] for reviews), conventional stochastic cooling using radio-frequency (RF) signals has achieved great success in increasing phase space density of particle bunches in storage rings, for heavier particles such as protons or anti-protons where synchrotron radiation damping is inefficient. Cooling time-scales typically range from minutes to hours. Ultra-fast stochastic cooling

(i.e., on microsecond time-scales) would be desirable in certain applications[164], for example, to boost final luminosity in the proposed muon collider, where the short particle lifetime severely limits the time available to reduce beam phase space. But fast cooling requires very high-bandwidth amplifiers so as to limit the incoherent heating effects from neighboring particles. A method of transit-time optical stochastic cooling (OSC) has been proposed[165, 166] which would employ high-gain, high-bandwidth, solid-state lasers to amplify the spontaneous radiation emitted from the charged particle bunch in a strong-field magnetic wiggler. This amplified light is then fed back onto the same bunch inside a second wiggler, with appropriate phase delay to effect cooling. But before amplification, the usable signal from any one particle is quite small, on average much less than one photon for each pass through the wiggler, suggesting that the radiation must be treated quantum mechanically, and raising doubts as to whether this weak signal even contains sufficient phase information required for cooling, and whether it can be reliably amplified to provide the needed cooling on each pass. A careful treatment of the dynamics, where the radiation and amplification processes are treated quantum mechanically, indicates that fast cooling is in principle possible, with cooling rates which essentially agree with a simple classical calculation, provided that the effects of the unavoidable amplifier noise arising from quantum mechanical uncertainty are included. Thus, quantum mechanical uncertainties do not appear to present any insurmountable obstacles to optical cooling, but do establish limits on cooling rates and achievable emittances, so the effectiveness of such schemes will probably be limited by more prosaic classical concerns over the required laser power and phase control. Both these sources of noise might be expected to act more or less like having some effective number of extra classical particles in the sample, affecting the cooling rate and equilibrium emittances in degree but not catastrophically, but they would nevertheless impose a lower bound on the achievable effective sample size and ultimately limit the effectiveness of efforts to dilute the beam before cooling.

5.2 Stochastic Cooling: General Features and Considerations

In stochastic cooling [161, 162, 163], an interaction of a charged particle beam with a "pickup" measuring device generates a weak electromagnetic cooling signal containing partial information about the phase-space structure of the particles, ideally in a non-perturbative limit without any appreciable distortion of or back-action on the beam itself.

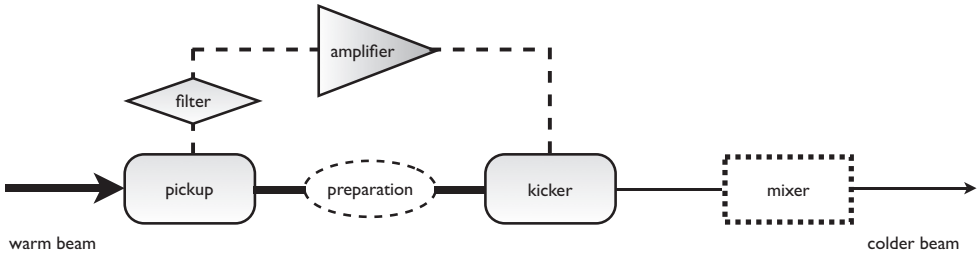

Figure 5.1. Schematic of an electromagnetically-based stochastic cooling scheme. A signal containing information about particle phase space deviations from the reference orbit are produced in the kicker, possibly filtered, then amplified, and fed back onto the beam in the pickup, possibly after suitable beam-optical preparations induced on the latter. If necessary, a mixer section is included between cooling passes to ensure that the incoherent heating effects contribute diffusively.

A general schematic is shown in Fig. 5.1. If, while the beam is properly diverted and prepared, this signal is then suitably amplified, manipulated, and then fed back and made to interact with the same beam, in some "kicker" section or device, then the overall dynamics of the reduced ($6D$) beam phase space are non-Liouvillian, and both the longitudinal and transverse beam emittance can be reduced. Stochastic cooling is non-evaporative, in the sense that it does not decrease phase space at the expense of removing the outlying particles; beam brightness can be increased as emittance is reduced.

In essentially all stochastic cooling schemes so far proposed or implemented, in which the perturbation to the particle phase space is approximately linearly proportional to some type of electromagnetic kicker fields, the damping time-scale τ_c for cooling of a particular degree of freedom, or equivalently the cooling rate τ_c^{-1}, at a particular time in the cooling process, can be at least roughly approximated by:

$$\tau_c^{-1}(t) \approx f_c \left[B(t)\sqrt{G(t)} - \frac{1}{2}A(t)G(t) \right], \tag{5.1}$$

where f_c is the frequency of passages through the cooling sections; $G(t)$ is the net power gain of the amplifier acting on the pick-up signal; $B(t)$ is a dimensionless positive parameter, depending on both the nature of the cooling scheme and on the current particle phase space distribution, which represents the cooling effects arising from interaction with the feedback signal in the kicker; and

$$A(t) = A_0 + A_1 \left(N_s + N_n \right) + \cdots , \tag{5.2}$$

represents the heating effects due to each particle's interaction with its own and neighboring particles' signals, where A_0 arises from self-fields and A_1 and higher-order terms are due to

fields from neighboring particles; N_n is a measure of extra noise introduced in the kicker signal by the amplifier, expressed as an equivalent number of extra particles; and N_s is known as the effective sample size, or number of samples, and represents the effective number of particles with whose kicker signal a given particle also interacts in addition to its own amplified self-field. Because the pick-up and amplifier system has finite bandwidth and therefore finite time-response, the signals of some number of neighboring particles always corrupt a given particle's signal, decreasing the effectiveness of cooling. In fact, the first term of (5.1) is the drift, or coherent term, and arises solely from the interaction of each particle with its own kicker signal, and which can be the only source of actual cooling; while the second term is the diffusive, or incoherent, contribution resulting primarily from amplifier noise and the interaction of a given particle from signals from the roughly N_s other particles on average in its sample, and contributes only to heating, as is apparent from its negative sign. (The incoherent term also includes a contribution from the self-field to account for the tendency of particles far out in the tails of the distribution of orbit deviations to receive too large of a correcting kick and over-shoot the target value.)

Because the incoherent heating contribution in (5.1) grows faster with amplifier gain than the coherent term, at any time there is then some locally optimal value of the gain,

$$G(t) \approx \left\{ \frac{B(t)}{A(t)} \right\}^2 \tag{5.3}$$

which maximizes the instantaneous cooling rate then given by

$$\tau_c^{-1}(t) \approx \tfrac{1}{2} f_c \frac{B(t)^2}{A(t)}. \tag{5.4}$$

Typically, this locally optimal gain will typically start relatively high when the beam is noisy and the coherent corrections are large, and then tend to decrease as the beam cools and approaches an asymptotic distribution with finite emittance in which the cooling and heating terms just balance. Note, however, because the range of possible subsequent cooling rates depend on the current particle phase space distribution, which in turn depends on the history of past cooling, locally optimal cooling rates do not necessarily result in global optima, i.e., in the fastest possible cooling. The overall optimal cooling history is an interesting but non-trivial problem in nonlinear control theory, left for future research.

5.3 Requirements On and Uses For Fast Stochastic Cooling

Both the achievable cooling rates and asymptotic equilibrium emittances therefore depend on the absolute power delivered by the feedback signal and on the relative power in

the usable "coherent" pickup signal from any single particle, which contains the phase space information necessary for cooling, compared to the corrupting "incoherent" signal arising from nearby particles or from noise in the amplifier, which actually contributes to heating during feed-back.

To increase the cooling rate and typically simultaneously decrease the equilibrium emittance achievable, the incoherent term represented by $A(t)$ must therefore be made smaller. From the form of $A(t)$ we can see that fast cooling times therefore require that the amplifier noise and effective sample size both be made as small as possible. The sample size will scale like

$$N_s \sim \rho_b \min\left[\pi\sigma_{b\perp}^2, S_c\right] v_0 \Delta\omega^{-1}, \tag{5.5}$$

where $\rho_b = \frac{N_b}{\pi\sigma_{b\perp}^2 L_b}$ is the spatial number density of particles in the bunch, $\sigma_{b\perp}$ is the transverse beam radius, L_b is the longitudinal beam length, $v_0 = c\beta_0$ is the mean longitudinal beam velocity, S_c is a measure of the transverse coherence area of the kicker fields, and $\Delta\omega$ is the limiting bandwidth of the pick-up/amplifier/kicker system.

Rapid cooling will therefore require relatively low beam densities and high-bandwidth, high-but-variable-gain, low-noise amplifiers. Existing stochastic cooling schemes relying on radio-frequency or microwave technology are limited by the O(GHz) bandwidths available for high-gain amplifiers at these frequencies , and typical cooling time-scales range from minutes to hours.

Yet much faster cooling might be necessary or desirable in some situations. For example, in order to achieve ultra-high luminosity proton beams, because phase space reduction can be offset by particle losses due to collisions, diffusion, etc, in the cooling ring at long time-scales. Typically time-scales for current radio-frequency (RF) stochastic cooling are $\tau_{\text{RF-SC}} \sim O(10^2 \text{ s})$ or $O(10^3 \text{ s})$, while with realistic technology OSC might achieve $\tau_{\text{OSC}} \sim O(1 \text{ s})$. For electron beams, OSC can work at lower energies where synchrotron damping is Inefficient. Rapid cooling for any proposed muon collider will be essential, because of finite lifetime of the muons; all stages of particle beam production, collection, collimation, acceleration, cooling, and experimentation must take place in only a few lab-frame decay times, which for muons at $O(10^2$ Gev) energies is only $O(1$ ms).

Such fast stochastic cooling, on microsecond time-scales, would require beam densities much lower than those typically achieved for particle beams of useful current with bunch sizes chosen for acceleration in RF structures, and would require amplifiers which can achieve high gain over very broad bandwidths with minimal added noise. Beam stretching and sub-

sequent compression can be used to suitably dilute the beam density during cooling and restore bunch sizes after cooling, and can be achieved in a stable, essentially reversible manner using conventional beam optics, but sufficiently broad gain bandwidths cannot be achieved with existing RF technology. Barring any unforeseen breakthroughs, fast cooling will require moving to optical wavelengths, where solid-state lasers amplifiers (such as those using Ti:Sapphire crystals) have achieved high gain over O(THz) bandwidths centered around $O(1\ \mu\text{m})$ wavelengths. Because of the high gain and high bandwidth achievable, optical stochastic cooling shows great promise, but also poses significant technological challenges. At such extremely fast time-scales, the pick-up signal cannot be manipulated electronically, but must be suitably collected, controlled, amplified, and directed into the kicker for feedback entirely through optical means; in order to reduce longitudinal emittance, transverse optical fields must be made to effect longitudinal momentum kicks, requiring very high gain; and particle beam optics must control particle positions within a fraction of an optical wavelength, presumably through some active monitoring and feedback. These pose important questions and difficult challenges, but none in themselves are feared to invalidate the possibility of OSC in principle. Here we focus primarily on a fundamental question of principle that has been raised, namely whether the wiggler signal from the beam in like regimes of operation contains adequate information to cool quickly or even cool at all, or instead whether it may be hopelessly corrupted by quantum "noise" or "fluctuations."

5.4 Why Consider a Muon Beam?

The possibility of a muon collider has received significant attention in the past decade. The muon is a fundamental particle that has been little studied at high energies in controlled experiments. In a muon-muon collider, physics similar to that studied in linear electron-positron colliders might be pursued, but because with a rest mass of $m_\mu = 105.7\ \text{MeV}/c^2$, the muon is ~ 207 times heaver than the electron, so synchrotron radiation is comparatively suppressed, and muons can be accelerated and stored in circular rings at high energies ($\sim O(10^3\ \text{GeV})$), as opposed to electrons or positrons which require linear accelerators. The higher mass also suppresses the so-called "beamstrahlung" effects which can lead to energy loss and/or energy spread, so larger bunch sizes and higher luminosities are in principle possible, while radiative corrections are smaller. In terms of their potential for particle creation, collisions between leptons are intrinsically more efficient than collision between particle like protons with significant internal structure. The larger mass of the

muon also translates into larger cross sections for certain interactions, especially for Higgs production, and precise measurements of the muon lifetime or $g-2$ factor, or searches for a muon electric dipole moment (EDM) or for lepton-flavor-violating decays offer promising avenues upon which to search for signatures of SUSY or other physics beyond the standard model.

The catch is that muons are unstable. They must be created through pion capture, so intense sources are expensive and produce initial beams with poor collimation (high transverse emittance) and large energy spread. A proper lifetime of only of $\tau_\mu \approx 2.2\ \mu$s, even with time dilation in the lab frame corresponding to relativistic factors of $\gamma \sim O(10^3)$, leaves only a matter of milliseconds to collimate, manipulate, cool, accelerate, and collide the muon bunch. This might in part turned to advantage by optimizing the ring not for muon collisions for the production of an highly-collimated, high-flux, terrestrially-based source of neutrinos via spontaneous decay of the muons.

In either mode of operation, as a muon-collider or a neutrino factory, the ring would pose many severe technological challenges, including the need for ultra-fast and intensive cooling, with a characteristic damping time $\tau_{\rm damp} \sim O(10^{-6}s)$ or less. Ionization-based rapid transverse cooling, as well as longitudinal cooling through shaped absorbers and/or emittance exchange, will be necessary, but transit-time optical stochastic cooling has been proposed as a possible way to boost final luminosity after the beam is already collimated and relativistic, beyond that which can be achieved by ionization cooling, which is limited by multiple scattering and trade-offs with longitudinal emittance.

5.5 Transit-Time Optical Cooling

Zolotorev, *et al.*[165, 166] have proposed and explored a possible method of ultra-fast transit-time optical stochastic cooling, in which both the pick-up and kicker consist of large-field magnetic wigglers; as shown schematically in Fig. 5.2. In the pick-up magnetic wiggler, Lorentz forces will produce transverse quiver motion of the charged beam particles, which in turn generates a small amount of spontaneous synchrotron radiation. This wiggler radiation is then collected and greatly amplified in a low-noise, solid state optical amplifier system and directed into the second wiggler. While the light is being amplified, the particles are directed through a bypass lattice, whose beam optics are designed so that each particle receives a time-of-flight delay proportional to the deviation of its longitudinal and/or transverse phase-space coordinates from the desired reference orbit. Particles then rejoin the amplified light

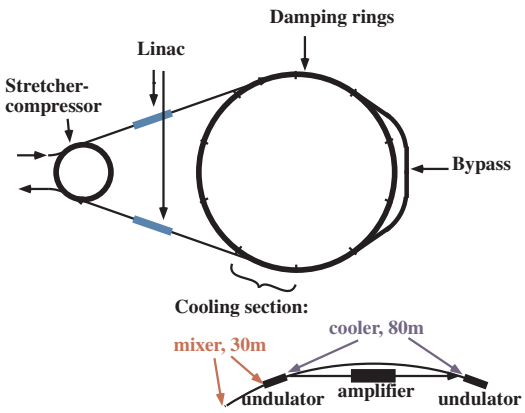

Figure 5.2. Schematic of a system for fast transit-time optical stochastic cooling. The stretcher/compressor ring, together with the linacs, expands the beam to lower the density before cooling, then reversibly compresses it after cooling. The cooling ring may contain one or more cooling sections.

in the kicker wiggler, where they again undergo transverse quivering with nearly the same polarization and at nearly the same frequency as the electric field of the optical radiation, resonantly exchanging some amount of energy, with the magnitude and sign of the net energy kick depending on the relative phase between particle quiver and field carrier oscillation.

Actual cooling can be effected only by the interaction of each particle with its own amplified field, and, if the transit-time in the bypass lattice is adjusted so that the delay relative to the self-field is proportional to the longitudinal momentum deviation, then the interaction in the kicker can produce a restoring force leading to reduced momentum spread. If particle time-of-flight is manipulated so as to also depend on a transverse betatron coordinates, then strong dispersion in lattice in the region of the kicker wiggler can also lead to transverse emittance reduction in that direction, and alternating the polarization of the wigglers or rotating the transverse phase space of the particle beam can then result in cooling the full transverse phase space.

Cooling is therefore critically sensitive to particle phase. Between the pickup and the kicker, particle positions must be carefully controlled, within a fraction of the micron-scale optical wavelength, presumably through active feedback on the particle beam optics, As cooling proceeds and deviations from the ideal orbit decrease, continued efficient cooling demands that the amplifier gain be decreased, and that the particle beam optics be adjusted in the bypass, if not continuously then at least intermittently, so that the probable range of particle deviations continues to be mapped into approximately one-half of an optical period of time-of-flight variation.

Between one cooling pass and the next, efficient cooling demands that the lattice must be designed to provide good mixing, or effective "randomization" of particles within each bunches, so that the incoherent signal arising from the sample particles neighboring any given particle becomes effectively randomized for each cooling kick. Because of the intrinsically low signal-to-noise ratios, stochastic cooling can really only work because the coherent drift term acts during every pass as a restoring force on average, tending to kick each particle towards the desired reference orbit on every pass, while the heating terms, while always present, ideally are more or less random from pass to pass, and therefore contributing in a diffusive fashion, with no bias in a particular direction, and with a net standard deviation accumulating only in proportion to the square root of the number of kicks.

However, if the net effect from all the sample particles is not independent from kick to kick, so a particle can experience a substantially similar incoherent signal for several passes, or even even partially correlated heating kicks in the same direction over many passes the resulting, then the heating can become super-diffusive, and cooling can be greatly slowed or even suppressed altogether.

Any given particle will be subject to the heating effects due to signals primarily from its neighboring sample of particles, whose size will scale inversely with the bandwidth of the cooling system. In most traditional stochastic cooling schemes, good mixing essentially requires that the identity of particles in a given particle's sample be randomized between each pass. To shuffle the make-up of each sample between each cooling pass, each particle should be shifted in position by a distance comparable to a sample length or more, while traveling from the kicker to the next passage through the pickup, in some manner that does not lead to persistent correlations between longitudinal particle position and the degrees-of-freedom undergoing cooling.

For the high-bandwidth, transit-time OSC method, this length is very short compared to the beam dimensions, about $L_s \sim N_u\lambda_0$,[1] so this kind of mixing should be even easier to achieve than in RF schemes, but its really more than is needed here. With transit-time optical stochastic cooling applied to highly relativistic beams of already moderately low emittance and energy spread, each particle will produce wiggler radiation of roughly the same envelope or spectrum, just with a different overall phase. The heating term for any given particle consists essentially of a non-stationary shot noise, the superposition of about N_s almost-identical wave-packets of duration $N_u\lambda_0$, with essentially random phases. Since

[1]This is what we will shortly recognize as the coherence length of the radiation, not the much longer wiggler length $N_u\lambda_u$.

the phase-space information useful for cooling is encoded in the phases of the signals, good mixing in this context does not really require that particles move among samples between kicks, but can be achieved merely by assuring that particle positions within a sample are effectively randomized, or equivalently that the relative distances between particles shift on the order of one optical wavelength λ_0 or perhaps a little more in a suitably "random" fashion, so as to randomize the phases of the approximately N_s wave-packets making up the incoherent kick signal in a given particle's sample. By "random," we here mean that, ideally, the shifts should be largely uncorrelated with relative longitudinal positions before the kick, but can actually depend on other, even cooled, degrees-of-freedom, such as transverse betatron coordinates or even longitudinal momentum. Because λ_0 is so short compared to the stretches of beam line between cooling sections available for mixing (typically $O(10\text{ m})$), this level of sample mixing should not be difficult to achieve, and in fact the greater challenge will be to suppress unwanted mixing on these optical length scales between the pickup and kicker. (Actually, if the longitudinal positions are initially uncorrelated with beam energy or betatron deviations, then the time-of-flight delays purposefully introduced in the bypass lattice between pickup and kicker during a single pass themselves can provide much of the needed mixing for the next pass.)

In order to sufficiently reduce the incoherent heating effects from neighboring particles, beam density is lowered before cooling, and then restored cooling. Even with high-bandwidth optical amplifiers and therefore relatively short sample lengths, typical densities for particle beams of interest are still much too large for microsecond-scale cooling with conceivable amplifier powers. RF acceleration requires reasonably short bunch lengths ($O(10\text{ cm})$ or so) for proper phasing, and high luminosity requires large bunch charge, $N_b \sim O(10^9)$ or higher if possible, so each bunch will enter the cooling section with rather high linear charge density, leading to unacceptably large sample sizes, on the order of $N_s \sim O(10^4)$ or greater. So before cooling, each bunch emerging from the acceleration sections must be greatly stretched, from tens of centimeters to a few hundred meters, so that so that cooling can take place at lower particle density and smaller sample sizes, say $N_s \sim O(10^2)$. After cooling, the bunch is de-compressed to restore luminosity. The reversible compression and stretching is effected using a linear accelerator (LINAC) and a specially designed lattice (see Fig. 5.2). Because the beam is highly relativistic, all beam particles travel at almost the same nearly-luminal velocity, but a distribution of relativistic momenta, so drift in free space will not efficiently expand the beam, but instead the beam may be first stretched by transport through a ring with very high momentum compaction

factor, such that particles with greater longitudinal momentum will travel along longer orbits and lag behind those with less momentum. Assuming such stretching separately conserves longitudinal action (clearly something of an approximation, since the compaction ring must correlate transverse coordinates with longitudinal momentum), this corresponds to a simple symplectic rotation in energy-duration beam phase space, so as it increases the beam length, it also introduces a head-to-tail energy chirp, i.e., a correlation between longitudinal intra-bunch longitudinal particle position and energy, and decreases the effective relative energy spread, having effectively transformed part of the random energy variation into an ordered energy correlation. The chirp would tend to impede particle mixing between cooling passes, and so is canceled by applying a suitably ramped current in an induction linac before the beam passes into the cooling ring proper. The beam stretching and energy compression is essentially completely Hamiltonian, analogous to the adiabatic expansion of a gas, and can be reversed after cooling, when a complementary energy chirp is introduced in the beam by another linac, and the beam is de-compressed to its original length in the same compaction ring, but now hopefully with reduced emittance and higher luminosity. For the muon cooling scenario, the stretching and compression phases would consume approximately half of the allotted cooling time, i.e., typically a few to several microseconds, but the benefits of beam-stretching more than make up for this extra time by increasing the achievable cooling rate.

5.6 Spontaneous Wiggler Radiation

The detailed dynamics of a transit-time optical cooling system will depend on the physics and form of spontaneous wiggler radiation. The central wavelength λ_0 of the wiggler radiation is downshifted from the wiggler period λ_u by relativistic effects; for a planar wiggler,

$$\lambda_0 \approx \frac{\lambda_u}{2\gamma_0^2}\left(1+\frac{a_u^2}{2}\right), \tag{5.6}$$

where $a_u = \frac{|q|B_u}{k_u mc^2}$ is the wiggler parameter, $k_u \equiv 2\pi/\lambda_u$, is the lab-frame wiggler wavenumber, B_u is the peak wiggler magnetic field, $\gamma_0 mc^2$ is the average energy of a beam particle and q is its charge. Given the average beam energy, the wiggler parameter and period are chosen to approximately match the central radiation wavelength λ_0 to the center of the gain bandwidth for the solid-state laser amplifiers, which is typically $O(1\ \mu\text{m})$. The homogeneous

"coherent" bandwidth[167] of such wiggler radiation is given by

$$\Delta\omega \approx \frac{1}{2N_u}\omega_0, \tag{5.7}$$

where $\omega_0 = ck_0 = 2\pi c/\lambda_0$ is the central radiation frequency and N_u is the number of undulator periods in the pickup wiggler. This so-called coherent component or coherent mode[2] of the radiation corresponds to a nearly diffraction-limited beam with angular spread

$$\delta\theta \approx \frac{\sqrt{1+\frac{1}{2}a_u^2}}{2\sqrt{N_u}\gamma_0}, \tag{5.8}$$

and spot size (waist)

$$w_0 \approx \frac{\lambda_0}{4\pi}\frac{1}{\delta\theta} = \frac{\sqrt{1+\frac{1}{2}a_u^2}}{4\pi}\frac{\sqrt{N_u}\lambda_u}{\gamma_0} \tag{5.9}$$

satisfying the optical version of the uncertainty principle.

Assuming that the light fields in the kicker are spatially coherent over the transverse extent of the particle beam, and the amplifier bandwidth $\Delta\omega_A$ centered on ω_A is matched to the coherent bandwidth $\Delta\omega$ of spontaneous radiation, centered on ω_0, the effective sample size N_s, meaning the average number of particles contributing appreciably to the incoherent signal experienced by any other particle, will scale like:

$$N_s \sim \rho_b\sigma_{b\perp}^2 L_s, \tag{5.10}$$

where $n_b = \rho_b\sigma_{b\perp}^2$ is the average number line density of particles in the bunch, and $L_s \sim N_u\lambda_0$ is the sample length length[3] over which particles will affect their neighbors. For fast cooling, the beam must be sufficiently stretched so that N_s is quite small compared to conventional stochastic cooling regimes, i.e.., $N_s \sim O(10^2)$ or so. In addition, the wigglers themselves will be optimized quite differently than for conventional applications. Most applications of wiggler radiation benefit from high coherence (narrow-bandwidth) spectra, and therefore rely on moderate (in the context of particle physics) energy beams traveling through wigglers of moderate wavelength but with many periods. In optical stochastic cooling, it is desirable to produce very broad bandwidth optical radiation from extremely

[2]Coherent here refers to radiation that is is approximately diffraction-limited transversely and approximately Fourier-limited longitudinally or temporally. The radiation from a single particle into this bandwidth and in this transverse mode is coherent in this sense, but the radiation from the bunch as a whole is incoherent in general, consisting of approximately $\frac{L_b}{N_u\lambda_0}$ longitudinal modes, essentially randomly-phased with respect to each other, where $N_u\lambda_0$ is the so-called coherence length, proportional to the inverse bandwidth.

[3]Actually, a particle trailing another particle by any distance will see almost none of the other particle's radiation, since the radiation is largely confined to a small forward angle by relativistic effects, and the radiation slips forward relative to the particles, while a particle leading another particle by about $N_u\lambda_0$ will interact with only a small fraction of that particle's radiation that slips sufficiently far ahead over the finite length of the kicker wiggler, so the effective sample length L_s will be somewhat shorter than $N_u\lambda_0$, as will be seen below.

relativistic beams, so the wigglers will consist of a relatively small number ($N_u \sim O(10)$) of long-period ($\lambda_u \sim O(50\text{ cm})$) magnets with field strengths essentially as high as is practical ($B_u \sim O(10\text{ T})$).

From classical radiation theory and Planck's law, the average number of photons emitted per particle into the coherent component of the wiggler radiation may be roughly approximated as

$$\mathscr{N}_{\text{ph}} = \alpha \tfrac{\pi}{2} \tfrac{a_u^2}{1+\frac{1}{2}a_u^2} \left[J_1\!\left(\tfrac{a_u^2}{4\left(1+\frac{1}{2}a_u^2\right)} \right) - J_0\!\left(\tfrac{a_u^2}{4\left(1+\frac{1}{2}a_u^2\right)} \right) \right]^2 \sim O(\alpha), \tag{5.11}$$

where $J_\ell(x)$ is the ℓth-order ordinary Bessel function, and $\alpha = \frac{e^2}{\hbar c} \approx \frac{1}{137}$ is the fine structure constant.

While the total power radiated as determined by the Larmor formula will be proportional to the number N_u of wiggler periods, the power radiated into the coherent mode remains constant (or nearly so), because the bandwidth scales inversely with N_u. The remaining power is radiated too far off-axis or at frequencies too far from the fundamental to be useful for cooling. Actually, equation (5.11) is not entirely accurate for $N_u < O(10)$, but in any case, for $\lambda_0 \sim O(1\ \mu\text{m})$, $N_u \sim O(10)$, and achievable magnetic field strengths corresponding to $a_u \sim O(1)$, the number of photons emitted per particle into the coherent mode should still be $O(\alpha)$.

5.7 "Naive" Quantum Mechanical Considerations

So optical stochastic cooling may poses serious technological challenges, at least for the very fast cooling required for muons, but in exploring the possibilities for such fast cooling based on wiggler radiation, serious concern arose over possible fundamental rather than merely practical limitations of this scheme. The apparent problem is that the actual cooling arises only through the interaction of each particle with its own wiggler radiation, which is extremely weak before amplification.

In any one pass through the pickup, each particle only radiates on average $O(\alpha) \sim 10^{-2}$ photons that can be collected, amplified, and fed back to actually effect cooling, the optical cooling signal from each particle will very weak and presumably may be subject to quantum mechanical effects. Yet naive quantum mechanical considerations raise fundamental doubts as to whether the pickup signal even contains the phase information needed for transit-time cooling, whether this information can be reliably amplified and extracted, and whether

quantum fluctuations in the incoherent signal from neighboring particles and arising in the optical amplifier itself lead to more significant heating than is accounted for classically.

5.7.1 Quantum Cooling Catastrophes?

With so few photons on average in the relevant cooling component of the pickup signal for any given particle, simple-minded quantum mechanical thinking suggests possible quantum catastrophes for cooling.

Do individual particles radiate at random in "quantum jumps"?

If, as in the standard treatment by Sands[168] of synchrotron radiation damping in electron storage rings, particles are assumed to radiate independently and at random, in a series of discrete, stochastic "quantum jumps" corresponding to Poissonian emission of a whole number of photons, at some average rate but at random times, and if the amplifier is imagined to faithfully multiply whatever photons are emitted, perhaps with the addition of some extra randomly-phased noise photons due to spontaneous emission or thermal noise,[4] then on average any one particle emits a photon only once in every $O(\alpha^{-1})$ passes through the pickup, while the neighboring particles in its sample emit on average a total of $O(N_s\alpha)$ photons, give or take $O\left(\sqrt{N_s\alpha}\right)$, So for $N_s \gtrsim O(10^2)$, one or more photons within a coherence length, randomly phased with respect to the particle in question will likely be present in the pickup signal and be amplified. It might seem that each particle will be subject to an appreciable heating kick on each pass, but usually experience no cooling kick whatsoever on most passes, but then every $O(\alpha^{-1})$ turns or so, suddenly receive a large cooling kick, comparable in magnitude to the typical heating kick per pass. Such stochastic discreteness effects would be expected to drastically lower the cooling rate compared to that calculated classically, or possibly even lead to unstable feedback preventing or suppressing cooling altogether.

[4]Note that the *independent* Poissonian photon emission model, in which all emissions are assumed uncorrelated, can predict any average photon number, the variance in photon number is always proportional to the mean. At least for the amplifier noise, we would imagine something closer to a so-called chaotic or thermal state, where due to interference effects the statistics are not those of shot noise, but rather standard deviation is proportional to the mean. This is the first hint that something might be wrong with or reasoning.

When individual particles do radiate, is the phase even well-defined?

Because the sample size N_s is quite small compared to that in most conventional stochastic cooling schemes, and the coherent signal intrinsically small, it also seems possible that the fluctuations in the pickup signal may no longer be dominated by the classical shot noise associated with random particle positions within the beam, but instead or additionally include quantum fluctuations of some sort. In particular, if the radiation emission from each particle consists of the occasional random emission of a photon in a "quantum jump" at some random time while the particle undulates in the pickup wiggler, then the phase of these photons would be expected to be very poorly determined. Even if the photon is perhaps more realistically considered to be emitted over some finite formation length, presumably the wiggler length itself e lab frame, in the relevant regime of small emission probability, this might seemingly make the phase uncertainty worse, not better. But in transit-time cooling schemes, the particle phase-space information used for cooling is encoded almost exclusively in the phase of the signal, so that even if the self-field is present, it seems that it might not carry the phase information necessary for transit-time cooling to work. Because the average photon number emitted by any one particle is small, the variation or uncertainty in this number is also small in absolute terms, and the "number-phase" Heisenberg uncertainty principle,

$$\Delta\mathscr{N}\Delta\phi \geq \frac{1}{2} \tag{5.12}$$

suggests that the phase[5] of these photons is very poorly determined. If photon emission is assumed to be at least approximately Poissonian, then $\Delta\mathscr{N} \approx \bar{\mathscr{N}}^{1/2}$, where $\bar{\mathscr{N}}$ is the average number of photons; but if $\bar{\mathscr{N}} \sim O(\alpha)$, the phase uncertainty must satisfy

$$\Delta\phi \geq \bar{\mathscr{N}}^{-1/2} \sim O\big(\alpha^{-\frac{1}{2}}\big) \gtrsim 2\pi. \tag{5.13}$$

But if essentially all of the particle phase-space information useful for cooling is contained in the optical phase of the pickup radiation, and if this is unavailable, cooling cannot occur.

What of spontaneous emission or thermal noise in the amplifier?

In addition to the discreteness effects in photon-emission and photon-phase noise, the amplifier is expected to add the equivalent one or more photons to the pre-amplified signal due to unavoidable spontaneous emission within the active gain medium, and perhaps as

[5] At this heuristic level, we can fortunately ignore the well-known technical difficulties associated with defining a self-adjoint phase operator in quantum mechanics conjugate to the usual Bosonic number operator.

well some thermal or other noise. These photons will be randomly-phased, so this source of noise acts more or less like having some extra number of particles N_n in the sample, in addition to, and for our parameters comparable or greater to, the actual number N_s. If the cooling does work despite our concerns, this may thereby affect the cooling rate in degree, but should not quell cooling altogether.

Stochastic Cooling in the Quantum-Jump Model

Fortunately, these fears of quantum effects catastrophically slowing or suppressing cooling will turn out to be misguided; the only source of quantum noise actually present is of the final sort mentioned: additive amplifier noise. But it will be informative to incorporate these considerations into very approximate quantitative estimates for the longitudinal (or rather, energy-spread) cooling rates with various naive quantum effects incrementally included for comparison, in order to make this intuitive reasoning more precise and to better understand later where it goes wrong.

The skeptical reader may suspect that these arguments are introduced merely as straw men, but they accurately represent the very concerns that motivated this investigation, and with some basis. As mentioned previously, a very similar discrete model of radiation by photon emission was used, and by all appearances quite successfully, by Sands[168] to treat radiation damping in a synchrotron ring, where quantum fluctuations offset damping, leading to finite equilibrium emittances, and is commonly used to analyze laser cooling of particle beams by Thomson scattering[169].

The energy kick given to the jth particle on a single pass is

$$mc^2\Delta\gamma_j = q\int_{t_j}^{t_j+\Delta t_j} dt\, \boldsymbol{E}\big(\boldsymbol{x}(t),t\big)\cdot\boldsymbol{v}_j(t), \tag{5.14}$$

where $\boldsymbol{E}$ is the transverse optical electric field seen by the particle in the kicker, $\boldsymbol{x}_j(t)$ is the particle's quivering spatial trajectory, governed predominately by the large magnetic field of the plane-polarized kicker wiggler, $\boldsymbol{v}_(t) = c\boldsymbol{\beta}_j(t) = \frac{d}{dt}\boldsymbol{x}(t)$ is its velocity, t_j is its arrival time at the front of the kicker wiggler, and Δt_j is the time spent inside.

As we are interested in a comparative scaling, we will use a simplified "back-of-the-envelope" model of longitudinal cooling where we assume all fields and trajectories oscillate sinusoidally, neglect end effects in the wigglers, diffractive and other transverse variation in the fields, dispersion and nonuniform gain in the amplifier, and various other details,

consider beam particles to be highly relativistic (i.e., $\gamma_j \gg 1$) but already relatively cold, (i.e., $|\gamma_j - \gamma_0| \ll \gamma_0$), and assume wiggler fields are large compared to radiation fields both before and after amplification. Keeping only the lowest order contributions in γ^{-1}, the (normalized) energy kick per pass can be estimated as

$$\Delta\gamma_j \approx \frac{qa_u N_u \lambda_u}{2mc^2 \beta_0 \gamma_0} \sqrt{G} \left[E_j \sin(\phi_{jj}) + \sum_{k \neq j} E_k \sin(\phi_{jk}) + \eta_j \right], \tag{5.15}$$

where $\beta_0 c$ and $mc^2\gamma_0$ are the average beam velocity and energy, respectively; E_j is the amplitude of the wiggler field produced by particle j before amplification, assumed to be a sinusoidal plane wave with a rectangular envelope corresponding to exactly N_u periods in the pickup wiggler; ϕ_{ij} is the relative phase delay or advance between the transverse quiver velocity of particle i and the amplified pickup radiation from particle j at the entrance to the kicker, as determined by their separation in the pickup and the bypass beam optics; G is the power gain of the amplifier, assumed constant over the relevant bandwidth for wiggler radiation, and η is a stochastic variable with vanishing mean, representing the additional amplifier noise arising from thermal effects and/or spontaneous emission. We have supposed that the spatial trajectory of each particle is classical, determined by the wiggler fields and initial conditions upon entering the cooling system, i.e., neither space charge nor other collective effects, nor the radiation fields appreciably perturb the spatial trajectories of the particles on any one pass, although by design the latter perturbs the energy of the particle. The sum in (5.15) is taken over some effective number N_s particles in particle j's sample, where to compensate for neglecting the assumed finite temporal extent of each particle's pickup radiation, such that any one particle is subjected to only part of the field arising from a neighboring particle while quivering in the kicker wiggler, this effective sample size will be somewhat smaller (i.e., by some numerical factor of $O(1)$) than the actual number of particles within a coherence length $N_u\lambda_0$ of the given particle. We have also assumed that after amplification the optical field can be treated classically, although it can retain the (amplified) stochastic fluctuations arising from its quantum origins.

Excluding phase noise in the radiated fields, ideally the phase-delay ϕ_{jj} is arranged through appropriate choice of bypass optics to be a function, ideally a nearly linear function, of the deviation $\delta\gamma_j$ in particle energy:

$$\phi_{jj} = \mu_1 \, \delta\gamma_j + \mu_2 \, \delta\gamma_j^2 + \ldots \tag{5.16}$$

where $\delta\gamma_j = \gamma_j - \gamma_0$ is the (normalized) energy deviation of the jth particle from the reference orbit, and the sign of μ_1 is chosen so that the coherent signal on average provides

a restoring force nudging the jth particle toward the reference energy. (If transverse phase space is also to be cooled, then the phase delay will have contributions proportional to the betatron errors as well. This is neglected here.)

The longitudinal cooling rate, or inverse cooling time-scale τ_c^{-1}, may be defined as the instantaneous rate at which RMS energy deviations are damped:

$$\tau_c^{-1} = -\tfrac{1}{2}\tfrac{d}{dt}\log\left\langle \delta\gamma_j^2 \right\rangle \approx -\tfrac{1}{2} f_c \left[\frac{\left\langle \Delta\gamma_j^2 \right\rangle + 2\left\langle \delta\gamma_j \Delta\gamma_j \right\rangle}{\left\langle \delta\gamma_j^2 \right\rangle} \right], \tag{5.17}$$

where f_c is the frequency of passes through the cooling system(s), $\Delta\gamma_j$ is the cooling kick as given by (5.15), and averages are performed over the current phase space distribution of the particles in the beam, and over any stochasticity in their radiation emission as well as over any noise from the amplifier.

Assuming purely classical emission, the self-field amplitudes E_j and phases ϕ_{jj} are deterministic functions of the single particle energy, (itself known only in a statistical or ensemble sense), while the phase differences ϕ_{ij} for $i \neq j$ are determined by shot noise describing the random particle positions within the beam. Further supposing the time-of-flight delays and amplifier gain are chosen to be locally optimal at all times during the cooling (not necessarily leading to the globally optimal minimum cooling time), and for simplicity taking particle energy deviations to be Gaussian in the lab frame, and assuming $N_s + N_n) \geq 1$, the longitudinal cooling rate becomes approximately

$$\tau_c^{-1} \approx f_c \frac{\frac{1}{e} - \frac{2}{e^5}\frac{1}{(N_s+N_n)^2}}{1 - \frac{1}{e^2} + (N_s+N_n) + \frac{4}{e^4}\frac{1}{(N_s+N_n)}}, \tag{5.18}$$

where N_n is proportional to the noise power $\langle \eta^2 \rangle$ in the amplifier, expressed in terms of an equivalent number of extra sample particles, and we have neglected corrections of order $O\left(\frac{1}{(N_s+N_n)^3}\right)$ or higher. Classically, N_n cannot vanish, but it could in principle be made arbitrarily small by sufficiently strong pumping and simultaneously cooling of the amplifier.

Instead, if we incorporate our naive quantum mechanical intuitions about emission fluctuations by imagining that the particles emit photons in a Poissonian fashion, but continue for the moment to ignore phase noise, then ϕ_{jj} is still regarded as deterministic (in the sense explained above) but E_j^2 may be taken to be proportional to a Poissonian random variable with some probability $p_{\text{rad}} \sim \alpha$ of photon emission per pass. Since $p_0 \ll 1$ in our regime, we can approximate each Poissonian random variable by a binomial variable, effectively neglecting the very rare emission of two or more photons, and we find after some algebra

that in this case the optimized cooling rate becomes

$$\tau_c^{-1} \approx p_{\rm rad}\, f_c \frac{\frac{1}{e} - \frac{2}{e^5}\frac{1}{(N_s+N_n)^2}}{1 - \frac{1}{e^2} + (N_s+N_n) + \frac{4}{e^4}\frac{1}{(N_s+N_n)}}, \tag{5.19}$$

which as expected is slower by a factor of $p_{\rm rad} \sim O\left(\alpha^{-1}\right)$ than the fully classical prediction. This slow-down occurs, of course, because the coherent cooling signal is by assumption only present on average in a fraction $\sim p_{\rm rad} \sim O(\alpha)$ of passages through the cooling section.

While throughout our simple analysis we have assumed linearity of the cooling kicks in the fields and particle deviations, in which case stochastic cooling can work at some rate regardless of the signal-to-noise-ratio, nonlinear effects can lead to instabilities beyond some finite range of signal-to-noise ratios, so it is also possible that this large multiplicative noise might not just drastically slow down cooling but frustrate cooling altogether.

Finally, to incorporate the phase noise in addition to the Poissonian emission into our model, it seems natural to also treat the ϕ_{jj} as random variables wit conditional means determined as above, but subject to random fluctuations, ostensibly with RMS deviations

$$\langle \delta\phi_{jj}^2 \rangle^{1/2} \sim \frac{1}{2\sqrt{N_{\rm ph}}} \sim O\left(\alpha^{-1/2}\right), \tag{5.20}$$

which is comparable to 2π. Taking these assumed phase fluctuations to be Gaussian for simplicity, the cooling rate becomes approximately

$$\tau_c^{-1} \approx p_{\rm rad} f_c e^{-\langle \delta\phi_{jj}^2 \rangle} \frac{1/e}{1 - e^{-2(\langle \delta\phi_{jj}^2 \rangle + 1)} + (N_s+N_n)}, \tag{5.21}$$

to leading order in the small quantities $e^{-\langle \delta\phi_{jj}^2 \rangle}$ and $\frac{1}{N_s+N_n}$. With such phase uncertainty, the cooing rate would be suppressed by more than a factor of $e^{-\langle \delta\phi_{jj}^2 \rangle} \sim O(10^{-3})$, because the cooling information carried by the phase of the self-fields has been corrupted.

Finally, quantum mechanics will enforce a lower bound for the amplifier noise $N_n \approx \frac{1}{2}\langle \eta(t)^2 \rangle$ due to unavoidable spontaneous emission in the pumped medium responsible for the gain. Various quantum mechanical and semi-classical arguments (see for example, [170] for a survey) suggest a minimum amount of added noise equivalent in its final effects to one-half photon per mode entering the amplifier along with any actual signal present and amplified along with it. With our simple model of windowed plane waves, and an effective interaction length in the kicker (accounting for slippage between the particles and fields) equal to the coherence length $N_u\lambda_0$ of the radiation, each particle effectively interacts with a single mode, so that one would predict that at best $N_n \sim \frac{1}{2}\frac{1}{N_{ph}} \sim O(\alpha^{-1})$. While not appearing as a multiplicative slow-down in the cooling rate and therefore not as devastating as the other

possible quantum effects described above, such spontaneous emission noise does indicate that at small sample sizes quantum fluctuations may be as important a contribution to the incoherent heating as classical shot noise from actual particles, and that these fluctuations will ultimately limit the cooling gains achievable by diluting or stretching the beam. So indeed we can we can roughly account for the spontaneous emission by increasing the effective sample size from N_s to $N_s + N_n$ for some effective noise number $N_n \sim N_A/N_{\rm ph}$, where $N_{\rm amp}$ is a measure of the effective added noise by the amplifier expressed as an equivalent number of photons at the amplifier front-end (i.e., prior to amplification.)

5.8 Towards a More Careful Treatment of Quantum Effects

Fortunately, a more careful quantum mechanical analysis will reveal that these naive intuitions are incorrect, and that in effect only the additive amplifier noise is present, so that cooling rates are accurately estimated by classical continuous emission results, provided allowance is made for the unavoidable amplifier noise, which is ultimately quantum mechanical in origin, and cannot be made arbitrarily small without violating the Heisenberg Uncertainty Principle or the unitary nature of quantum dynamics. That is, in a more careful quantum mechanical analysis neither the multiplicative emission-noise leading to the $O(\alpha)$ slow-down in the naive Poissonian-emission model (5.19), nor the catastrophic slowdown appearing in the naive phase-noise model (5.21) actually occurs; in effect, only additive noise arising from spontaneous emission or thermal effects in the amplifier appears, leading to a cooling rate of the approximate form (5.18) for some effective noise number N_n, whose minimum value is essentially constrained by the uncertainty principle.

A more rigorous treatment of OSC will force us to examine carefully each stage of the cooling dynamics: the particle motion in the pickup wiggler, the resulting radiation emitted, the quantum mechanics of the optical amplification process, the particle-radiation interaction in the kicker wiggler, and the resulting changes in the beam phase-space distribution. Because we are here interested in a proof-of-principle question rather than a detailed design assessment, we will make a number of simplifying assumptions, yet still incorporate the essential classical, quantum, and statistical physics of the processes so as to address the fundamental question of whether optical stochastic cooling based on amplification of small pickup signals is intrinsically flawed due to the effects of quantum noise, and ultimately arrive at a less pessimistic answer.

A fully self-consistent quantum mechanical (or worse, QED) treatment of beam parti-

cles, radiation fields, amplifiers and other optical elements would be prohibitively difficult, but fortunately is not necessary either. Rather than explaining the quantum mechanical features of the dynamics of the beam particles, we will carefully explain them away, arguing that particles in the beam can be treated classically in their interaction with the wiggler, radiation, and any external focusing fields.

Then we will verify that such particles do not radiate in to photon number states, but rather Glauber coherent states, which are actually the states closest to classical radiation fields allowed by quantum mechanics, and quite different in their statistics from the states consisting of whole numbers of photons. Making certain idealizations, The radiation essentially retains this form as it passes through dielectric optic elements.

Of course, the inverted population of atoms in the active medium of a laser amplifier behaves in a highly non-classical manner, but the precise dynamics of the amplifier need not be evolved explicitly. Very general considerations of amplifier action will be sufficient to determine the action of the amplifier on the pickup field in an "input-output" formalism' where the specifics of the intermediate dynamics can be neglected, and to characterize the best-case limits on the additional noise introduced of the amplification process without resorting to an explicit microscopic model of the amplifying medium.

Once amplified, the radiation behaves entirely like a classical field, both in its statistics and its interaction with the beam particles in the kicker.

So once these results are established, a simple estimate of the cooling rates can be made essentially along classical lines, just accounting for any extra noise in the field due to amplified spontaneous emission.

5.9 Particle Dynamics are Classical

For beams of electrons, muons, or protons, at relevant energies and emittances, and with realistic wiggler strengths, the de Broglie wavelengths associated with the particles' longitudinal and transverse motion are extremely small compared to the radiation wavelength, wiggler period, beam dimensions, and other relevant scales, so we will argue that the particles can be treated as classical point particles obeying classical relativistic kinematics. The subsequent analysis of the radiation will be greatly simplified by assuming the dynamics of the particles in the pickup wiggler to be classical, and in fact with prescribed

classical trajectories determined by external fields only, so the effort needed to carefully establish and justify the accuracy of these assumptions will be subsequently rewarded.

Clearly, in order to adequately describe particles classically, all statistical and dynamical manifestations of quantum-mechanical or quantum-electrodynamical effects on the particle degrees-of-freedom must be negligible. Often, statistical and dynamical effects are not clearly distinguished, but it not difficult to find regimes where either the classical statistical limit or classical dynamical limit may be valid, but not both. However, it is often the case for a system consisting of many particles that classical statistical noise can also tend to mask the dynamical manifestations of quantum noise. Both limits of course are related in part to the size of the typical de Broglie wavelength λ_{dB} associated with a particle, but ignoring quantum statistical effects typically requires that λ_{dB} be small compared to average inter-particle spacings, while ignoring quantum dynamical effects typically requires that λ_{dB} remain small compared to all relevant dynamical length scales. For example, in a plasma or ionized gas at rest (on average), if $\lambda_{dB} \sim \frac{h}{\sqrt{mk_{\rm B}T}} \ll n^{-\frac{1}{3}}$, where n is the particle density, T is the temperature, and $k_{\rm B}$ is Boltzmann's constant, then quantum statistical effects can be ignored and the particles can be taken to satisfy Maxwell-Boltzmann statistics, as if they were classical point particles, whether they are identical Fermions or Bosons. But unless $\lambda_{dB} \ll \ell_{\rm L}$ where $\ell_{\rm L} \sim \frac{q^2}{k_{\rm B}T}$ is the Landau length, or typical distance of closest approach between two plasma particles, then the Coulomb scattering between particles should be treated quantum mechanically.

5.9.1 Quantum Statistical Degeneracy

In the case at hand of a relativistic, charged particle beam of indistinguishable Fermions, the safe neglect of quantum statistical effects requires that the beam be non-degenerate in the average rest frame:

$$(\rho_b')^{-\frac{1}{3}} \gg \lambda_{dB}', \tag{5.22}$$

where $\lambda_{dB}' \sim \frac{h}{\sqrt{mk_{\rm B}T'}}$ is the typical thermal deBoglie wavelength, T' is an effective temperature, which may be different for the longitudinal and transverse degrees-of-freedom, and m is the invariant particle rest mass. We will assume throughout that the average beam energy $mc^2\gamma$ and beam energy spread $mc^2\delta\gamma$ satisfy $\gamma \gg 1$ and $\frac{\delta\gamma}{\gamma} \ll 1$, even before stochastic cooling is applied. Since $a_u \sim O(1)$, this also implies that the longitudinal momentum $\tilde{p}$ inside the wiggler is only negligibly smaller than the average beam momentum $p = mc\gamma\beta$ outside the wiggler. Some of the longitudinal beam kinetic energy is converted

into energy of transverse quiver in the wiggler magnetic field, but assuming that transverse canonical momentum is conserved in the presumed transversely-uniform wiggler fields, and that particles enter the wiggler approximately on axis, one finds

$$\frac{\tilde{p}}{p} \sim 1 - \frac{1}{4}\frac{a_u^2}{\beta^2\gamma^2}, \tag{5.23}$$

while by assumption $\beta \sim 1$ and $\frac{a_u}{\gamma} \ll 1$, so we will ignore this distinction when convenient.

Since motion in the wiggler field will not significantly effect the beam temperature, average density, or energy, for simplicity we can analyze the beam while it drifts freely, prior to its entering the pickup wiggler. In this limit of high energy and relatively low energy spread, one finds from relativistic velocity addition that the rest-frame longitudinal temperature can be expressed as

$$k_{\mathrm{B}}T'_{\parallel} \sim mc^2\beta'^2_{\parallel} \sim mc^2\frac{1}{\beta^2\gamma^2}\left(\delta\gamma\right)^2, \tag{5.24}$$

where βc is the average beam velocity in the lab frame, with $\beta \sim 1$, and $\delta\gamma \approx \frac{\delta p_{\parallel}}{mc}$ is the lab-frame beam energy spread. Since momenta perpendicular to a Lorentz boost remain unchanged, the transverse rest-frame temperature can be written as

$$k_{\mathrm{B}}T'_{\perp} = \frac{\left\langle p'^2_{\perp}\right\rangle}{2m} \approx \frac{\delta p_{\perp}^2}{m}. \tag{5.25}$$

where $\delta p_{\perp}$ is the root-mean-squarer transverse momentum spread in the lab frame, related to the conventional transverse beam emittance $\epsilon_{\perp} \sim \frac{\sigma_{\perp}\delta p_{\perp}}{mc\beta\gamma}$, where $\sigma_{\perp}$ is the transverse (radial) beam size. From Lorentz contraction, it follows that the particle density must transform as $n'_b = n_b/\gamma$, so the conditions for non-degeneracy, written in terms of lab-frame quantities, become

$$\rho_b \ll \gamma\frac{1}{\lambda_c^3}\left[\frac{\delta\gamma}{\gamma}\right]^3 \sim \left(\frac{\epsilon_{\perp}}{\lambda_c}\right)^3\frac{\gamma}{\sigma_{\perp}^3}, \tag{5.26}$$

and

$$\rho_b \ll \gamma\frac{\delta p_{\perp}^3}{h^3} = \gamma\frac{1}{\lambda_c^3}\left[\frac{\delta p_{\perp}}{mc}\right]^3 \tag{5.27}$$

where $\lambda_c = \frac{h}{mc}$ the Compton wavelength. For achievable beam densities and emittances, even after cooling, these conditions are typically easily achieved by many orders of magnitude; existing particle beams are very far from being degenerate, and quantum statistical effects arising from the Pauli Exclusion Principle can be completely ignored.

5.9.2 Pair Creation or other QED Effects

Although the particles are assumed to be highly relativistic in the lab frame, they are all streaming in the same general direction, with small relative variations in momentum, so

particle-scattering will involve small center-of-momentum energies, and hence particle/anti-particle pair creation or other exotic QED scattering effects should be negligible. In the average rest frame, this requires

$$mc^2\langle\beta'^2\rangle \ll 2mc^2, \tag{5.28}$$

equivalent in the lab frame (after dropping some factors of two) to the requirement that

$$\frac{\delta\gamma}{\gamma} \ll 1, \tag{5.29}$$

which has already been assumed. Of course, muons will spontaneously decay in a random, Poissonian fashion, which is a completely non-classical process mediated by electroweak interactions, but the resulting electrons will be lost from the beam as soon as it is bent in magnetic fields calibrated for the heavier muons, so this process can be ignored for particles which remain in the beam throughout the full cooling process, requiring several turns in the cooling ring.

5.9.3 Spin Effects

Since spin degrees of freedom are intrinsically quantum mechanical, at least for non-composite particles such as muons or electrons,[6] all spin effects should be negligible in a truly classical treatment. The energy associated with the spin of an elementary fermion in the wiggler field B_u is $U_{spin} \sim \frac{q\hbar}{2mc}B_u = \frac{1}{2}a_u\hbar k_u c$. In order to safely neglect spin effects, this energy U_{spin} should be smaller than all other relevant energy scales, including: the total particle particle energy γmc^2; particle quiver kinetic energy, which for $\gamma \gg 1$ and $a_u \sim O(1)$ is about $\frac{1}{2}\gamma m v_\perp^2 = \frac{1}{2}\gamma m\left[c\frac{a_u}{\gamma}\right]^2 = \frac{1}{2}mc^2\frac{a_u^2}{\gamma}$; the mean photon energy $\hbar\omega_0$; and the total radiated energy per particle, U_{rad}, which is larger by about $O(N_u)$ than the power radiated into the coherent mode. Using the relativistic Larmor formula, the total radiated energy per particle in the pickup wiggler can be estimated as

$$U_{rad} = P_{rad}\Delta t = \frac{2}{3}\frac{q^2}{m^2c^3}\frac{dp^\mu}{d\tau}\frac{dp_\mu}{d\tau}\frac{N_u\lambda_u}{c} \approx \frac{q^2}{c}\gamma^4\dot{\beta}^2\frac{N_u\lambda_u}{c} \approx \alpha N_u a_u^2\hbar\omega_0. \tag{5.30}$$

[6] While the energy of a spin state can be made arbitrarily large by increasing the magnetic field, the *action* associated with any spin states of a single lepton is bounded in magnitude by $\frac{\sqrt{3}}{2}\hbar$ is bounded so cannot participate in an sort of Correspondence Principle limit. See the discussion in the sequel.

These requirements imply, respectively, that

$$\hbar\omega_0 \ll mc^2\frac{\gamma^3}{a_u}; \tag{5.31a}$$

$$\hbar\omega_0 \ll \gamma a_u mc^2; \tag{5.31b}$$

$$\gamma^2 \gg a_u; \tag{5.31c}$$

$$\gamma^2 \gg \frac{1}{\alpha N_u a_u}. \tag{5.31d}$$

Since $\hbar\omega_0 = \frac{\lambda_c}{\lambda_0}mc^2$, the first two conditions can be written equivalently as $\lambda_c \ll \frac{\gamma^3}{a_u}\lambda_0$ and $\lambda_c \ll \gamma a_u \lambda_0$. For optical or near infrared frequencies, a wiggler parameter $a_u \sim O(1)$, $N_u \sim O(10)$ wiggler periods, and $\gamma \sim O(10^2)$ or $O(10^3)$, these conditions are all easily satisfied.

To be thorough, we should examine not just energies but forces associated with spin degrees of freedom. As pointed out by Bohr and Mott as early as 1929 [171], consistency of the Correspondence Principle demands that any non-composite particle whose spatial degrees of freedom are behaving classically should also exhibit no observable spin effects, because the magnitude of the spin and associated action is bounded, and there is therefore no way to consider its classical (high-action) limit independently from that of the spatial degrees of freedom. Evidently, this suppression of spin effects must be enforced by a "conspiracy" between the Heisenberg Uncertainty Principle and the Maxwell-Lorentz equations.[7] In particular, for charged elementary Fermions, the magnetic field produced by the spin of one particle and felt by another particle will be swamped by the magnetic field produced by the moving charge associated with the first particle. At a characteristic distance $r' \sim n'^{-\frac{1}{3}}$, in the average rest frame, the magnetic field due to one spin is roughly $B'_{\text{spin}} \sim \frac{q\hbar}{2mc}\frac{1}{r'^3}$, while the magnetic field due to the same particle's motion will scale like $B'_v \sim \frac{qv'}{c}\frac{1}{r'^2}$. If the uncertainty in position $\delta r'$ satisfies $\delta r' \ll r'$, consistent with the particle's spatial degrees of freedom behaving classically, then it can easily be shown that $B'_{\text{spin}} \ll \delta B'_{v'}$, where $\delta B'_{v'}$ is the uncertainty in the Biot-Savart field $B'_{v'}$ due to the Heisenberg uncertainty in the velocity $\boldsymbol{v}'$. Therefore, spin-spin effects will be negligible in a non-degenerate beam of charged leptons.

Although ideally the magnetic field of the wiggler is taken to be transverse, transversely uniform, and plane polarized, such spatial variation is not strictly consistent with Maxwell's equations, which demand that the wiggler magnetic fields are both curl-free and divergence-free in the neighborhood of the beam line, requiring in addition to the principle sinusoidal

[7] Incidentally, this makes the prospects for using optical stochastic cooling to perform beam polarization look rather bleak. Actually, partially polarized beams would seem to offer a counter-example to the Bohr-Mott Conjecture, except that only collective spin observables on many particles are ever really measured, which can behave classically.

transverse component at least some small longitudinal component and some transverse gradients with scale-length $O(\lambda_0)$ in the neighborhood of the beam axis. With such self-consistent field profiles, it can be directly shown that the transverse component of the dipole force on the spin, given by, $\boldsymbol{F}_{dp} = \frac{g\hbar}{2mc}(\boldsymbol{s}\cdot\boldsymbol{\nabla})\boldsymbol{B}_u$, will be small compared to the transverse component of total Lorentz force on the charge, which is given by $\boldsymbol{F}_L \frac{q}{c}\boldsymbol{v}\times\boldsymbol{B}_u$, if

$$\delta v_\perp \ll v_z, \tag{5.32a}$$

$$\hbar k_0 \ll \gamma^2 \beta mc, \tag{5.32b}$$

where $\delta v_\perp \approx \frac{1}{\gamma m}\delta p_\perp$ is the transverse velocity spread, and $k_0 = \frac{2\pi}{\lambda_0}$ is the central optical wavenumber; while the longitudinal dipole force on the spin will be negligible compared to the longitudinal component of the Lorentz force, provided

$$\hbar\omega_0 \ll \gamma mc^2, \tag{5.33a}$$

$$\frac{1}{\gamma}N_u^2 a_u \lambda_c \ll \lambda_0, \tag{5.33b}$$

$$\hbar k_0 \ll \frac{1}{N_u a_u}\gamma^3 \beta mc, \tag{5.33c}$$

all readily satisfied. Spin degrees-of-freedom are here unimportant to all aspects of the particle dynamics.

5.9.4 Transverse Motion

Next, we argue that quantum mechanical effects in the transverse motion are unimportant. In order that this be the case, the spread in a particle's transverse wavefunction must remain small relative to other relevant length-scales, including the transverse beam size, the amplitude of the transverse quiver motion, the range of transverse variation of the wiggler field, as well as the optical wavelength. Also, the transverse velocity fluctuations associated with quantum mechanical uncertainty demanded by the Heisenberg uncertainty principle must remain small compared to the transverse quiver velocity and to the transverse velocity spread in the beam. Neglecting any initial transverse particle velocity and any transverse spatial variation in the wiggler fields, conservation of canonical momentum implies that the transverse quiver velocity will be $v_\perp \sim c\frac{a_u}{\gamma}$. Since we assume $\frac{a_u}{\gamma} \ll 1$, the transverse quiver motion can be taken to be non-relativistic, provided we replace the rest mass m everywhere with the "effective" relativistic mass γm to account for the increase in inertia arising from the high longitudinal velocity in the lab frame. Furthermore, we will neglect the transverse forces due to the wiggler fields or external beam-optical fields.

Because classically these fields will provide focusing (at least in one direction) on average, semi-classical considerations suggest that they will act mostly to inhibit the spread of the wavefunction, so ignoring them altogether should constitute a conservative approximation.

To get an idea of the scaling, we suppose each particle is described by a Gaussian wavepacket, initially (at $t = 0$) separable in the transverse (x and y) coordinates, and having some minimum initial variance in each component of position and momentum consistent with the Uncertainty Principle, but then freely streaming during the total interaction time $\Delta t \approx \frac{N_u \lambda_u}{\beta c} \sim \frac{N_u \lambda_u}{c}$ in the wiggler. The dynamics for the different transverse directions are identical, so we need follow only one component, say in the $\hat{\boldsymbol{x}}$ direction. The variance in position at a subsequent time t will be given by

$$\Delta x^2(t) = \Delta x^2(0) + \tfrac{1}{4}\tfrac{\hbar^2 t^2}{\gamma^2 m^2}\Delta x^2(0) = \Delta x^2(0) + \tfrac{\Delta p_x^2(0) t^2}{\gamma^2 m^2}, \tag{5.34}$$

while the variance in momentum,

$$\Delta p_x^2(t) = \Delta p_x^2(0). \tag{5.35}$$

is constant, since transverse forces have been neglected, and we have further assumed that the initial conditions satisfy

$$\Delta x(0)\Delta p_x(0) \approx \frac{\hbar}{2}. \tag{5.36}$$

By differentiating the expression for $\Delta x^2(t)$ with respect to $\Delta x^2(0)$, one finds that the spatial variance is minimized at any subsequent time $t > 0$ by

$$\Delta x^2(t) = \tfrac{1}{2}\tfrac{\hbar}{\gamma m} t. \tag{5.37}$$

Such a wave-packet with minimal spatial spread is the closest thing to a classical point-particle allowed by quantum mechanics, and evaluating the spatial variances at $t = \Delta t$, we have

$$\Delta x^2(\Delta t) \sim \tfrac{1}{4\pi}\tfrac{N_u \lambda_c \lambda_u}{\gamma}. \tag{5.38}$$

In order that the particle behavior remain classical during this interaction time, this variance should be sufficiently small in several senses: it is necessary that this transverse spatial spread be small compared to the transverse beam dimensions, i.e., that $\Delta x(\Delta t) \ll \sigma_\perp$, or

$$\lambda_c \lambda_u \ll \tfrac{4\pi}{N_u}\gamma\sigma_\perp^2; \tag{5.39}$$

that it be small compared to the transverse extent of individual particle orbits, i.e., that $\Delta x(\Delta t) \ll c\frac{a_u}{\gamma}\frac{\lambda_u}{c}$, or

$$\lambda_c \ll \tfrac{4\pi}{N_u}\tfrac{a_u^2}{\gamma}\lambda_u; \tag{5.40}$$

that it be small compared to the *transverse* range of variation for the wiggler magnetic field, which from Maxwell's equations (specifically, $\boldsymbol{\nabla} \times \boldsymbol{B} = \boldsymbol{0}$) will be comparable to the longitudinal length-scale of field variation, namely λ_u; i.e., that $\Delta x(\Delta t) \ll \lambda_u$, or

$$\lambda_c \ll \tfrac{4\pi}{N_u}\gamma\lambda_u; \tag{5.41}$$

and finally, in order that the particle's phase in the radiation field be well-defined, we require that the phase uncertainty associated with this transverse uncertainty in the presence of wavefront curvature remain small: $k_0\,\delta\theta\Delta x \ll \pi$, or roughly

$$\lambda_c \ll \tfrac{\pi}{2}\lambda_0. \tag{5.42}$$

For sufficiently long wigglers ($N_u \gg 10$), some of these conditions could be violated, but in that case a more sophisticated analysis including transverse focusing effects would be needed to assess the classicality of the particle orbits.

So far we have followed the conventional reasoning, but we actually need to be more careful, because up to now we have only considered spatial spread, but in fact the wavepacket that minimizes transverse spatial variances has arbitrarily large variance in transverse momenta as $t \to 0$, violating the requirement that quantum mechanical transverse momenta uncertainties should also be small compared to the transverse momentum spread in the beam and to the wiggler-induced quiver momentum. It turns out that with a lengthy calculation for an appropriate Gaussian wavepacket, the uncertainties in both transverse position and momentum can be made sufficiently small simultaneously provided the above conditions remain satisfied, and in addition

$$\tfrac{\lambda_c}{\lambda_0} \ll 4\pi\tfrac{\delta p_\perp}{mc}, \tag{5.43a}$$

$$\tfrac{\lambda_0}{\lambda_c} \ll 4\pi a_u. \tag{5.43b}$$

which are readily satisfied.

Including the effects of the wiggler fields should not change these conclusions, provided the wiggler field itself acts classically. In the average rest frame, the wiggler fields can be treated via the Weizsäcker-Williams approximations as a traveling electromagnetic plane wave consisting of virtual photons, which then scatteroff the particles to produce the real wiggler radaition. Over length of the wiggler and the transverse area of the coherent radiation, the number of virtual virtual wiggler photons is very large:

$$\alpha^{-1} N_u^2 \tfrac{a_u^2}{\gamma^2}\tfrac{\lambda_u^2}{\lambda_c^2} \gg 1. \tag{5.44}$$

Actually, even over the much smaller volume defined by the wiggler length, the transverse extent of the particle's quiver, and the classical muon radius r_μ, the number of virtual wiggler photons is still large:

$$N_u \frac{a_u^3}{\gamma} \frac{\lambda_u}{\lambda_c} \gg 1. \tag{5.45}$$

Additionally, by definition the wiggler field is very coherent, with "quantum" uncertainties in the virtual photon number density very small in a relative sense, and uncertainties in phase very small compared to 2π. One could not demand more from a classical field in what is really a quantum mechanical world.

5.9.5 Longitudinal Dynamics

In a similar manner, we can show that the longitudinal dynamics can be taken to be classical, although unlike the transverse motion, the longitudinal motion is highly relativistic. For the moment, we will assume that the particles are freely streaming longitudinally, unperturbed by wiggler, space-charge, radiation, or other fields. Because spin, ZBW, and other relativistic quantum effects can be neglected, we can fortunately forego use of the Dirac or even Klein-Gordon equations, and simply use a one-dimensional Schrödinger equation with the positive-energy branch of the relativistic dispersion relation:

$$\hbar\omega(k) = \sqrt{m^2c^4 + c^2\hbar^2k^2} - mc^2. \tag{5.46}$$

We assume the initial state is a minimum-uncertainty Gaussian wavepacket:

$$\psi(z,\, t=0) = \tfrac{1}{\sqrt{\sqrt{2\pi}\Delta z(0)}} e^{i\bar{k}z} e^{-\frac{z^2}{4\Delta z^2(0)}}, \tag{5.47}$$

where $\Delta z(t)^2$ is the longitudinal spatial variance at time t, and $\hbar\bar{k} = mc\gamma\beta$, is the average longitudinal beam momentum. Taking a Fourier transform, this can be written as

$$\psi(z,\, t=0) = \frac{1}{\sqrt{2\pi}} \int dk\, \psi(k) e^{ikz}, \tag{5.48}$$

where

$$\psi(k) = \left[\tfrac{2\Delta z(0)}{\sqrt{2\pi}}\right]^{\frac{1}{2}} e^{-\left(k-\bar{k}\right)^2 \Delta z^2(0)}. \tag{5.49}$$

Using the dispersion relation (5.46), the freely-propagating solution at later times can then be written as

$$\psi(z,\, t) = \tfrac{1}{\sqrt{2\pi}} \int dk\, \psi(k) e^{ikz - i\omega(k)t}. \tag{5.50}$$

Since $\frac{\delta\gamma}{\gamma} \ll 1$, the relativistic dispersion relation (5.46) can be expanded in a Taylor series about $\bar{k}$:

$$\begin{aligned}\omega(k) &\approx \tfrac{mc^2}{\hbar}\left[\gamma - 1\right] + c\beta\left(k - \bar{k}\right) + \tfrac{1}{2}\tfrac{\hbar}{\gamma^3 m}\left(k - \bar{k}\right)^2 \\ &= \tfrac{mc^2}{\hbar}\left[\tfrac{1}{\gamma} - 1\right] + c\beta k + \tfrac{1}{2}\tfrac{\hbar}{\gamma^3 m}\left(k - \bar{k}\right)^2 + \ldots .\end{aligned} \tag{5.51}$$

After a little algebra, we find the longitudinal variances to be

$$\Delta p_z^2(t) = \Delta p_z^2(0) = \tfrac{1}{2}\tfrac{\hbar^2}{\Delta z^2(0)}, \tag{5.52}$$

and

$$\Delta z^2(t) \approx \Delta z^2(0) + \tfrac{1}{4}\tfrac{\hbar^2 t^2}{m^2\gamma^6 \Delta z^2(0)} = \Delta z^2(0) + \tfrac{\Delta p_z^2(0) t^2}{m^2\gamma^6}. \tag{5.53}$$

Differentiating, we find that the spatial spread is minimized at any subsequent time $t > 0$ by the choice

$$\Delta z^2(t) = \tfrac{1}{2}\tfrac{\hbar t}{m\gamma^3}. \tag{5.54}$$

Over the interaction time $\Delta t \approx \frac{N_u \lambda_u}{c}$, this spread $\Delta z(t)$ must remain small compared to the smallest length-scale associated with variations in longitudinal forces experienced by the particle, namely the optical wavelength λ_0, so that the particle can behave like a point particle and can be taken to have a well-defined phase in the optical field. That is, classical behavior requires $\Delta z(\Delta t) \ll \lambda_0$, or

$$\tfrac{N_u}{4\pi}\tfrac{\lambda_c \lambda_u}{\gamma^3} \ll \lambda_0^2. \tag{5.55}$$

Since, $\lambda_u \sim \gamma^2 \lambda_0$, this condition is usually written as

$$\tfrac{N_u}{4\pi}\tfrac{\lambda_c}{\gamma} \ll \lambda_0. \tag{5.56}$$

In the FEL literature, this condition (5.56) is often the single requirement mentioned for allowing a classical treatment of electron dynamics, but again more care is needed, because both position and momentum uncertainty must remain sufficiently small for a classical treatment to be accurate. Consistency demands that the quantum mechanical uncertainty in particle momentum $\Delta p_z(t)$ at least remain negligible compared to the classical longitudinal momentum spread of the beam, $\delta p \approx \frac{\delta\gamma}{\gamma} p \approx mc\,\delta\gamma$. For a plane-polarized wiggler field, another small longitudinal momentum scale exists, namely the extent of longitudinal momentum variation due to the "figure-eight" particle orbits in the plane-polarized wiggler fields, which, from energy conservation, can be shown to be approximately $\delta p_1 \approx \frac{a_w^2}{\gamma^2}\, p_0$ in the lab frame. A little algebra reveals that uncertainties in position and momentum can be

made simultaneously negligible provided (5.56) is satisfied, as well as

$$\frac{\lambda_c}{\lambda_0} \ll 4\pi\,\delta\gamma \tag{5.57a}$$

$$\frac{\lambda_c}{\lambda_0} \ll 4\pi\,\frac{a_u^2}{\gamma}. \tag{5.57b}$$

We have so far neglected the effects of any longitudinal forces. The fast harmonic motion along z in a planar wiggler should only slow the spatial spreading, or at least not significantly increase it, as should any ponderomotive bunching in the electromagnetic fields. Coulomb repulsion from the space-charge fields will be de-focusing, and might lead to additional spreading of the wave-packets, but classically this force is typically negligible compared to the other forces in the present parameter regime, as we will see shortly.

5.9.6 Radiation Reaction

However, in both the longitudinal and transverse dynamics, we have so far ignored a potentially important force, that of radiation reaction, or direct recoil. In order to consistently treat the particles classically but the radiation emission quantum mechanically, the effects of radiation reaction, or particle recoil, must be negligible.[8] Intuitively, we expect that the total effect on any particle due to the spontaneous wiggler radiation fields, including self-fields, should be negligible simply because the the radiation field strengths are much weaker than those of the external wiggler fields, i.e., $a_1 \ll a_u$, where a_1 is the normalized vector potential for the spontaneous wiggler radiation. Using (5.30) as an estimate for the total power radiated per particle, assuming each particle emits into a cone with synchrotron opening angle $\delta\theta \sim \frac{1}{\gamma}$ (larger than the coherent opening angle by $\sqrt{N_u}$), and assuming the particle positions are governed by shot noise, so the emission from different particles is randomly phased and therefore adds incoherently in intensity, we find that $a_1 \ll a_u$ provided

$$\alpha^2 N_u \left[n\sigma_\perp^2 \lambda_c\right] \frac{\lambda_c}{\lambda_u} \sim \frac{1}{\gamma^2}\alpha^2 N_u n \sigma_\perp^2 \lambda_c \frac{\hbar\omega}{mc^2} \ll 1. \tag{5.58}$$

While easily satisfied in the parameter regimes of interest here, this is essentially a far-field condition, while the radiation reaction force is manifestly a near-field effect, so more care is again needed. In order that we can ignore recoil effects in the longitudinal motion,

[8] Curiously, there is an apparently consistent formalism for evolving matter quantum mechanically but radiation classically which includes radiation reaction effects, the so-called neoclassical radiation theory of E. T. Jaynes, which can capture many effects, such as the photoelectric effect and the Lamb shift, that are purported to be evidence for the quantum nature of *light*. Some predictions of quantum optics, such as perfect anti-correlation after a beam-splitter for single photon states, or photon polarization states that violate Bell's inequalities, cannot be described in this model. Anyway, here we are concerned with the opposite "hemi-classical" limit, with particles treated classically and radiation quantum mechanically.

the average energy of radiation per particle should be small compared to the average particle energy, i.e.,

$$\alpha N_u a_u^2 \hbar\omega_0 \ll \gamma mc^2, \tag{5.59}$$

or equivalently

$$\alpha\, a_u^2 \tfrac{N_u \lambda_c}{\gamma} \ll \lambda_0. \tag{5.60}$$

Because the observed power radiated per particle (if it could be isolated) would be subject to large (Poissonian) fluctuations, we should really demand that the energy of any single radiated photon be small compared to the average particle energy:

$$\hbar\omega_0 \ll \gamma mc^2, \tag{5.61}$$

or equivalently

$$\tfrac{1}{\gamma}\lambda_c \ll \lambda_0. \tag{5.62}$$

Actually, we should impose still stricter conditions. In the average rest-frame, the energy change due to the recoil from a particle scattering one virtual photon into one real photon ought to be small compared to the typical RMS kinetic energy of a particle. Because $\delta\gamma' \approx 2\frac{\delta\gamma}{\gamma}$, back in the lab frame, this requirement entails

$$\tfrac{\lambda_c^2}{\lambda_u^2}\gamma^2 \ll \tfrac{\delta\gamma}{\gamma}, \tag{5.63}$$

setting an ultimate bound on how cold the particle beam can be or become and retain classical behavior, which is nevertheless easily satisfied for any remotely accessible beam temperatures. Since the particle cannot actually absorb energy from the wiggler magnetic field, perhaps we should even more strongly require that the energy of a single emission or absorption event by a particle (ignoring inconsistencies with momentum conservation) is small compared to the RMS particle kinetic energy, in the average rest frame. That is, demand $\hbar\omega' \ll mc^2\delta\gamma'$, or

$$\tfrac{\lambda_c}{\lambda_u}\gamma \ll \tfrac{\delta\gamma}{\gamma}, \tag{5.64}$$

which apart from neglected factors of 2 amounts to the same thing as requiring that the lab-frame radiated photon energy is small compared to the lab-frame energy spread, i.e.,

$$\hbar\omega_0 \ll mc^2\delta\gamma, \tag{5.65}$$

which is reassuring since the Correspondence Principle Limit should be relativistically invariant. This condition is mathematical equivalent to the previous condition (5.57a) despite being deduced from very different physical considerations.

In order that the particle continue to have well-defined phase in the optical field after emission, it is also necessary that the total particle recoil over the interaction time, due to one photon emission, remains small compared to the optical wavelength. For a photon radiated in the forward direction, the momentum kick is roughly

$$\hbar k_0 = \Delta p = mc\Delta\left(\gamma\beta\right) \sim mc\left(\gamma\Delta\beta + \gamma^3\beta^2\Delta\beta\right) \sim mc\gamma^3\Delta\beta. \tag{5.66}$$

We demand that $c\Delta\beta\frac{N_u\lambda_u}{c} \ll \lambda_0$, or

$$\frac{N_u\lambda_c}{\gamma} \ll \lambda_0. \tag{5.67}$$

This is precisely the same condition found above for the longitudinal quantum mechanical spreading of the wave-packet, without consideration of recoil effects or other longitudinal forces, to remain negligible.

Transversely, off-axis photons are emitted with a polar angle θ of at most about $\theta \sim \frac{1}{\gamma}$. The resulting momentum kick will be small compared to the transverse momentum spread if $\frac{1}{\gamma}\hbar k_0 \ll \delta p_\perp$, or

$$\hbar\omega_0 \ll \gamma^2\frac{\delta p_\perp}{p}mc^2, \tag{5.68}$$

where again $p = mc\gamma\beta$ is the average beam momentum. The transverse recoil should also be small compared to the transverse quiver, $\frac{1}{\gamma}\hbar k_0 \ll mc\,a_u$, or

$$\hbar\omega_0 \ll a_u\gamma mc^2; \tag{5.69}$$

and finally, the total transverse recoil over the interaction time must be sufficiently small so that the resulting perturbation in the particle's transverse position does not appreciably affect the phase of any subsequently emitted radiation. The transverse momentum kick is approximately

$$\frac{1}{\gamma}\hbar k_0 \sim \Delta p_\perp \approx mc\gamma\Delta\beta_\perp. \tag{5.70}$$

The total perturbation in transverse position due to this recoil is

$$\Delta z \sim c\Delta\beta_\perp\frac{N_u\lambda_u}{c} = \frac{1}{\gamma^2}\frac{\hbar\omega_0}{mc^2N_u\lambda_u}. \tag{5.71}$$

For nearly on-axis particles, the resulting perturbation in the phase of the emitted radiation, as collected at the end of the wiggler, is somewhere between $\Delta\phi \sim k_0\Delta s$ and $\Delta\phi \sim k_0\frac{\Delta s^2}{2N_u\lambda_u}$, depending on the particle's longitudinal position in the wiggler, so demanding $\Delta\phi \ll 2\pi$, we find

$$\hbar\omega_0 \ll \gamma^2 N_u\frac{\lambda_0}{\lambda_u}mc^2 \sim N_u mc^2 \tag{5.72}$$

and

$$\hbar\omega_0 \ll \frac{\gamma^2}{\sqrt{N_u}} \left[\frac{\lambda}{\lambda_u}\right]^{1/2} mc^2 \sim \frac{\gamma mc^2}{\sqrt{N_u}}. \tag{5.73}$$

We have seen that recoil effects on any one particle are negligible for its subsequent dynamics, but in order that recoil can be fully ignored, its effects on the radiation field must also be small. From the Compton scattering relation, we see that in the average rest frame of the beam, the shift in frequency $\delta\omega'$ due to particle recoil during the scattering of a virtual wiggler photon into a real radiation photon in a Weizsäcker-Willliams treatment will be small compared to the central radiation frequency ω_0' provided that, in the lab frame

$$\frac{1}{2\gamma}\frac{\hbar\omega_0}{mc^2} \ll 1 \tag{5.74}$$

or equivalently

$$\lambda_c \ll 2\gamma\lambda_0. \tag{5.75}$$

Actually, this Compton shift should also be small relative to the coherent bandwidth, requiring

$$\frac{1}{\gamma}\frac{\hbar\omega_0}{mc^2} \ll \frac{\hbar\omega_0}{N_u}, \tag{5.76}$$

equivalent to the by now very familiar requirement that

$$\frac{N_u}{\gamma}\lambda_c \ll \lambda_0. \tag{5.77}$$

We must also demand that the recoil from emission of a photon will not change the energy of any one particle so much that any subsequent emission by that particle will likely lie outside the original coherent bandwidth at the average beam energy, but a little algebra shows that this condition is again (5.77); we begin to understand why in the literature this is often regarded as *the* condition defining the classical/quantum boundary for wiggler physics.

As pointed out by Benson and Madey [172] and Brau [173], these standard recoil arguments really only indicate that *expectation values* of particle observables can be calculated classically. We must also verify that variances or fluctuations about mean values are dominated by classical effects such as shot noise in the beam. We expect that the most stringent condition for higher-order moments will arise from considerations of longitudinal motion, so we concentrate on this behavior here. We have seen that the average recoil of a particle due to photon emission is sufficiently small for a classical description to be valid, but it should also be the case that additional quantum fluctuations due to the variance in the photon emission are negligible.

Now, it is known that the photon counting statistics for Compton scattering from plane waves is Poissonian, so for sufficiently uniform wiggler fields and a sufficiently relativistic beam, such that in the average beam rest frame the wiggler field can be represented as virtual photons in a plane-wave mode according to the Weizsäcker-Williams approximation, photon counting statistics for photons scattered by any one particle should be Poissonian or close to Poissonian. Of course, it is precisely the correct understanding of the field statistics, before and after amplification, as experienced not by hypothetical photon counters, but by actual individual particles in the beam which is at issue in our entire analysis, and we are not yet equipped to make a careful calculation. But to arrive at a conservative upper bound for the variance or fluctuations due to recoil, we can assume that the photon emission statistics for each particle in the wiggler are Poissonian, and then demand that the extra quantum mechanical uncertainty in energy due to fluctuations in photon emission is negligible compared to the momentum uncertainty already needed to satisfy the uncertainty principle, which in the absence of these recoil fluctuations has already been shown to be negligibly small. Even though most of the total power emitted by a given particle will be radiated at frequencies more than one coherent bandwidth below the central photon energy $\hbar\omega_0$, or outside the coherent solid angle $\delta\theta \sim \frac{1}{\sqrt{N_u\gamma}}$, to be conservative we will here assume that all power is emitted close to the forward direction near the frequency ω_0. The expected total number of photons radiated per particle (including those outside the coherent mode), in the pickup is then estimated as $\bar{n}_{\rm rad} \sim \alpha N_u a_u^2$. The additional uncertainty in particle energy due to these fluctuations in radiation, assuming Poissonian statistics, is then

$$\delta\gamma_{\rm rad} \sim \frac{\hbar\omega_0}{mc^2}\sqrt{\alpha N_u a_u^2} = \frac{\lambda_c}{\lambda_0}\sqrt{\alpha N_u a_u^2}. \tag{5.78}$$

For a classical-behaving wave-packet, the uncertainty in longitudinal position at the end of the pickup interaction should be no more than a few times the minimum uncertainty allowed as given by (5.54), which corresponds to an approximate lower bound on the Heisenberg uncertainty in beam energy:

$$\delta\gamma_{\rm HUP} \geq \sqrt{\frac{\lambda_c\gamma}{\lambda_0 N_u}}. \tag{5.79}$$

Demanding that $\delta\gamma_{\rm rad} \ll \delta\gamma_{\rm HUP}$, we find that

$$\frac{\lambda_c}{\lambda_0}\frac{\alpha}{\gamma}N_u^2 a_u^2 \ll 1, \tag{5.80}$$

which will be readily achieved in situations of interest.

5.9.7 Summary of Arguments for Classicality of Particle Degrees-of-Freedom

We have uncovered a plethora of conditions in support of our claim that the particle dynamics in the pickup wiggler can be treated classically. Perhaps we should review the most important: $\rho_b \ll \left(\frac{\epsilon_\perp}{\lambda_c}\right)^3 \frac{\gamma}{\sigma_\perp^3}$ and $\rho_b \ll \frac{1}{\lambda_c^3}\left(\frac{\delta\gamma}{\gamma}\right)^3$, so that the beam particles are non-degenerate; $\frac{\delta\gamma}{\gamma} \ll 1$, so that pair-creation and other exotic QED effects can be ignored regardless of beam density; $\frac{a_u}{\gamma} \ll 1$, so that certain spin effects can be neglected; and

$$\frac{\hbar\omega_0}{mc^2} \ll \min\left[\delta\gamma, \gamma, a_u, a_u\gamma, \frac{a_u^2}{\gamma}, \frac{\gamma^3}{a_u}, N_u, \frac{\gamma}{\sqrt{N_u}}, \frac{\gamma}{N_u}, \frac{\gamma}{N_u^2 a_u}, \frac{\gamma^3}{N_u a_u}, \frac{\gamma}{\alpha N_u^2 a_u^2}, \frac{\delta p_\perp}{mc}, \frac{\gamma\delta p_\perp}{mc}\right], \tag{5.81}$$

so that a variety of other potential spin, transverse, and longitudinal effects may be ignored

Note that the condition $\frac{N_u\lambda_c}{\gamma} \ll \lambda_0$, or $\frac{\hbar\omega_0}{mc^2} \ll \frac{\gamma}{N_u}$, is often mentioned as the fundamental condition for classical behavior of particles in wigglers, because it simultaneously ensures at least two fundamental conditions for classicality: that the longitudinal spreading of a single-particle wave-packet over the length of the wiggler in the absence of emission remains small relative to the shortest force length-scale, which is the optical wavelength; and that if a photon is emitted, then the resulting recoil is sufficiently small so that the resulting shift in particle position over the interaction time in the wiggler is small compared to an optical wavelength. However, for short wigglers this condition is not typically the most stringent of the ones we have established.

It should also be acknowledged that, strictly speaking, all the inequalities deduced above, individually or collectively, constitute neither sufficient nor necessary conditions for the observation of exclusively classical behavior on the part of the beam particles. The failure of any one condition does not necessarily indicate that quantum behavior will be observed, but only that it will be necessary to perform an actual quantum-mechanical calculation to determine whether the quantum corrections to classical behavior will be appreciable or not. That is, they are necessary conditions for avoiding a detailed quantum treatment.

Conversely, certain of these conditions determine whether sufficiently classical-looking quantum states for the beam exist, but even when such states are available there is no guarantee that the particles will be prepared in or remain in such a state. However, it remains a poorly understood but widely observed tendency for systems, especially macroscopic systems consisting of either massive objects or of many interacting constituent parts, to actually behave classically whenever they can behave classically. The emergence of a classical world from quantum physics remains a fundamental and only very partially understood problem, although some recent progress has been made in this direction by Zurek [174, 175, 176]

and others concerning the analysis of decoherence and so-called environmentally-induced superselection rules.

In the case of particle beams at least, all empirical evidence suggests this is true. We believe that the resolution lies in the realization that one does not typically track or make observations on individual particles but rather the particle beam as a whole, or at least some slice or segment of the beam, which always contains many particles and is always subject to uncertainty due to uncontrolled classical noise from the environment and/or our limited initial knowledge and subsequent measurement precision. Recall that the closest classical analog of a quantum state, even a single-particle state, is not a particle trajectory but rather a phase-space distribution, or statistical ensemble of trajectories. If there are sufficiently classically-looking quantum states for individual particles, consistent with what is known macroscopically about the beam, typically consisting of a just a few parameters such as average dimensions, density, average energy, and longitudinal and transverse emittances, and if furthermore the classical emittances are sufficiently large, then classical statistical uncertainty can simply swamp quantum fluctuations, and the density matrix for the beam as a whole may be essentially indistinguishable from a proper statistical mixture of the near-classical pure states, regardless of whether individual particles were actually prepared or remain in these particular classical-looking states.

5.10 Classical Single-Particle Dynamics Are Adequate

We safely conclude that particles in the beam can be taken to behave classically throughout the pickup wiggler, and furthermore, assert that they will follow classical trajectories determined by the initial conditions and the external wiggler magnetic fields (and other external focusing fields, if present) and perhaps mean self-fields, with negligible perturbation from the spontaneously-radiated pickup radiation itself, either through recoil (radiation-reaction) or coherent forcing or or incoherent scattering.

Under these conditions, we will see that it is possible to consistently treat the physics in what we will call a "hemi-classical" formalism, where the particles are treated classically but the radiation is described quantum mechanically, in contrast to "semi-classical" regime, a description usually applied either to a system in some sense poised between quantum and classical behavior and approaching a correspondence principle limit or otherwise amenable to a long wavelength asymptotic analysis, or else to interactions between matter and fields

in which the matter must be described quantum-mechanically but the radiation can be described classically.[9]

While not strictly necessary for the validity of the hemi-classical description, it will nevertheless be quite convenient to also argue away even the Vlasov (mean) fields, so that in fact the beam can be simply described by the single-particle physics of classical charges moving in external fields

5.10.1 Radiation Effects Are Small

We have already seen above that radiation reaction effects may be neglected, so the effects on a given particle of its own wiggler radiation are unimportant, But this radiation actually slips ahead of the particle, and might interact with other beam particles downstream. The simplest way to check whether these effects might be important is to compare the density of actual radiated photons inside the wiggler to the effective density of virtual wiggler photons. In the average rest frame, the latter photon density is given approximately by

$$n'_u \sim \pi\alpha^{-1}\frac{\gamma a_u^2}{\lambda_u\lambda_c^2}. \tag{5.82}$$

Using (5.30), and assuming the radiation from each particle is predominately radiated within the synchrotron opening angle $\frac{1}{\gamma}$ in the lab frame, and remembering that a given particle will feel the fields from about N_s other particles while traveling through the wiggler, the density of radiated photons experienced by a particle in the average rest frame is at most about

$$n'_{\rm rad} \sim 4\alpha\frac{\gamma^3 N_s a_u^2}{N_u^2\lambda_u^3}, \tag{5.83}$$

and demanding $n'_{\rm rad} \ll n'_u$, after a little algebra we deduce the requirement

$$\alpha^2\gamma^2 N_s \ll \tfrac{\pi}{4}N_u^2\tfrac{\lambda_u^2}{\lambda_c^2} \tag{5.84}$$

which is expected to be satisfied with many orders-of-magnitude to spare.

With this density of radiation photons, the mean free path ℓ_C for Thomson/Compton scattering from any one beam particle may be roughly estimated in the lab frame by setting

$$\ell_c\alpha^2\lambda_c^2\tfrac{1}{\gamma}n'_{\rm rad} \approx 1, \tag{5.85}$$

[9] We will also see that within the hemi-classical formalism, the behavior and observable properties of the radiation will be almost entirely classical, so the hemi-classical theory is semi-classical in the first sense of the word, but complementary to the second, widely-used meaning.

so we will have $\ell_c \ll N_u \lambda_u$ provided

$$\alpha^2 \frac{\lambda_c^2}{\lambda_u^2} \gamma^2 \frac{N_s}{N_u} a_u^2 \ll 1 \tag{5.86}$$

which is readily satisfied, so scattering off the pickup radiation by downstream particles is completely negligible. Besides, even if subsequent Thomson or Compton scattering from the radiated field does occur, the momentum transfers will be of comparable magnitude to the direct radiation recoil terms during the original emission, already known to be negligible.

5.10.2 Mean Space-Charge Forces are Mostly Small

Assuming a long, nearly azimuthally-symmetric beam with uniform particle density n_b, the average transverse Coulomb-Lorentz force due to self-fields on a beam particle at a distance r from beam axis is

$$F_r(r) = \frac{2\pi q^2 n_b}{\gamma^2} r, \tag{5.87}$$

so that the average electric and magnetic transverse forces due to mean self-fields cancel to $O\left(\frac{1}{\gamma^2}\right)$ for a symmetric beam. The residual repulsive force will be negligible compared to the Lorentz force from the wiggler provided $F_r(\sigma_\perp)\frac{N_u \lambda_u}{c} \ll a_u mc$, or

$$\alpha \frac{1}{a_u \gamma^2} N_u n \sigma_\perp \lambda_u \lambda_c \ll 1 \tag{5.88}$$

which is readily satisfied in the stretched beam. To be completely negligible, this repulsive self-force should also result in transverse beam spreading which remains small compared to the optical wavelength over the interaction time, i.e., $\frac{1}{2}\frac{1}{\gamma m} F_r(\sigma_\perp)\left[\frac{N_u \lambda_u}{c}\right]^2 \ll \lambda_0$, or

$$\alpha \frac{1}{\gamma} N_u^2 n \sigma_\perp \lambda_u \lambda_c \ll 1, \tag{5.89}$$

which is also satisfied provided the beam is sufficiently stretched. For a long uniform beam, the average longitudinal force on an randomly chosen vanishes by symmetry, but the RMS force does not, and particles away from the midpoint of the beam will experience Coulomb repulsion. The velocity is highly relativistic, so particles are inertially stiff, but on the other hand, the relevant length-scale for perturbations is only the optical wavelength λ_0 over the full interaction time Δt in the wiggler. At about one standard deviation into the tail of a Gaussian beam, the Coulomb repulsion will scale like $F_z \sim \pi q^2 n \frac{\sigma_\perp^2}{\sigma_z}$. The resulting spread over the interaction time should be small compared to the optical wavelength, i.e.,

$$\frac{1}{2\left[\gamma + \beta^2\gamma^3\right] m} F_z \left[\frac{N_u \lambda_u}{c}\right]^2 \ll \lambda_0 \tag{5.90}$$

which becomes

$$\alpha N_u^2 \frac{1}{\gamma} n \sigma_{perp}^2 \lambda_u \frac{\lambda_c}{\sigma_z} \ll 1, \tag{5.91}$$

which is easily satisfied for our parameters.

5.10.3 Fluctuations and Collective Oscillations are Small

In addition to establishing that the average space-charge effects are negligibly small, to we must additionally show that the fluctuations and collective oscillations about this average behavior, arising from self-fields, are also negligible. Collective plasma oscillations will be unimportant if either the relativistic plasma frequency ω_p' in the average rest frame is sufficiently low so that oscillations simply do not have time to develop during the interaction time, or if any Langmuir waves that might arise are heavily damped via Landau damping mechanisms, which will occur if the typical wavelength λ_p' for plasma waves is small compared to the Debye screening length $\lambda_{\rm D}'$. In the average rest frame, the relativistic plasma frequency is given in Gaussian units by

$$\omega_p' = \left[\frac{4\pi n' q^2}{m}\right]^{\frac{1}{2}} = \left[\frac{4\pi n q^2}{\gamma m}\right]^{1/2}, \tag{5.92}$$

where from relativistic length contraction $n' = n/\gamma$ and $\Delta t' \approx \frac{N_u \lambda_u}{\gamma c}$, so demanding $\omega_p' \Delta t' \ll 2\pi$, we find

$$\left[2\alpha \frac{1}{\gamma^3} N_u^2 n \lambda_u^2 \lambda_c\right]^{1/2} \ll 2\pi, \tag{5.93}$$

which is expected to be satisfied, at least for sufficiently massive particles like muons or protons.

In the average rest frame, the Debye length is $\lambda_D' = \sqrt{\frac{k_{\rm B} T'}{4\pi q^2 n'}}$, where T' is the effective longitudinal rest-frame temperature. From relativistic velocity addition, we have seen that

$$k_{\rm B} T' \approx mc^2 \langle \beta'^2 \rangle \approx mc^2 \frac{\delta\gamma^2}{\beta^2 \gamma^2}, \tag{5.94}$$

assuming $\gamma \gg 1$ and $\frac{\gamma}{\gamma} \ll 1$. Therefore the rest-frame Debye length can be written in terms of lab-frame quantities as

$$\lambda_{\rm D}' = \left[\frac{\gamma}{\alpha \lambda_c n}\right]^{\frac{1}{2}} \frac{\delta\gamma}{\gamma}. \tag{5.95}$$

The rest-frame wavelength λ_p' for plasma oscillations is expected to be comparable to that of the wiggler period in this frame, $\lambda_p' \sim \lambda_u' = \lambda_u/\gamma$, so that plasma oscillations will be suppressed provided $\lambda_p' \ll \lambda_D'$, or

$$\left[\frac{2\alpha \lambda_u^2 \lambda_c n}{\gamma \delta\gamma^2}\right]^{\frac{1}{2}} \ll 1. \tag{5.96}$$

Because the wiggler period in the envisioned OSC scheme is so long compared to most other applications, this may be only marginally satisfied or may actually be violated, but because the beam is diffuse and the total number of wiggler periods is so small and the resulting interaction time so short, waves that are not suppressed by Landau damping will not really have any time to arise, as demonstrated in as demonstrated in (5.93) above.

While we can and will therefore neglect collective oscillations here, these conditions could fail for certain beam parameter regimes, and then the effects of collective space-charge oscillations would in principle need to be included in the particle dynamics for a fully consistent treatment.

Collisions, or binary (but possibly shielded) Coulomb interactions between discrete particles not accounted for in the Vlasov (mean) fields, could also cause deviations from the ideal wiggler orbits. Including Debye screening effects, the effective Coulomb scattering frequency ν_c' in the average rest frame (including the cumulative effects of many small-angle scattering events) is approximately

$$\nu_c \approx \frac{8\pi n' q^4}{m^2 c^3 \delta\beta'^3} \ln\left(\Lambda'\right), \tag{5.97}$$

where $\delta\beta'$ is the RMS velocity in the average rest frame, and where

$$\Lambda' = \lambda_D'^3 n'^3 = \left[\frac{\delta\gamma}{\gamma}\right]^3 \left[\frac{\gamma}{2\alpha}\right]^{\frac{3}{2}} \left[\frac{1}{\lambda_c^3 n}\right]^{\frac{1}{2}} \tag{5.98}$$

is the rest-frame plasma parameter, and satisfies $\Lambda' \gg 1$, while typically $\ln(\Lambda') \leq O(10)$ or so. Comparing to the plasma frequency, we see that

$$\frac{\nu_c'}{\omega_p'} \approx \frac{\ln(\Lambda')}{2\pi\Lambda'},$$

so that the scattering frequency is expected to be much smaller than the plasma frequency, and should to be unimportant. Specifically, Coulomb scattering will be unimportant if $\nu_c'\Delta t' \ll 1$, or

$$\left[2\alpha \frac{1}{\gamma^3} N_u^2 n \lambda_u^2 \lambda_c\right]^{\frac{1}{2}} \frac{\ln(\Lambda')}{2\pi\Lambda'} \ll 1, \tag{5.99}$$

which is readily achieved for beams of interest.

Finally, while we have seen that we can neglect the effects of the space-charge forces on the average density n, or equivalently on the one-particle distribution function, we seek to establish that space-charge effects will not lead to significant correlations in the two-particle distribution function, so that the relative positions of beam particles can still be described by classical shot noise. Such shot noise is characterized by Poissonian statistics

within any given spatial region, or equivalently by particle positions chosen randomly and independently in proportion to the average density n, and so shot noise in any one Lorentz frame will appear as shot noise in any other frame, just with an appropriately Lorentz-transformed density. In the average rest frame, Coulomb repulsion is expected to lead to a certain amount of anti-correlations in particle positions, which may develop over the full lifetime of the beam, not just the interaction time in the wiggler. Roughly on the time-scale ν_c^{-1} corresponding to the Coulomb scattering and dielectric screening times in a plasma, we expect the beam to have approximately approach a state of thermodynamic equilibrium in its average rest frame, at least in the absence of cooling and when neglecting the slow expansion of the beam, and hence in order to estimate the magnitude of these particle correlations we can make use of well-known results from equilibrium statistical mechanics, without having to solve a much more difficult time-dependent problem in kinetic theory.

In the average rest frame, consider a sub-volume $\delta V'$ small compared to the beam as a whole but large enough to contain many particles on average. (Strictly speaking, we actually need not demand that $\delta N' \gg 1$, but only that $\delta N' \ll N_b$.) Treating this region as a thermodynamic system in thermal and diffusive equilibrium with the rest of the beam temporarily regarded as a heat and particle reservoir, it is easy to see that in the absence of particle interactions, the particle number $\delta N'$ in this volume will be subject to Poissonian fluctuations. In the non-degenerate (low-n and/or high-T) limit of indistinguishable free particles, the partition function for the region of interest can be taken to be

$$Z'(T', \delta N', \delta V') = \frac{\left[n'_Q \delta V'\right]^{\delta N'}}{\delta N'!}, \tag{5.100}$$

where

$$n_Q = n'_Q(T') = \left[m\frac{k_\mathrm{B}T'}{2\pi\hbar^2}\right]^{\frac{3}{2}} \sim \frac{1}{\lambda'^3_{dB}}$$

is the quantum concentration, corresponding to one particle in a volume whose linear dimensions scale like the thermal de Broglie wavelength λ'_{dB}, and where for simplicity we assume the transverse and longitudinal temperatures are equal. Using the usual Stirling approximation for the factorial term, the Helmholtz free energy is given by

$$F'(T', \delta N', \delta V') = -k_\mathrm{B}T' \ln Z' = \delta N' k_\mathrm{B}T' \left[\ln\left(\frac{\delta N'/\delta V'}{n'_Q}\right) - 1\right],$$

so the chemical potential is

$$\mu'(T', \delta N', \delta V') = \frac{\partial}{\partial \delta N'} F'(T', \delta N', \delta V') = k_\mathrm{B}T' \ln\left(\frac{\delta N'/\delta V'}{n'_Q}\right).$$

The grand partition function can then be written as

$$\mathcal{Z}'(T',\mu',\delta V') = \sum_{\delta N'=0}^{\infty} \lambda'^{\delta N'} \frac{\left[n'_Q \delta V'\right]^{\delta N'}}{\delta N'!} = e^{\lambda' n'_Q \delta V'},$$

where $\lambda' = e^{\frac{\mu'}{k_B T'}}$ is the absolute activity. The average particle number within $\delta V'$ is then

$$\delta\bar{N}'(T',\mu',\delta V') = \left\langle \delta N'(T',\mu',\delta V') \right\rangle = \lambda' \frac{\partial}{\partial \lambda'} \ln \mathcal{Z}'(T',\lambda',\delta V') = \lambda' n'_Q \delta V',$$

and the probability distribution for particle number can be written as

$$P'(\delta N') = \frac{1}{\mathcal{Z}'} \frac{\left[\lambda' n'_Q \delta V'\right]^{\delta N'}}{\delta N'!} = \frac{\left\langle \delta N' \right\rangle^{\delta N'} e^{-\langle \delta N' \rangle}}{\delta N'!}, \tag{5.101}$$

which is a Poisson distribution, as expected, so that particle fluctuations satisfy

$$\left\langle \left(\delta N' - \left\langle \delta N' \right\rangle\right)^2 \right\rangle = \left\langle \delta N' \right\rangle.$$

When the effects of Coulomb interactions are included, the general problem becomes somewhat intractable, but a perturbation-type expansion for the partition function or free energy can be found in the typical plasma regime, implying large Debye shielding, $\Lambda \equiv \lambda_D'^3 n' \gg 1$, or equivalently weak coupling:

$$\frac{q^2 \, n'^{\,1/3}}{k_B T'} \ll 1, \tag{5.102}$$

so that that the average Coulomb potential energy between neighboring particles is small compared to their mean thermal energy. For a single charged species, the Helmholtz free energy becomes

$$F'(T',\delta N',\delta V') = \delta N' k_B T' \left[\ln\left(\frac{\delta N'/\delta V'}{n'_Q}\right) - 1\right] - \frac{2}{3}\sqrt{\pi} q^3 \delta N' \left[\frac{\delta N'}{\delta V'}\right]^{\frac{1}{2}} + \cdots. \tag{5.103}$$

Because of the long-range nature of Coulomb force, successive terms in the expansion include various fractional powers and even a logarithm of the local density $\frac{\delta N'}{\delta V'}$, so this is not a simple virial expansion in the density, and because of the appearance of q^3, neither is it an analytic perturbation expansion in the natural coupling parameter q^2. But assuming the perturbation parameter (5.102) is very small, we need include only the leading correction, the so-called Debye term, and need not worry here about non-analyticity or convergence of higher-order terms. Typically, this free energy for a Coulomb gas is derived using textbook Debye-shielding arguments (see for example [177]) leading to an effective screened potential

energy of Yukawa form, but such derivations assume net neutrality of the plasma. However, the expansion (5.103) remains valid even for non-neutral plasmas, and can be derived from quantum or classical statistical mechanics by re-summation techniques, or in a kinetic theory formalism by truncating the BBGKY hierarchy beyond the two-particle distribution function, a valid approximation when the inter-particle correlations in fact do remain small, which can be self-consistently checked at the end of the calculation.

At least in the plasma or Debye-shielded regime, the long-range nature of the Coulomb force is both blessing and curse; the statistical correlations which can develop between a given particle and any one other randomly chosen particle in the plasma will on average be small because of the screening effects of all the other intervening particles, and the long-range correlations which do persist result from the self-averaged effects of many particles and can be incorporated through a self-consistent mean-field (in our case the beam space-charge field), leading to the well-known collective effects, as distinct from true many-body correlations.

While it is somewhat troublesome to find a convenient expansion for the full probability distribution for $\delta N'$, it is straightforward to find the so-called Fano factor, or ratio of variance to average, sufficient for our purposes. Using the Helmholtz free energy (5.103), the chemical potential for the Coulomb gas becomes

$$\mu(T', \delta N', \delta V') = \frac{\partial}{\partial \delta N'} F'(T', \delta N', \delta V') = k_{\mathrm{B}} T' \ln\left(\frac{\delta N'/\delta V'}{n'_Q}\right) - \sqrt{\pi} q^3 \left[\frac{1}{k_{\mathrm{B}} T'} \frac{\delta N'}{\delta V'}\right]^{\frac{1}{2}}. \tag{5.104}$$

Shifting from the canonical to the grand canonical ensemble, the chemical potential $\mu(T', \overline{\delta N}', \delta V')$ evaluated at the average particle $\overline{\delta N'}$ is formally invertible, and can be taken to implicitly define the average particle number $\overline{\delta N'}(T', \mu', \delta V') = \langle \delta N'(T', \mu', \delta V') \rangle$ as a function of the chemical potential μ'. The variance in particle number can then be expressed as

$$\left\langle \left(\delta N' - \langle \delta N' \rangle\right)^2 \right\rangle = k_{\mathrm{B}} T' \frac{\partial}{\partial \mu'} \langle \delta N'(T', \mu', \delta V') \rangle = \frac{1}{\frac{\partial}{\partial \overline{\delta N'}} \mu'(T', \overline{\delta N'}, \delta V')}, \tag{5.105}$$

which after a little algebra can be written as

$$\begin{aligned} \left\langle \left(\delta N' - \langle \delta N' \rangle\right)^2 \right\rangle &= \frac{\overline{\delta N'}}{1 - \frac{1}{2}\sqrt{\pi} q^3 \left(\frac{1}{k_{\mathrm{B}} T'}\right)^{\frac{3}{2}} \left(\frac{\overline{\delta N'}}{\delta V'}\right)^{\frac{1}{2}}} \\ &= \frac{n' \delta V'}{1 - \frac{1}{2}\sqrt{\pi} \left(\frac{q^2 n^{1/3}}{k_{\mathrm{B}} T'}\right)^{3/2}}, \end{aligned} \tag{5.106}$$

so the actual Fano factor may be written simply as

$$\frac{\left\langle \left(\delta N' - \langle \delta N' \rangle\right)^2 \right\rangle}{\langle \delta N' \rangle} = \frac{1}{1 - \frac{1}{16\pi}\frac{1}{\Lambda'}}. \tag{5.107}$$

Provided the weak-coupling condition (5.102) holds, we see that any deviations from Poissonian fluctuations are indeed small, at least at the level of first and second-order moments.

In summary, to a very good approximation, we may therefore assume that that particles in the beam behave classically in the pickup wiggler, even as the particles radiate, and that their relative positions are well approximated by classical shot noise, and that their trajectories are determined almost exclusively by the external electromagnetic fields and their individual initial conditions at the entrance to the wiggler, and are largely unaffected by collisions or space-charge forces or by the radiation field itself, either through coherent or incoherent scattering or through radiation reaction.[10] With this classical model of prescribed particle trajectories, the subsequent analysis can be tremendously simplified. Off-axis effects and all transverse motion of particles are ignored accept idealized 1D quiver motion in the wiggler field, as determined by canonical momentum conservation. Higher harmonics of particle motion in planar wiggler are ignored, and longitudinal motion is deduced from conservation of kinetic energy and transverse canonical momentum in the wiggler fields.

5.11 Quantum Mechanical Description of Wiggler Radiation

An exact analytic description of the quantum mechanics of emission would be prohibitively difficult, requiring a complete treatment of the dynamics of the coupled particle-field system, and in general leading to highly entangled states for the matter and radiation. But in the regime established above, were recoil effects are negligible and particles follow classical orbits that can be determined independently of the actual radiation fields, the problem is simplified tremendously. For a review of quantum and classical features of radiation from free of quasifree electrons, see [134].

However, before we can deduce the quantum state of the radiation, we must digress to summarize a quantum optics formalism that can capture the collimation, collection, transport and amplification of the wiggle radiation.

[10] As we have seen, in some regimes, collective Langmuir oscillations might be marginally appreciable, but as we are interested in a fundamental proofs-of-principle regarding the effects of radiation noise, these collective effects will be subsequently ignored.

5.11.1 Quantum Optics for Dielectric-Guided Paraxial Beams

After emission in the pickup wiggler, travels essentially paraxially down a mostly open beam line, but which does contain various passive dielectric lenses used collect, and transport the radiation, as well as one or more apertures to select transverse mode characteristics, in addition to the active gain medium of the amplifier or amplifiers. If the field is quantized in the vacuum, then the effects of dielectric elements must be included as interaction terms in the Hamiltonian coupling a large number of vacuum modes, complicating the dynamics.

Instead, we can quantize directly the "macroscopic fields" in the presence of a linear medium described by susceptibility tensors. Numerous approaches to this problem may be found in the literature, for example those in [178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193] and even earlier antecedents. Most are restricted to special cases where the medium is homogeneous and isotropic, or the dielectric tensor is frequency-independent, or only certain kinds of frequency response are allowed. The general problem, involving anisotropy, slow spatiotemporal variations in the medium response, and arbitrary frequency dispersion consistent with the Kramers-Kronig relations, remains an area of current research, and certain fundamental debate persists.[11]

Here we consider the case where $\mu = 1$ while the dielectric tensor $\epsilon(\boldsymbol{x})$ is real (i.e., lossless), positive, isotropic, dispersionless over the wiggler and amplifier bandwidths, but possibly spatially inhomogeneous, i.e., piecewise smooth in its dependence on position $\boldsymbol{x}$. In practice, some dispersion will be present, so here $\epsilon(\boldsymbol{x})$ is chosen to best represent the average effects of the actual lenses or other optical media along the beam path. Perfectly-conducting boundaries[12] can also be included to model apertures or beam pipes. The resulting eigenmodes in general posses more complicated spatial structure than simple plane waves, but will retain harmonic time dependence. The exact form of these modes is not

[11] Disagreement continues over whether the electric field $\boldsymbol{E}(\boldsymbol{x}, t)$ or the displacement field $\boldsymbol{D}(\boldsymbol{x}, t)$ should be canonically conjugate to the vector potential $\boldsymbol{A}(\boldsymbol{x}, t)$. Also, in the microscopic theory it is well known that with the addition of sources, the commutation relations of the field observables are unchanged, and commute with all observables for the material degrees-of-freedom, which does not always seem to be the case with the quantized macroscopic theories. These questions are not fully resolved, but [194] suggests that some of the disagreement may just be due to different gauge choices, or to different dressed operators going by the same name. Anyway, this mostly concerns the meaning of "photons" and the nature of measurements inside the dielectric medium, which are minor concern here.

[12] We have worked out in detail the case where the quantization region is simply-connected and simply-bounded. Some subtleties arise if it is not simply-connected, because then the kernel of the curl operator can contain certain vector fields which are not gradients of a single-valued scalar field, reflecting the non-trivial cohomology of the interior region. However, generalizations of the Helmholtz-Hodge Decomposition Theorem can be used in these cases, and this is a subject of ongoing study on our part.

actually needed for our purposes, here, but in the vacuum regions can be taken to be the usual paraxial Gauss-Hermite basis set.

We had worked out this particular quantum theory of dielectrics in some detail, before realizing it was nearly identical to that developed by Glauber and Lewenstein in [178], so we jut summarize the main results. In the presence of the hypothesized medium, the macroscopic fields can still be written (in Gaussian units) as $\boldsymbol{B} = \boldsymbol{H} = \boldsymbol{\nabla} \times \boldsymbol{A}$ and $\boldsymbol{E} = \epsilon^{-1}\boldsymbol{D} = -\boldsymbol{\nabla}\phi - \frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{A}$. We choose to work in the generalized Coulomb, or ϵ-transverse, gauge, where

$$\boldsymbol{\nabla} \cdot [\epsilon \boldsymbol{A}] = 0. \tag{5.108}$$

For any piecewise-smooth dielectric function satisfying $0 < \epsilon(\boldsymbol{x}) < \infty$, it can be rigorously shown using generalizations of the Helmholtz theorem and existence and uniqueness results for the Laplace equation, that this is a well-defined gauge choice leading to a unique decomposition of the electric field into an ϵ-transverse component containing all (but not only) radiation and an irrotational (curl-free) non-radiative component. From the macroscopic Maxwell's Equations, we deduce that the scalar and vector potentials potentials must satisfy

$$\boldsymbol{\nabla} \cdot [\epsilon \boldsymbol{\nabla}\phi] = -4\pi\rho, \tag{5.109a}$$

$$-\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{A}) - \tfrac{\epsilon}{c^2}\tfrac{\partial^2}{\partial t^2}\boldsymbol{A} = -\frac{4\pi}{c}\boldsymbol{J} + \frac{\epsilon}{c}\frac{\partial}{\partial t}\boldsymbol{\nabla}\phi \tag{5.109b}$$

where ρ and $\boldsymbol{J}$ are the "free" charge and current flux densities, respectively, for any charges not implicitly accounted for through $\epsilon(\boldsymbol{x})$, namely those in the particle beam.

GIven the assumed good behavior of the dielectric function, it can be verified that the generalized Poisson equation (5.109a) has existence and unique properties essentially identical to the standard Poisson boundary value problem. The scalar potential ϕ is therefore determined by the instantaneous positions of all the additional charges, as in the usual Coulomb gauge, so does not embody independent field degrees-of-freedom

Separation of variables is used to express the vector potential $\boldsymbol{A}$ in terms of paraxial eigenmodes $\boldsymbol{u}_k(\boldsymbol{x};\omega)$ of the generalized Helmholtz equation:

$$\epsilon(\boldsymbol{x})^{-1}\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times \boldsymbol{u}_k(\boldsymbol{x};\omega) = \tfrac{\omega^2}{c^2}\boldsymbol{u}_k(\boldsymbol{x};\omega); \tag{5.110}$$

Here the set of integers labeled by k is taken to index the transverse modal structure, assumed to be discrete, as in a paraxial beam or a waveguide, including and polarization

degeneracy, so k may have multiple components which are implicitly hidden in the notation for brevity. The frequency ω labels[13] what is in general a continuum of longitudinal modes.

The generalized Helmholtz operator appearing in (5.110) is not Hermitian with respect to the usual $\mathcal{L}_2$ inner product on $\mathbb{R}^3$, but it is Hermitian with respect to the ϵ-weighted inner product:

$$(\boldsymbol{u}\,,\,\boldsymbol{v})_\epsilon = \int d^3\boldsymbol{x}\,\epsilon(\boldsymbol{x})\,\boldsymbol{u}(\boldsymbol{x})^*\cdot\boldsymbol{v}(\boldsymbol{x}), \tag{5.111}$$

and with suitable boundary conditions can be naturally extended to a self-adjoint operator, and the eigenmodes may then be chosen to be orthogonal with respect to this inner product:

$$\big(\boldsymbol{u}_{k'}(\boldsymbol{x};\omega')\,,\,\boldsymbol{u}_k(\boldsymbol{x};\omega)\big)_\epsilon = \delta_{kk'}\delta(\omega-\omega'), \tag{5.112}$$

where $\delta_{kk'}$ and $\delta(\omega)$ are the Kronecker delta symbol and Dirac delta function, respectively.

These modes will be complete in the sense of spanning the space of ϵ-transverse vector fields in the quantization region, so the vector potential $\boldsymbol{A} = \boldsymbol{A}_{\epsilon\perp}$ for the radiation fields can be decomposed as

$$\boldsymbol{A}(\boldsymbol{x},t) = \sum_k \int_0^\infty d\omega\, c\sqrt{\tfrac{2\pi\hbar}{\omega}}\left[a_k(\omega)\boldsymbol{u}_k(\boldsymbol{x};\omega)e^{-i\omega t} + a_k(\omega)^\dagger \boldsymbol{u}_k^*(\boldsymbol{x};\omega)e^{+i\omega t}\right], \tag{5.113}$$

where the macroscopic field operators $a_k(\omega)$ satisfy canonical commutation relations

$$\left[a_k(\omega), a_{k'}(\omega')^\dagger\right] = \delta_{kk'}\delta(\omega-\omega') \tag{5.114a}$$

$$\left[a_k(\omega), a_{k'}(\omega')\ \right] = 0. \tag{5.114b}$$

In the presence of the dielectric medium, the "quasi-free" Hamiltonian governing propagation in the absence of free charges becomes

$$H_0 = \tfrac{1}{8\pi}\int d^3\boldsymbol{x}\,\boldsymbol{E}\cdot\boldsymbol{D} + \boldsymbol{B}\cdot\boldsymbol{H} = \sum_k \int_0^\infty d\omega\,\hbar\omega\, a_k(\omega)^\dagger a_k(\omega), \tag{5.115}$$

where constants and other terms which do not depend on the field operators have been dropped, and which remains equivalent to that of a collection of uncoupled harmonic oscillators as in the vacuum case. The source/field interaction Hamiltonian remains

$$H_{\text{int}} = -\tfrac{1}{c}\int d^3\boldsymbol{x}\,\boldsymbol{J}\cdot\boldsymbol{A}. \tag{5.116}$$

[13]Following both Caves[195], and general conventions for paraxial optics, we labeled the modes by the frequency ω at which they would oscillate if allowed to freely evolve, because in general they are not characterized by a true fixed wavenumber. The explicit dependence on ω is simply a label, and is not meant to indicate that these are in general Fourier transforms of time-domain quantities. In fact, we remain in the time-domain, where functions labeled by ω nevertheless evolve in t.

Note, however, that the macroscopic operators $a_k(\omega)$, are "dressed," in general containing in regions with dielectric contributions from both the usual creation and annihilation operators associated with the microscopic fields in free space.

Recall that here the background dielectric function must be real and independent of time and/or frequency. Any effects of dispersion or attenuation must be incorporated explicitly through additional interaction terms. The gain in the active amplifying medium must also be treated quantum mechanically, but a specific atomic model will not be needed; very general features of quantum mechanics will be sufficient to characterize its best-case noise properties without explicitly evolve the radiation fields through the system.

One could argue that by neglecting dispersion we are throwing away a source of noise for the radiation. Dispersion at some frequencies implies absorption at other frequencies according to the Kramers-Kronig relations, while fluctuation-dissipation theorems suggest that this absorption must also be associated with noise. Since quantum noise in the radiation field is of central concern, one might reasonably fear that ignoring dissipation and the attendant fluctuations in the lenses or other dielectric elements is a terrible idea. However, there are no fundamental lower bounds on the losses and fluctuations introduced by passive dielectric elements in any given bandwidth, so this perhaps is a useful idealization, because the frequencies at which absorption occurs and noise is injected are hopefully well removed by design from the bandwidth of interest. If not, because we will assume the radiation propagation and amplification are entirely linear, we can actually lump any dispersion, attenuation, gain, and noise into a single generalized amplification process, and need not ascribe them to individual elements.

5.11.2 "Hemi-Classical" Model for Particle-Field Dynamics

At last we are able to determine the quantum state of the radiation field due to emission in the pickup wiggler. We have seen that, with high accuracy, the dynamics can be treated in a "hemi-classical" approximation, where the particles and static wiggler field are treated classically, without any radiation reaction forces, while the light is treated quantum mechanically, a regime complementary to the usual semi-classical approximation, in which the matter and its Coulombic interactions are treated quantum mechanically but the radiation fields are assumed classical.

First we briefly review some important properties of the coherent states, first discussed

by Schrödinger but introduced and extensively analyzed by Glauber[196, 197, 198, 199] in the context of the quantum theory of optical coherence.

Coherent States

As eigenstates of the modal annihilation operators for the fields, such coherent states possess many useful mathematical and physical properties, a few of which will be briefly reviewed here. (See, for example, [200] for a thorough modern discussion.) Given a complete orthonormalized set of field modes, the (multi-mode) continuum coherent state $|\{\alpha_k(\omega)\}\rangle$ is by definition an eigenstate of each modal annihilation operator $a_k(\omega)$ with complex eigenvalue $\alpha_k(\omega)$, i.e.,

$$a_k \left|\{\alpha_k(\omega)\}\right\rangle = \alpha_k(\omega) \left|\{\alpha_k(\omega)\}\right\rangle . \tag{5.117}$$

In the case of a discrete set of modes, this multi-mode Coherent state is manifestly separable, and can be written as tensor product $\bigotimes_{ik} |\alpha_k(\omega_i)\rangle$. The choice $\alpha_k(\omega) = 0$ corresponds to the vacuum $|0\rangle$.

By convention, phases are chosen so that the coherent state can be written as

$$|a_k\rangle = D\left[\alpha_k(\omega)\right]|0\rangle , \tag{5.118}$$

where

$$D\left[\alpha_k(\omega)\right] = e^{\sum_k \int d\omega \left[\alpha_k(\omega) a_k(\omega)^\dagger - \alpha_k(\omega)^* a_k(\omega)^\dagger\right]} \tag{5.119}$$

is the unitary modal phase-space displacement operator, which satisfies $D\left[-\alpha_k(\omega)\right]^{-1} = D\left[\alpha_k(\omega)\right]^\dagger = D\left[\alpha_k(\omega)\right]^{-1}$ and $D\left[\alpha_k(\omega)\right]^\dagger a_k(\omega) D\left[\alpha_k(\omega)\right] = a_k(\omega) + \alpha_k(\omega)$, We see that Coherent states just consist of displaced vacuum states, so they possess the same uncertainties as the vacuum but with a non-zero bias to mode amplitudes. The states are well-defined and normalizable provided that $\sum_k \int d\omega \ |\alpha_k(\omega)|^2 < \infty$.

Using this definition of the coherent state, it is straightforward to show for a discrete mode the coherent state can be written as

$$|a_k(\omega_i)\rangle = \sum_{n=0}^{\infty} \frac{\alpha_k(\omega_i)^n e^{-|\alpha_k(\omega_i|^2}}{\sqrt{n!}} |n\rangle \tag{5.120}$$

where $|n\rangle = \frac{1}{\sqrt{n!}} \left[a_k(\omega_i)^\dagger\right]^n |0\rangle$ is the usual number eigenstate. From (5.120), it is evident that coherent state will have Poissonian photon-counting statistics under ideal measurement conditions.

Glauber coherent states are also minimum uncertainty states in two important senses. For simplicity , let us consider the case of discrete modes. For any choice of overall phase $\phi \in$

realsymbol, the Hermitian quadrature operators, or sine-like and cosine-like components of the field amplitude operator $a_k(\omega_i)$ are defined as

$$X_k(\omega_i) = X_k(\omega_i)^\dagger = \tfrac{1}{2}\left[e^{-i\phi}a_k(\omega_i) + e^{i\phi}a_k(\omega_i)^\dagger\right] \tag{5.121}$$

which is said to be in-phase, and

$$Y_k(\omega_i) = Y_k(\omega_i)^\dagger = \tfrac{1}{2i}\left[e^{-i\phi}a_k(\omega_i) + e^{i\phi}a_k(\omega_i)^\dagger\right], \tag{5.122}$$

which is said to be in quadrature. The Hermitian quadrature operators can be shown to satisfy

$$\left[X_k(\omega_i), X_{k'}(\omega_i')\right] = 0, \tag{5.123a}$$

$$\left[Y_k(\omega_i), Y_{k'}(\omega_i')\right] = 0, \tag{5.123b}$$

$$\left[X_k(\omega_i), Y_{k'}(\omega_i')\right] = \tfrac{i}{2}\delta_{kk'}\delta_{ii'}. \tag{5.123c}$$

Defining $\Delta X_k(\omega_i) = X_k(\omega_i) - \langle X_k(\omega_i)\rangle$ and $\Delta Y_k(\omega_i) = Y_k(\omega_i) - \langle Y_k(\omega_i)\rangle$, it follows form the generalized Heisenberg-Robertson uncertainty principle that

$$\langle \Delta X_k(\omega_i)\rangle \langle \Delta Y_k(\omega_i)\rangle \geq \tfrac{1}{4}\left|[X_k(\omega_i), Y_k(\omega_i)]\right|^2 = \tfrac{1}{16}. \tag{5.124}$$

It is easily shown that any coherent state symmetrically saturates this uncertainty-principle limit, in the sense that the variances satisfy $\left\langle \Delta X_k(\omega_i)^2\right\rangle = \left\langle \Delta Y_k(\omega_i)^2\right\rangle = \frac{1}{4}$. Also, while it is impossible to define a true Hermitian phase observable $\Phi_k(\omega_i)$ conjugate to the number operator $\mathscr{N} = a_k(\omega_i)^\dagger a_k(\omega_i)$, under reasonable definitions of a quasi-phase operator, such as the Barnett-Pegg phase [200], the coherent states also saturate or nearly saturate the "number-phase" uncertainty principle: $\left\langle (\Delta \mathscr{N}_k(\omega_i))^2\right\rangle (\Phi_k(\omega_i))^2 \geq \frac{1}{4}$.

Coherent states also have important overlap and over-completeness properties, but these will be introduced as needed, as well as beautiful group theoretic structure which we will not need here.

Optical Correspondence Principle

Central to our arguments is the so-called Optical Correspondence Principle, a theorem first established by Glauber [198] in early work on the theory of coherent states, and then generalized independently by Glauber[201] and Sudarshan and Mehta[202, 203]. As it turns out, essentially the identical theorem holds within our formalism for "macroscopic" radiation propagation in a linear medium described by $\epsilon(\boldsymbol{x})$ and possibly bounded by perfect conductors.

If the initial state of the radiation field is a multi-mode Glauber coherent state (possibly but not necessarily the vacuum state), and the free sources all follow prescribed classical trajectories, then the final state of the radiation field is also a multi-mode Glauber coherent state, whose expectation value is given by the corresponding solution to Maxwell's equations with the same sources and propagating through the same medium, with classical initial conditions given by the expectation value of the initial quantum state.

If the sources for interest for the radiation, namely the charges in the particle beam, can be taken as a prescribed classical current,then the interaction Hamiltonian (5.116) can be written as

$$H_{\rm int}(t) = -\sum_k \int_0^\infty d\omega \left[j_k(\omega;t)^* a_k(\omega) + j_k(\omega;t) a_k^\dagger \right], \tag{5.125}$$

where

$$j_k(\omega;t) = \sqrt{\tfrac{2\pi\hbar}{\omega}} \int d^3\boldsymbol{x}\, \boldsymbol{u}_k(\boldsymbol{x};\omega)^* \cdot \boldsymbol{J}(\boldsymbol{x},t) \tag{5.126}$$

As long as the current density $\boldsymbol{J}(\boldsymbol{x},t)$ is considered a prescribed current density, rather than the result of self-consistent trajectories of charged sources moving in the fields, it need not be expressed in terms of canonical particle coordinates, and the above form for the interaction Hamiltonian thus holds even if the actual sources are arbitrarily relativistic.

We see that each mode in the medium still evolves independently if subjected to prescribed sources, so for simplicity we will drop all subscripts and indices cluttering the notation and look at a single mode, which evolves according to

$$H(t) = \hbar a^\dagger a - \left[j(t)^* a + j(t) a^\dagger \right] \tag{5.127}$$

for a given c-numberr function $j(t)$ Note that this Hamiltonain looks exaclty like the corresponding Hamiltonain in absence of the dielectric medium but with the same prescribed currents, but here the definition and meaning of each dressed modal annihilation operator a is generally somewhat different than in a vacuum; by design, the effects of the medium are incorporated (hidden) in the spatial structure of the mode, determining what kind of photon a annihilates. If the initial state for this mode is described a Glauber coherent state up to some overall phase, i.e., if $|\psi(0)\rangle = e^{i\phi(0)} |\alpha(0)\rangle$ for some real number $\phi(0)$ and complex $\alpha(0)$, then we claim that in the Schrödinger picture, the state at later times is still given by some

$$|\psi(t)\rangle = e^{i\phi(t)} |\alpha(t)\rangle, \tag{5.128}$$

where $|\alpha(t)\rangle$ is a coherent state with standard phase. Because the coherent states are overcomplete, the state (5.128) will be a solution to the time-dependent Schrödinger equation

corresponding to the Hamiltonian (5.127), provided

$$\langle\beta|\, i\hbar\tfrac{\partial}{\partial t}\,|\psi(t)\rangle = \langle\beta\,|H(t)|\,\psi(t)\rangle \tag{5.129}$$

holds for any multiple of any arbitrary time-independent Glauber state $|\beta\rangle$, i.e., for any state such that $\langle\beta|\, a^\dagger = \beta^*\,\langle\beta|$ and $\frac{d}{dt}\beta = 0$. Using the eigenvector conditions for the coherent states $|\alpha(t)\rangle$ and $|\beta\rangle$, the inner product in the form

$$\langle\beta\,|\,\alpha\rangle = e^{-\frac{1}{2}\left(|\beta|^2+\alpha^2-2\beta^*\alpha\right)}, \tag{5.130}$$

and the over-completeness relation

$$\int \tfrac{d^2\beta}{\pi}\,|\beta\rangle\,\langle\beta| = \mathcal{I}, \tag{5.131}$$

where $d^2\beta = d\,\mathrm{Re}(\beta)d\,\mathrm{Im}(\beta)$ and $\mathcal{I}$ is the identity operator on the relevant modal Fock space, we find that indeed $ket\psi(t)$ is a solution for all subsequent time, provided that $\alpha(t)$ satisfies

$$\tfrac{d}{dt}\alpha(t) = -i\omega\alpha(t) + \tfrac{i}{\hbar}j(t), \tag{5.132}$$

and $\phi(t)$ satisfies

$$\tfrac{d}{dt}\phi(t) = +\tfrac{1}{2\hbar}\left[j(t)^*\alpha(t) - j(t)\alpha(t)^*\right]. \tag{5.133}$$

Note that the real quantity $\phi(t)$ is the overall phase of the quantum state vector for this mode, which does not appear in the corresponding density matrix, and should not be confused with the actual optical phase of the mode itself, whose expectation value may be taken as $\arg[\alpha(t)]$, and is not directly related to $\phi(t)$. Since the modal equation of motion (5.132) is identical to that associated with the classical evolution of $H(t)$ with the same sources, the expectation value $\langle a(t)\rangle = \alpha(t)$ simply tracks the classical field dynamics driven by the given sources, starting from a classical field given by the initial expectation value $\langle\alpha(0)\rangle$ of the quantum field a. Since this is true for each of a complete set of radiation modes, the result is proved.

The essential idea is that the Heisenberg equations of motions for the electromagnetic field operators are exactly analogous to the classical Maxwell equations, or equivalently to evolution equations for harmonic oscillators (one for each EM mode), and so if the sources are the same prescribed c-number functions in both cases, we expect the solutions, we expect the solutions to be as nearly identical as allowed by quantum mechanics. The average of the quantum mechanical solution coincides with the classical solution, while the uncertainties, quantified in terms of the variances and covariances of field quantities, remain just the

minimum allowable by the uncertainty principle (because the amplitude and phase of each mode are conjugate), and are conventionally ascribed to "vacuum fluctuations."

This is an crucial result to our argument, and merits some elaboration. Coherent states can be thought of as phase-space displacement of the vacuum, i.e., a translation of the vacuum fluctuations by some average amplitude and phase angle, so that the resulting state retains the symmetric minimum-uncertainty or "zero-point" fluctuations, only centered about the new average amplitude and phase. For classical, prescribed sources, an initial coherent state remains a minimum uncertainty state, with the centroid just following the classical trajectory

This correspondence between the classical solution and the average of the quantum mechanical state holds however dim the radiation; that is, regardless of how weak the sources are or how small the resulting average photon density is. The relative uncertainty increases, or in other words the signal-to-noise ratio decreases, but the average behavior is still exactly classical.

Note that Glauber's Correspondence Principle holds for arbitrarily relativistic sources, and the "prescribed" classical particle trajectories can include the effects of charges moving in non electromagnetic classical force fields, external classical EM fields such as the wiggler magnetic field, mean space-charge fields, or even the instantaneous (non-radiative) Coulomb fields of other charges, as represented in the scalar potential in the generalized-transverse gauge, just no feed back from the radiation field itself. In general, any appreciable recoil during the radiation process, or subsequent scattering off of other particle's real (as opposed to virtual) radiation, can destroy Glauber coherence of the radiation field. In particular, any gain in the pickup wiggler due to the FEL instability must be negligible, so that all radiation is dominated by spontaneous rather than stimulated wiggler emission.

Optical Equivalence Theorem

While the trajectories of the particles are describable by classical mechanics and assumed definite, they will not be known exactly but only probabilistically. If sources are classical but stochastic in this sense, i.e., definite in principle but unknown, then the radiation field will be in a statistical mixture of coherent states, each with positive statistical weight, i.e., a so-called proper mixture of coherent states, with the weights interpretable as a standard probability distribution reflecting ignorance.

It turns out that *any* density matrix can be written as a weighted sum of projections

into coherent states[198, 199, 204, 205, 206, 207, 207], if we are sufficiently generous in allowing the weights to become negative or singular. Specifically, for the case of discrete modes, we have the diagonal coherent-state representation

$$\rho_{\rm rad} = \int d^2\alpha_{k_1}(\omega_1) d^2\alpha_{k_2}(\omega_2) \cdots P\big(\alpha_{k_1}(\omega_1), \alpha_{k_1}(\omega_1), \dots\big)\, |\alpha_{k_1}(\omega_1); \cdots\rangle\langle\, \alpha_{k_1}(\omega_1); \cdots|\,, \tag{5.134}$$

where $P\big(\alpha_{k_1}(\omega_1), \alpha_{k_1}(\omega_1), \dots\big)$ is known as the Glauber-Sudarshan P-function, or normally-ordered quasi-distribution function.

Radiation from classical prescribed sources is described by states where the P function is non-negative and normalizable, so may be interpreted as a classical probability measure over the corresponding coherent states. In addition, the quantum measurement statistics of any state with such a nonnegative P function can be reproduced in a semi-classical model (quantum detectors, classical field) by using the analogous classical field. Conversely, all known manifestations of truly quantum mechanical behavior in radiation, such as squeezing in number or quadrature, anti-bunching, violation of Bell's inequalities, etc., that cannot be reproduced in a classical or semi-classical model, involve states with negative or even infinitely negative P functions, leading to Sudarshan's Equivalence Principle[206, 207], which asserts that non-negativity and normalizability of the P function are jointly necessary and sufficient[14] for classical behavior of the radiation field.

The closest analogs to classical radiation allowed by quantum mechanics correspond to coherent states or their proper statistical mixtures, i.e., to positive, normalizable P functions; and conversely, any state for which P is normalizable and everywhere positive, so that it can be interpreted as a classical probability distribution, is in its field statistics and properties virtually indistinguishable from the corresponding stochastic classical field, regardless of how small the average photon number. Note that textbook requirement of large degeneracy (in the sense of large mean modal occupation, i.e., $\langle \mathscr{N}_k(\omega_i)\rangle \gg 1$) of the relevant participating modes is, strictly speaking, neither necessary nor sufficient for classical behavior of radiation. However in practice, increasing the degeneracy does tend to make observation of remaining quantum signatures (i.e., statistics which could not be reproduced classically) more difficult. Conversely, at low degeneracy, any state, including coherent states or thermal states, will be subject to large relative uncertainties in photo-

[14]Necessity is now widely recognized. Sufficiency is sometimes still debated for very weak or occasionally other states. But any discreteness or other quantum signatures in the measurements of such as a state can always be attributed to quantum mechanical features of the detectors rather than the fields, so we stand by sufficiency as well.

counts, but this can be consistently attributed to the quantum nature of the atoms in the detector, because the same statistics are reproduced in a semi-classical theory.

In many respects coherent states can be thought of as equivalent to a (classical) superposition of a deterministic classical field (itself representable as a deterministic superposition of plane waves or other classical modes of the optical system in question) and an unavoidable amount of stochastic noise associated with quantum mechanical uncertainty. Thus, we conclude that classical (but possibly stochastic) sources produce classical-looking (but possibly stochastic) radiation, as might be expected.

Quasi-Distribution Functions and Quantum Characteristic Functions

The Glauber-Sudarshan P function is the most natural quasi-distribution function from which to assess the classical/quantum correspondence for radiation, but others are useful in different settings. Consider the case of a single discrete mode for simplicity.[15] We can define the s-ordered quantum characteristic function[208, 205] in terms of the von Neumann density matrix $\rho_{\rm rad}$ as

$$\chi(\xi;s) = \mathrm{Tr}\left[\rho_{\rm rad} D(\xi)\right] e^{s\frac{1}{2}|\xi|^2} \tag{5.135}$$

in principle for any $s \in \mathbb{R}$, but in practice the most useful choices are $s = +1, 0, -1$, corresponding, as we will see, to normal-ordering, symmetric-ordering, and anti-normal ordering of the mode operators. As in classical statistics, the derivatives of the characteristic function are related to the appropriately ordered expectation values:

$$\left\langle \left[a^{\dagger\ell} a^m\right]_s \right\rangle = \left(\frac{\partial}{\partial\xi}\right)^\ell_{\xi^*} \left(-\frac{\partial}{\partial\xi^*}\right)^m_{\xi} \chi(\xi;s)\Big|_{\xi=0} \tag{5.136}$$

where $\left[a^{\dagger\ell} a^m\right]_s$ denotes the s-ordered product, so that, for example, $\left\langle \left[a^\dagger a\right]\right\rangle = \left\langle a^\dagger a\right\rangle + \frac{1}{2}(1-s)$.

The corresponding quasi-distribution functions are obtained by Fourier transforms, as in the classical case:

$$F(\alpha;s) = \tfrac{1}{\pi^2}\int d^2\xi\, \chi(\xi;s) e^{\alpha\xi^* - \alpha^*\xi}, \tag{5.137}$$

so that s-ordered moments are calculated just like normal moments of ordinary random variables:

$$\left\langle \left[a^{\dagger\ell} a^m\right]_s \right\rangle = \int d^2\alpha\, \alpha^{*\ell}\alpha^m F(\alpha;s). \tag{5.138}$$

[15] The results here can be generalized to continuum modes, involving quasi-distribution functionals, but the notation involves functional derivatives and path integrals and gets rather cumbersome.

The quasi-distribution functions are always real, but not necessarily non-negative for $s > -1$, and may be highly singular. Each is is a smoothed version of the preceding via convolution with a Gaussian:

$$F(\alpha, s-1) = \tfrac{2}{\pi}\int d^2\beta F(\beta; s) e^{-2|\alpha-\beta|^2}. \tag{5.139}$$

The normally-ordered choice $s = 1$ corresponds to the Glauber-Sudarshan P-function introduced above, the symmetrically-ordered choice corresponds to the Wigner Function $F(\alpha; 0) = W(\alpha) = \frac{1}{\pi}\,\mathrm{Tr}\left[D(\alpha)^\dagger \Pi D(\alpha)\rho_{\rm rad}\right]$, where Π is a parity operator which flips the sign of the "momentum-like" quadrature component but leaves the "position-like" component unchanged, and the anti-normally ordered choice $s = -1$ corresponds to $F(\xi; -1) = Q(\alpha) = \frac{1}{\pi}\left\langle \alpha \left| \rho_{\rm rad} \right| \alpha \right\rangle \geq 0$.

From the P function, we can directly determine whether the state can exhibit non-classical behavior. The Husimi Q function is useful to graphically represent the modal phase space "occupied" by the state, since it is non-negative. The Wigner function corresponds to symmetrically-ordered (or Weyl-ordered) observables which often are of concern, and is the distribution that people implicitly have in mind when they speak of the vacuum fluctuations or zero-point motion as being equivalent to "half a photon" per mode. To see this, note that the characteristic functions for a thermal, or blackbody state containing an average of $\bar{n}$ photons are

$$\chi_{\bar{n}}(\xi; s) = e^{-\frac{1}{2}(2\bar{n}+1-s)|\xi|^2}, \tag{5.140}$$

so if $s = 0$, we see that in a thermal state, the contribution from vacuum fluctuations (i.e., the part independent of $\bar{n}$) is exactly equivalent to the addition of half a photon. While obviously the P-function of a coherent state corresponds to a Dirac delta function, the corresponding W function is a Gaussian occupying about $\frac{1}{2}\hbar$ of quadrature phase space.

Summary of Field State Prior to Amplification

In practice, rather than working directly with the resulting density matrix, it will be useful to defer the averaging over the classical uncertainty in the particle trajectories until the end of the calculation. That is, because of the linearity of the quantum evolution and of any averaging over classical ignorance, we can propagate the conditional state given the particle trajectories, apply these fields to the particles in the kicker, and then average over the initial particle distribution function, to determine the actual statistics for the beam cooling at each pass.

This conditional state of the radiation in the coherent wiggler mode, given the particle orbits, will be a pure state, in particular a multi-mode[16] Glauber coherent state, such that before any amplification, its expectation value is

$$\langle \boldsymbol{E}(\boldsymbol{x},t)\rangle_{\text{pre}} = \sum_j \langle \boldsymbol{E}_j(\boldsymbol{x},t)\rangle \tag{5.141}$$

is equal to the corresponding classical solution consisting of the sum of the contributions from all the particles, while its variance will scale like

$$\left\langle |\Delta \boldsymbol{E}(\boldsymbol{x},t)|^2 \right\rangle_{\text{pre}} \approx \tfrac{1}{2}\tfrac{\hbar\omega_0}{V_u}, \tag{5.142}$$

where $V_u \sim \pi w_0^2 N_u \lambda_0 \approx \frac{N_u^2 \lambda_u \lambda_0^2}{8\pi}$ is the original coherent mode volume in real space (not phase space).

5.11.3 Amplification of Pickup Wiggler Radiation

The exact quantum mechanical evolution of a mixture over such multi-mode coherent states during the amplification process will of course depend on precise details of the laser amplifiers, but by considering idealized quantum mechanical amplifiers, the best-case statistical properties of the amplified radiation and the minimum amount of added amplifier noise can be derived using just a few very basic assumptions about the input-relations. We will follow the very illuminating fully quantum treatment of Caves[195], who extended and clarified pioneering analyses (both semi-classical and quantum) in [209, 210, 211, 212, 213, 214, 215, 216, 217, 218] and elsewhere. (See also [219, 220] and references therein for other insightful modern treatments of quantum amplifier noise.)

It will be supposed that the amplifier output mode field operators depend only linearly on the input mode operators, implying that saturation effects in the output are ignored, although it need not be assumed that the internal dynamics within the gain medium or the interactions with the pumping or any other supplementary field modes are actually linear. The amplifier is further assumed to be time-stationary, or frequency-preserving and phase-preserving; i.e., an input mode at a given frequency is mapped into an amplified output mode at the same frequency (typically just the same mode, extrapolated forward along the

[16] Viz a viz the radiation of concern to us, each particle radiates into a single coherent wiggler mode consisting approximately of a diffraction-limited and Fourier-limited wave packet of length $N_u\lambda_u$, but the eigenmodes of the optical system were taken to be time-harmonic, so many of these modes are required to represent the coherent wiggler mode. However, only part of the zero-point noise from each of these harmonic eignenmodes appears over the actual extent of the radiated wave-packet felt by any given electron, so we can still count roughly one mode in the "half photon per mode" noise rule.

paraxial optical path), and if the input is shifted in phase by some constant offset, then the output will be shifted by the same amount, so the amplifier has no absolute internal clock nor any preferred phase bias. For transit-time OSC, this time-stationarity is desirable, or even essential, because the information used by the cooling mechanism to reduce particle phase space is stored almost exclusively in the phases rather than amplitudes of the light, and efficient field/particle coupling in the kicker will require that the light frequency remains near the resonant frequency of the particles in the wiggler.

To develop the needed gain, the amplifier can be multi-stage (cascaded), but must be single-pass for two reasons. Firstly, the amount of total time-delay that can be realistically introduced between pickup and kicker compared to the proper time between these points (i.e., the time it would take light to travel between them in vacuum) is not very great for heavier particles like muons traveling at near-luminal velocities, so there simply is not time to circulate the light. Secondly, a regenerative amplifier operating at high pump power can easily be pushed over the oscillation threshold, where it converts from an amplifier preserving the phase and modulations of the input to a laser oscillating very cleanly but at essentially at some random phase.

Quantum Mechanics of Linear Optical Amplifiers

In the Heisenberg picture, where time-evolution is ascribed to the operators, it is supposed that the amplifier output mode field operators depend only linearly on the input mode operators The amplifier is further assumed to be time-stationary, that is, frequency and phase-preserving. Consistent with the 1D dynamics adopted throughout the current analysis, it is also taken to preserve transverse coherence, in the sense of maintaining a one-to-one relationship between input and output transverse modes. As demanded by quantum mechanics, the time-evolution of the field must be unitary, and the field observables must satisfy appropriate commutation relations and Heisenberg uncertainty relations at all times before, during, and after amplification.

It turns out that these basic assumptions about linearity and lack of phase bias, along with the axioms of quantum mechanics are sufficient to characterize the action of and best possibly noise figures for any linear amplifier, without needing to know the specific details of the amplifier system, or explicitly following the evolution of all the degrees-of-freedom in its gain medium.

In the Heisenberg picture, we take $a_k(\omega)$ to represent the continuum field annihilation

operator for a mode *before* amplification (i.e., the input mode operator) and $b_k(\omega)$ to represent the annihilation operator *after* amplification (i.e., the output operator). The modes are specified by a continuum frequency ω as well as some discrete index k meant to represent different transverse paraxial modes, or otherwise lift degeneracy. The input modes satisfy the canonical Bosonic commutation relations (5.114) as must the output modes, i.e.

$$\left[b_k(\omega), b_{k'}(\omega')^\dagger\right] = \delta_{kk'}\delta(\omega - \omega') \tag{5.143a}$$

$$\left[b_k(\omega), b_{k'}(\omega')\ \right] = 0, \tag{5.143b}$$

consistent with the quantum mechanical evolution remaining unitary. By linearity, we mean that the input and output modes are related by affine transformation, so we are disregarding any sort of saturation effects in the gain medium. The most general possible linear input-output relation is then

$$b_k(\omega) = \sum_{k'} \int d\omega' \left[M_{kk'}(\omega, \omega')a_{k'}(\omega') + L_{kk'}(\omega, \omega')a_{k'}(\omega')^\dagger\right] + F_k(\omega) \tag{5.144}$$

Next we demand that the input-output relations are time-stationary, meaning both frequency and phase-preserving. By frequency preserving we mean that input at frequency ω is mapped to output at frequency ω, while by phase-preserving we mean that a phase rotation of the input leads to the same phase rotation of the output, and that if the input has time-stationary noise statistics, then so will the output. These are desirable properties for any amplifier to be used for transit-time OSC, and together imply that

$$M_{kk'}(\omega, \omega') = M_{kk'}(\omega)\delta(\omega - \omega') \tag{5.145a}$$

$$L_{kk'}(\omega, \omega') = 0. \tag{5.145b}$$

Finally, we assume a one-to-one relationship between input and output transverse modes, so that the matrix $M_{kk'}(\omega)$ is unitarily diagonalizable at any frequency ω. Actually, without any loss of generality under our assumptions, and to avoid introducing new notation, we may suppose that the $a_k(\omega)$ and $b_k(\omega)$ have *already* been chosen in the proper linear combination to correspond to the transverse eigenfunctions of the amplifier, so that $M_{kk'}(\omega)$ is already diagonal:

$$M_{kk'}(\omega) = \delta_{kk'}M_k(\omega), \tag{5.146}$$

and the input-output relations for the amplifier become simply

$$b_k(\omega) = M_k(\omega)a_k(\omega) + F_k(\omega). \tag{5.147}$$

Obviously $M_k(\omega)$ is the amplifier transfer function for the kth mode of frequency ω, such that $M_k(\omega)^\dagger M_k(\omega) = G_k(\omega)$ is the power gain for the mode. In general, $M_k(\omega)$ can be an operator (q-number) commuting with all input mode operators:

$$\left[a_k(\omega), M_{k'}(\omega')\right] = \left[a_k(\omega)^\dagger, M_{k'}(\omega')\ \right] = 0. \tag{5.148}$$

The additive noise operators $F_k(\omega)$ must be present to ensure unitarity. They represent independent degrees-of-freedom corresponding to some internal or auxiliary states of the amplifier, so can be taken commute with the input and output modes. In fact, in order that the Bosonic commutation relations are preserved, we can deduce the commutation relations

$$\left[a_k(\omega), F_{k'}(\omega')^\dagger\right] = \left[a_k(\omega), F_{k'}(\omega')\ \right] = 0, \tag{5.149a}$$

$$\left[F_k(\omega), F_{k'}(\omega')\ \right] = 0, \tag{5.149b}$$

$$\left[F_k(\omega), F_{k'}(\omega')^\dagger\right] = \delta_{kk'}\delta(\omega-\omega')\left[1 - M_k(\omega)M_k(\omega)^\dagger\right], \tag{5.149c}$$

for the noise operators $F_k(\omega)$.

To complete the description of the amplifier, we must specify an initial state (density operator) for the scaling and noise operators. In the Heisenberg picture, time evolution resides in the operators, so this operating state is fixed, and should be independent of any possible input signal. That is, the density matrix of the total system (modes to be amplified plus all auxiliary modes internal degrees-of-freedom of the amplifier) factorizes: $\rho_0 = \rho_a \otimes \rho_{\rm amp}$.

No fundamental physical principle imposes any lower bound on the variances and covariances of the $M_k(\omega)$, so in the ideal case they can be taken to be c-numbers, implying no *multiplicative* quantum noise in the output, that could arise from variations in the gain via fluctuations in the pump strength, for example. Besides, such noise can easily be added back later, if needed. With $G_k(\omega) = |M_k(\omega)|^2 \geq 1$ and deterministic, note that we can write $F_k(\omega) = \sqrt{G_k(\omega)-1}\,f_k(\omega)^\dagger$, where $f_k(\omega)$ is a Bosonic annihilation operator for what can be interpreted as the effective internal amplifier mode that contributes the extra noise. Nothing prevents an assumption excluding bias in the noise, such that $\langle F_k(\omega)\rangle = 0$, and in fact our previous assumptions about phase preservation require it.

But the commutation relations together with a generalized Heisenberg Uncertainty Principle applicable to creation and annihilation operators constrain the second-order noise statistics. For the symmetrized covariances, we find the inequality

$$\tfrac{1}{2}\left\langle F_k(\omega)^\dagger F_{k'}(\omega') + F_{k'}(\omega')F_k(\omega)^\dagger\right\rangle - \left\langle F_k(\omega)^\dagger\right\rangle\left\langle F_{k'}(\omega')\right\rangle \geq \tfrac{1}{2}\left|\left[F_k(\omega), F_{k'}(\omega')^\dagger\right]\right| \tag{5.150}$$

In particular, for coherent-state input, the output noise statistics, in terms of symmetrized covariances become,

$$\left\langle b_k(\omega)b_{k'}(\omega')\right\rangle - \left\langle b_k(\omega)\right\rangle \left\langle b_{k'}(\omega')\right\rangle = 0; \tag{5.151}$$

and

$$\begin{aligned}\tfrac{1}{2}\left\langle b_k(\omega)^\dagger b_{k'}(\omega') + b_{k'}(\omega')b_k(\omega)^\dagger\right\rangle - \left\langle b_k(\omega)^\dagger\right\rangle \left\langle b_{k'}(\omega')\right\rangle = \\ G_k(\omega)\delta_{kk'}\delta(\omega-\omega')\left[\tfrac{1}{2} + S_k(\omega)\right].\end{aligned} \tag{5.152}$$

The first term on the right-hand side represents the amplified "vacuum fluctuations"[17] present in the input, equivalent in noise to "half a photon per unit bandwidth" before amplification, while the second term represents the extra noise added by the amplifier (typically through amplified spontaneous emission in the gain medium), in which the added-noise spectral density $S_k(\omega)$ must satisfy:

$$S_k(\omega) = \tfrac{1}{2}\left|1 - \tfrac{1}{G_k(\omega)}\right|\sigma_k^2(\omega), \tag{5.153}$$

where $\sigma_k^2(\omega) \geq 1$, with equality in the Heisenberg-limited case, so at large gain the added noise power is the equivalent of at least one additional "half photon" per mode of input to the amplifier.

For a single mode, note that the fluctuations in photon number in the output are super-Poissonian. If the input is coherent, the Fano factor starts off as unity: $\frac{\left\langle(\Delta\mathscr{N}_a)^2\right\rangle}{\langle\mathscr{N}_a\rangle} = 1$ but at output, becomes $\frac{\left\langle(\Delta\mathscr{N}_b)^2\right\rangle}{\langle\mathscr{N}_b\rangle} \approx G + 1 \sim G \gg 1$, where here $\mathscr{N}_a = a^\dagger a$ and $\mathscr{N}_b = b^\dagger b$. The output state is chaotic, not coherent.

5.11.4 Quasi-Distribution Functions

For simplicity, we will drop all the indices and arguments and for the moment treat a single discrete input mode a, output mode b and internal amplifier mode f, assuming a gain factor $G \gg 1$. The multi-mode results are exactly analogous, only the notation is quite cumbersome. While we worked with symmetrically-ordered operators above, it is more convenient here to begin with normally-ordered characteristic functions and the corresponding Glauber-Sudarshan quasi-distribution function.

[17] We will follow standard convention and refer to the quantum mechanical uncertainties in the vacuum state or coherent states *qua* displaced vacuum states as "vacuum fluctuations" or "zero-point motion," even though the mathematical formalism clearly identifies them as predictive uncertainties, not physical fluctuations.

We assume the input mode a is in a Glauber coherent state. Actually, it should be described in some statistical mixture of Glauber states, weighted by the particle distribution function, but again, that averaging will be deferred until later – in effect, we are here conditioning on supposed knowledge of the classical particle trajectories in the pickup wiggler.

Recalling that the combined density matrix for the input and amplifier factorize, and the input and internal operators all commute, the normally-ordered characteristic function of the output mode is then

$$\begin{aligned}\chi_b(\xi,1) &= \mathrm{Tr}\left[\rho_0 e^{\xi b^\dagger + \xi^* b}\right] e^{\frac{1}{2}|\xi|^2} = \mathrm{Tr}\left[\rho_a \otimes \rho_f e^{\xi(M^* a^\dagger + F^\dagger) + \xi^*(Ma+F)}\right] e^{\frac{1}{2}|\xi|^2} \\ &= \mathrm{Tr}\left[\rho_a e^{\xi M^* + a^\dagger \xi^* M a}\right] \mathrm{Tr}\left[\rho_f e^{\xi\sqrt{G-1} f + \xi^*\sqrt{G-1} f^\dagger}\right] e^{\frac{1}{2}|\xi|^2} \\ &= \chi_a(M^*\xi,1) e^{-|\xi|^2(G-1)} \chi_f(\sqrt{G-1} x i^*, 1)\end{aligned} \tag{5.154}$$

Now, for a coherent state input, it is easy to see that $\chi_a(M^*\xi,1) = \chi_{Ma}(\xi,1)$, i.e., the first term is the characteristic function for a deterministically scaled coherent state with expectation $\langle Ma\rangle = M\langle a\rangle$. The second term on the RHS is recognized as the normally-ordered characteristic function for a thermal (chaotic) state with an average of $\bar{n}_n = G-1$ photons. The final term is the characteristic function for the internal operator f. Considerations of the central limit theorem or of maximum entropy suggest that the added noise appearing through the initial state of f should, in the absence of further specific information, also be taken as thermal: $\chi_f(\xi,1) = e^{-\bar{n}_f|\xi|^2}$, where $\langle f\rangle = \langle f^2\rangle = \cdots = 0$ and we define $\bar{n}_f = \langle f^\dagger f\rangle \geq 0$, so the characteristic function becomes

$$\chi_b(\xi,1) = \chi_{Ma}(\xi,1) e^{-(\bar{n}_f+1)(G-1)|\xi|^2} \tag{5.155}$$

According to the convolution theorem, we see that after amplification, the resulting Glauber-Sudarshan P-function is a displaced chaotic state, or equivalently the sum of a multiple of the input coherent state an independent additive chaotic noise term corresponding to about $(\bar{n}_f+1)\left(1-\frac{1}{G}\right)$ thermal photons at input, where $\bar{n}_f \to 0$ in the ideal limit. It is easy to verify that the Fourier transform of (5.155), i.e., the P-function, is nonnegative and normalizable – in fact it is just a Gaussian.

A Heuristic (Single-Mode) Picture

Keeping in mind that the quantum state of the radiation is not actually diagonal in a number state basis and that a more careful multi-mode analysis is really essential because

the relative bandwidth is not terribly small, we can nevertheless paint the following suggestive picture: before amplification, the radiation field from the pickup wiggler that would be seen by any one particle "consists" of, on average, the equivalent of: $O(\alpha)$ self-radiated photons constituting the cooling signal; $\frac{1}{2}$ photon of "vacuum fluctuations," and $O(N_s\alpha) \gtrsim 1$ photons radiated from other particles in the sample. The relevant signal-to-noise ration (SNR) is approximately $\frac{\alpha}{\alpha N_s + \frac{1}{2}}$.

After amplification in an ideal, quantum mechanical phase-preserving linear amplifier, with overall gain factor $G \gg 1$, the field seen by the same particle is essentially classical but stochastic, containing the equivalent of $O(G\alpha)$ photons with proper phasing on average for cooling, $\frac{1}{2}G$ randomly-phase photons from the amplified vacuum fluctuations, at least another $\frac{1}{2}G$ photons from the noise added by the internal amplifier mode or modes, and $O(GN_s\alpha)$ from amplification of the other particle's radiation, with approximately $O(G)$ photons at each of about $O(N_s\alpha)$ different random phases. The SNR becomes approximately $\frac{\alpha}{\alpha N_s + 1}$, comparable to, but slightly worse than, the pre-amplified SNR.

Quantum amplifiers do not improve the SNR, but ideally amplify the input without serious degradation of the SNR, up to some sufficiently high level, where it can be be measured, manipulated, or employed by course-grained classical means without any significant further loss of SNR by quantum back-action. But the essential point is that neither the pre-amplified or post-amplified SNR is not "catastrophically" affected by quantum mechanical fluctuations or uncertainties inherent in emission, or amplification, but the effects are just roughly equivalent to having an extra $O(\alpha^{-1})$ particles in the sample. Of course, this is not precisely true, because the noise photons added by the amplifier will not have exactly the same spectrum or quasi-distribution function as the photons actually radiated in the wiggler, but if $N_s \gtrsim O(10)$, the Central Limit Theorem suggests that the net field from all the other particles in the sample will be close to a Gaussian chaotic state.

The Bottom Line on Quantum Amplification and the Amplified Field

Conventional linear phase-preserving amplifiers do not first measure then multiply photon number state, but instead linearly amplify the entire input field, while unavoidably adding additional noise, but not necessarily any multiplicative noise. In each mode, the additive noise is equivalent at input to the original vacuum fluctuations, themselves equivalent to about "half a photon," plus a second independent set of vacuum fluctuations, roughly equivalent to another "half a photon."

As Caves[195] aptly put it, "quantum mechanics extracts its due twice." After amplification in a high-gain amplifier, the two non-commuting quadrature components of a field mode may be measured classically with little further degradation in the SNR. But simultaneous measurement of non-commuting variables must involve auxiliary variables that do commute, and whose uncertainties add to the uncertainty already present.[18]

Once linearity is demanded of the amplifier, it must amplify input vacuum fluctuations along with any signal, and indeed cannot distinguish between what we, the cooling system, or the beam particles might consider signal versus noise. There are of course states of the combined system consisting of the input fields modes and the amplifier for which the additional noise in the form of a second set of vacuum fluctuations is not present in the output, but they will necessarily involve correlations between the input and internal modes, but because any internal amplifier modes must be prepared independently of the input, their zero-point motion will be amplified and added to the output (physically, through the mechanism of spontaneous emission.)

But apart from this doubled additive noise, quantum mechanics does not require anything like the speculated multiplicative Poissonian noise due to "quantum jumps" that we found so deleterious in our naive model.

Despite naive fears to the contrary, the signal does contain the needed phase information, albeit corrupted by a large amount of noise which, however, is random in phase (unbiased), independent of the input state, and purely additive.

The noise added by the amplifier can be characterized by an effective temperature assuming it roughly thermal in distribution, or as an equivalent number of photons at input, or perhaps most approximately but most usefully, as an equivalent number of extra particles in the sample. Because the cooling signal in the coherent portion of the wiggler radiation spectrum was approximately $O(\alpha^{-1})$ photons on average before amplification, while the total noise power is about $O(1)$ photons, the amplifier noise is expected to be the equivalent of approximately $N_n \sim O(\alpha^{-1})$ extra particles in the sample. While not nearly as detrimental to cooling as the noise predicted from the quantum jump model, this will eventually saturate improvements expected from further beam stretching, and limit the ultimate rates of cooling and the equilibrium emittances and energy spreads achievable.

When the gain is large, the radiation density matrix for the amplified field , conditioned on a given set of particle trajectories, is described as a biased chaotic state, essentially

[18] As a corollary, we deduce that if somehow only one phase is amplified, then no additional noise at input need be added.

indistinguishable from a classical but stochastic light field consisting of the superposition of the deterministically-scaled classical pickup radiation, together with some admixture of chaotic light characterized by a certain noise temperature proportional to the overall gain. Because the Glauber-Sudarshan P function of such a displaced chaotic state is everywhere nonnegative and well-behaved, the field can be interpreted classically.

In fact, the post-amplified field values can be taken as Gaussian c-number random variables with means

$$\langle \boldsymbol{E}(\boldsymbol{x},t)\rangle_{\text{post}} \approx \sqrt{G}\,\langle \boldsymbol{E}(\boldsymbol{x},t)\rangle_{\text{pre}} \tag{5.156}$$

and variances

$$\left\langle |\Delta\boldsymbol{E}(\boldsymbol{x},t)|^2 \right\rangle_{\text{post}} \approx G \left\langle |\Delta\boldsymbol{E}(\boldsymbol{x},t)|^2 \right\rangle_{\text{pre}} \tag{5.157}$$

so the SNR is approximately preserved. Calculation of the cooling dynamics can proceed classically, only with the added stochastic contributions included.

5.12 Cooling Dynamics for a Simplified Model

After using the hemi-classical radiation model and this very general formalism for linear amplifiers, determination of the interaction of the particles and amplified optical field within the kicker and calculation of the resulting statistics for cooling kicks on each pass can then proceed essentially classically. Because we are here primarily interested in questions of principle, and endeavor to demonstrate that quantum effects properly considered will not catastrophically affect cooling rates, the cooling statistics can be worked out explicitly, albeit for a highly idealized, one-dimensional model, wherein we make a number of simplifying assumptions in order to achieve tractability.

The average beam energy $\gamma_0 \gg 1$ is assumed highly relativistic, and the energy deviations $\delta\gamma_j = \gamma_j - \gamma_0$ are assumed sufficiently small, so that $|\delta\gamma_j| \ll \gamma_0$, and $|\gamma_j - \gamma_k| \ll \gamma_0$, and also $|v_{zj}(t) - v_0| \ll c$, where $v_0 = c\beta_0 \approx c\sqrt{1 - \frac{1}{\gamma_0^2}}$ is the mean longitudinal velcoity outside the wigglers. All particles then approximately have the same longitudinal velocity and the same magnitude of transverse quiver momentum inside the wigglers, and their trajectories can be be determined from conservation of energy and by transverse canonical momentum. Only longitudinal cooling is considered; the possibility of cooling of transverse degrees-of-freedom, by making part of the time-of-flight delay proportional to betatron errors, and exploiting dispersion in the lattice, is here omitted. Because the beam is highly

relativistic, the change in longitudinal momentum is related to the energy kick by

$$\frac{\Delta p_{zj}}{p_{z_j}} \approx \frac{\Delta p_{zj}}{p_0} \approx \frac{\Delta\gamma_j}{\gamma_0} \approx \frac{\Delta\gamma_j}{\gamma_j} \tag{5.158}$$

where $p_0 \approx mc\gamma_0\beta_0$ is the average longitudinal momenta of the beam particles.

In the kicker, the wiggler fields are much stronger than the amplified radiation fields from the pickup and vastly stronger than the newly-emitted spontaneous radiation. So the classical wiggler fields together with the particle initial conditions at the wiggler entrance are assumed to exclusively determine particle spatial trajectories, with negligible perturbations from the radiation field or space-charge fields. Both wiggler fields are assumed to be planar and perfectly transverse, with no transverse variation over the extent of the beam, and exactly sinusoidal in their longitudinal variation, requiring $\lambda_u \gg \sigma_\perp$ and $\frac{a_u}{\gamma_0}$ 1, both satisfied with orders-of-magnitude to spare. That is, the wiggler magnetic fields are assumed to be of the form $\boldsymbol{B}_u(\boldsymbol{x}) = \hat{\boldsymbol{y}}\sin\big(k_u(z-z_u)\big)$ over the extent $z_u \le z \le z + N_u\lambda_u$ of each wiggler, with fringe fields neglected.

Off-axis effects and all transverse motion of particles are ignored accept idealized 1D quiver motion in the wiggler field, as determined by canonical momentum conservation. Higher harmonics of particle motion in planar wiggler are ignored, and the longitudinal components of the unperturbed orbits are deduced from the transverse quiver and conservation of kinetic energy in a magnetic field. The effects of space-charge forces and any plasma waves or other collective effects the beam distribution are expected to be negligible and are ignored. Any external focusing is ignored as well, although in practice these forces may not be small.

Specifically, with the effects of energy variation and possible off-axis injection ignored, the transverse motion for each particle inside the wiggler is determined within our approximations by the conservation of transverse canonical momentum conservation in the wiggler field only:

$$\beta_{\perp j}(z) \approx \frac{1}{\gamma_0}\frac{q}{|q|}a_u\sin\big(k_u(z-z_u)\big) \qquad \text{for} \qquad z_u \le z \le z_u + N_u\lambda_u. \tag{5.159}$$

Because variations in longitudinal velocity between different particles are neglected, and the fast oscillating component of this velocity due to motion in a planar wiggler is ignored, for the longitudinal motion one finds

$$\beta_{zj}(t) \approx \beta_0\tfrac{1}{1+\lambda_0/\lambda_u}, \tag{5.160}$$

so that $\frac{1-\beta_{zj}}{\beta_0} \approx \frac{\lambda_0}{\lambda_u} \ll 1$, and the time spent in the kicker wiggler is then $T_j \approx \frac{N_u\lambda_u}{c\beta_0}$ for any particle.

Spectral broadening or other effects on the spontaneous emission spectra from the pickup, due to off-axis injection or the angular deviations associated with the transverse quiver motion are ignored, and the expected (i.e., classical) emission from each particle is assumed to be exactly N_u plane-polarized, sinusoidal oscillations at central frequency ω_0, emitted on-axis, with amplitude E_0, expected phase determined by the random longitudinal position of the particle in the beam, and extending transversely over some spot size w_0 larger than the transverse beam size and extent of a single particle's quiver orbit, i.e., for which $w_0 \geq \sigma_{b\perp}$ and $w_0 \geq \frac{a_u}{\gamma}\frac{\lambda_u}{2\pi}$. At best, the latter conditions are probably only marginally satisfied with the present parameters, but for simplicity we do neglect for now any transverse structure or imperfect transverse coherence in the radiation fields.

A single plane-polarized fundamantal transverse mode of the pickup radiation is assumed to have been amplified by an ideal laser with a flat gain $G > 1$ and negligible linear or nonlinear dispersion over the bandwidth $\Delta\omega_L \gtrsim \Delta\omega$. Outside the amplifier, all radiation is assumed to propagate purely longitudinally according to the vacuum dispersion relation. After amplification, the field is stochastic due both to the original beam shot noise and the added amplifier noise, but is essentially classical in its statistics and dynamic behavior, consistent with hemi-classical quantum optics detailed above.

The bypass lattice is assumed to introduce a time-of-flight variation linearly proportional to energy deviation $\delta\gamma_j = \gamma_j - \gamma_0$. The energy change in the kicker is calculated to leading order in $\frac{1}{\gamma}$ using the *unperturbed* quiver orbit of the particle in the kicker field. Any particular cooling kick on a single pass for a single particle will also be relatively small compared to the particle's kinetic energy, i.e., $|\Delta\gamma_j| \ll \gamma_0 - 1$, but not necessarily very small compared to the energy deviation $\delta\gamma_j$.

The bunch as a whole is assumed to contain many particles, $N_b \gg 1$, and after stretching the beam density ρ_b is assumed to be uniform over the entire stretched length, which is presumed very long compared to the sample length, i.e., $L_b \gg N_u\lambda_0$, so that the relative size of any sample is small: $N_s \ll N_b$. A Gaussian bunch with $\sigma_{b\|} \sim L_b$ is also analytically tractable.

Nearby particles will experience similar kicks and become partially correlated, but if the sample sizes are small this effect is negligible when averaged over all particles in the beam. So if good mixing between cooling kicks is assumed, so the noise effects in a given particle's kicks can be taken to be statistically independent from kick to kick, and the full kicks (including coherent and incoherent parts) can be taken to be effectively uncorrelated for different particles, and only the single-particle reduced distribution function at the present

moment in time need be tracked. In fact, the energy distribution will be assumed to start remain Gaussian, so that only first-order and second-order moments need by retained.

Many of these assumptions are fully justifiable in our parameter regime, while others are rather drastic and will eventually need to be relaxed in a more realistic calculation, but should be adequate for our intended proof-of-principle.

5.12.1 Bunch Stretching and Compression

We model the initial bunch stretching and final bunch re-compression very simply, as symplectic maps which ultimately leave the transverse phase space distribution unchanged, and preserve the action of the longitudinal phase space. Recall that the stretching is accomplished by drift in a ring with a high-momentum-compaction-factor ring to disperse the beam based on energy spread, and the use a linear accelerator (LINAC) to remove the resulting head-to-tail correlation between position and energy. After cooling, the process is reversed to re-compress the beam. Obviously, the transverse phase space cannot remain unaffected or uncorrelated with the longitudinal phase space at intermediate times, during the actual stretching or compression, because particles with greater longitudinal momenta are being forced onto longer orbits by highly-dispersive beam optics, but it can hold to a reasonably good approximation after the expansion or contraction is completed. We assume that the ring is sufficiently long to avoid "wrap-around" effects, which will lead to striations in longitudinal phase space. The dispersive drift shears the occupied phase space horizontally, while the ramped kick from the linac can shear it more or less vertically. Assuming all the head-to-tail energy variation is removed by the linac, and correlations between transverse and longitudinal degrees-of-freedom that build up at intermediate times are undone by proper lattice design, then from Liouville's theorem, and if t_b and t_a are some times before and after the stretching, then

$$L_b(t_a)\delta\gamma(t_a) \approx L_b(t_b)\delta\gamma(t_b). \tag{5.161}$$

If the bunch length satisfies $L_b t_a \gg L_b(t_b)$, then the energy spread satisfies $\delta\gamma(t_a) \ll \delta\gamma(t_b)$. Not only does stretching lower the beam density n_b and the resulting sample size N_s, but it also (reversibly) reduces the range of energy variation $\delta\gamma$, which relaxes design constraints on the bypass lattice. If OSC (irreversibly) reduces the energy spread by some factor μ by time t_c without affecting the overall bunch length, i.e., $\delta\gamma(t_c) \approx \mu\delta\gamma(t_a)$ while $L_b(t_c) \approx L_b(t_a)$, and then the bunch is re-compressed to its original length by the final time t_f, then

$$\delta\gamma(t_f) \approx \delta\gamma(t_c)\frac{L_b(t_c)}{L_b(t_f)} = \mu\delta\gamma(t_b), \tag{5.162}$$

so the final energy spread after expansion remains reduced by the same factor compared to the initial spread.

If $N_b \sim O(10^9)$ and $L_b(t_a) \sim O(10^{-1}$ m), then to get down to sample sizes of $N_s \sim O(10^2)$ for $N_u \sim O(10)$ and $\lambda_0 \sim O(10^{-6}$ m) requires a stretching factor of $\frac{L_b(t_b)}{L_b(t_a} \sim O(10^3)$. But the time required for a relativistic beam to traverse a compaction ring of circumference $C_0 \gtrsim L_b(t_a) \sim O(10^2$ m) is already about $t \sim O(10^6$s). The problem is that if we try to damp for several e-foldings such that $\mu = O(10^{-2})$, then the when it comes time to re-compress the pulse after cooling, we naively expect the compression to take $O(\mu^{-1}) \sim O(10^2)$ longer in the same lattice (with the same dispersion function), because the energy deviations are proportionally smaller. So the stretching and compression is not entirely trivial, but we leave linger questions for another day.

5.12.2 Energy Kick Statistics for Individual Particles

Within this simple model, the first and second order statistics of the cooling kicks for individual particles can be worked out analytically by lengthy but straightforward calculations, at least to leading order in certain small quantities, so we will summarize the results without presenting the unilluminating mathematical details.

First-Order Statistics

Within the kicker, recall that the effects of both beam self-fields and any newly-emitted spontaneous radiation can be assumed negligible compared to those of the wiggler fields and amplified radiation from the pickup wiggler. The latter fields, while highly amplified, still have a small effect on the actual *spatial* trajectories followed by the particles compared to the effect from the static wiggler field, so each single-particle orbit can be assumed to be determined by initial conditions at the entrance to the kicker and by the Lorentz forces produced by the wiggler field only. Of course, the whole point of the kicker interaction is to change the *energy* of the particle, and any such kick must arise from work done by a component of the optical electric field $\boldsymbol{E}(\boldsymbol{x}, t)$ in the kicker parallel to the particle velocity.

The total energy change inside the kicker for the jth particle during a given pass through the cooling section can then be estimated as:

$$mc^2 \Delta\gamma_j = q \int_{t_j+\Delta t_j}^{t_j+\Delta t_j+T_j} dt \; \boldsymbol{v}_{\perp j}(t) \cdot \boldsymbol{E}\big(\boldsymbol{x}_j(t), t\big), \tag{5.163}$$

where q is the (signed) charge of the particle; $\boldsymbol{x}(t)$ is the particle trajectory inside the wiggler and $\boldsymbol{v}_j(t) = v_{zj}(t)\hat{z} + \boldsymbol{v}_{\perp j}(t)$ is the particle velocity, both of which can be approximated assuming the dynamics are governed only by the wiggler field; $\boldsymbol{E}(\boldsymbol{x}, t)$ is the electric field of the amplified pick-up radiation as it propagates through the second wiggler; t_j was the arrival time of the particle at the entrance of the pickup wiggler; Δt_j is the total time delay between arrival at the first wiggler and arrival at the second wiggler, depending on the time-of-flight delay introduced in the bypass lattice; and T_j is the duration of time spent in the second wiggler. That is, the energy kick due to the interaction of optical electric field with the transversely quivering particle can be calculated in a perturbative fashion, in which to leading-order the effects of the changing energy on the particle orbit are neglected.

Neglecting various small terms , the conditional average energy kick for the jth particle, given values for the reference energy and of its own pre-kick energy deviation, and averaged over all intrinsic quantum mechanical uncertainty in the radiation as well as over the classical shot noise, i.e., random arrival times of other particles in the its sample, is given by

$$\langle \Delta\gamma_j \,|\, \gamma_j, \gamma_0 \rangle \approx \tfrac{q a_u N_u \lambda_u \sqrt{G} E_0}{2mc^2 \beta_0 \gamma_0} \Theta\left(1 - \tfrac{|\phi_{jj}|}{2\pi N_u}\right) \left[1 - \tfrac{|\phi_{jj}|}{2\pi N_u}\right] \sin(\phi_{jj}), \tag{5.164}$$

where $\Theta(s)$ is the Heaviside step function, $\phi_{jj} = \phi_{jj}(t_j, \gamma_j, \gamma_0)$ is the phase of the jth particle's amplified self-field as seen by that same particle at the entrance to the kicker, while t_j is the time of arrival of particle j at the pickup.

Note that ϕ_{jj} is determined by the particle's time-of-flight through the pickup wiggler and bypass lattice compared to the time-of-flight for the pickup radiation during propagation and amplification and any further phase-shift introduced by the laser amplifier system. If the delay in the bypass lattice is arranged such that for sufficiently small energy deviations $\delta\gamma_j$, the relative delay leads to $q \sin(\phi_j) \propto -\delta\gamma_j + O(\delta\gamma_j^3)$, then the particle will experience a restoring force on average, tending to push the energy toward that of the reference orbit. Ideally, the probable range of variation in $\delta\gamma_j$ before the cooling kick is mapped into just $\pm\frac{\pi}{2}$ of phase delay, so that the magnitude of the damping force grows monotonically with the field gain $\sqrt{G}$ and the magnitude $|\delta\gamma_j|$ of the energy deviation.

Second-Order Statistics

After some more algebra, the corresponding conditional expectation of $\Delta\gamma_j^2$ becomes approximately:

$$\begin{aligned}\left\langle (\Delta\gamma_j)^2 \,|\, \gamma_j, \gamma_0 \right\rangle \approx \left(\tfrac{q a_u N_u \lambda_u \sqrt{G} E_0}{2mc^2\beta_0\gamma_0}\right)^2 \Bigg[& \Theta\Big(1-\tfrac{|\phi_{jj}|}{2\pi N_u}\Big)\Big(1-\tfrac{|\phi_{jj}|}{2\pi N_u}\Big)^2 \sin^2(\phi_{jj}) \\ & + \tfrac{N_b}{L_b}\tfrac{N_u\lambda_0}{3}\Big(1-\tfrac{3}{8\pi^2 N_u^2}\Big)\Theta\Big(1-\tfrac{|\phi_{jj}|}{k_0 L_b}\Big)\Big(1-\tfrac{|\phi_{jj}|}{k_0 L_b}\Big) \\ & + \Big\{ \tfrac{1}{N_{\rm ph}}\tfrac{1}{\pi^2}\tfrac{N_u^2}{b_L(b_L^2-4N_u^2)}\sin^2(\pi b_L) \\ & + \quad \tfrac{1}{N_{\rm ph}}\tfrac{1}{8\pi}\left\{2\,\mathrm{Si}\,(2\pi b_L) - \mathrm{Si}\,(2\pi[b_L-2N_u]) - \mathrm{Si}\,(2\pi[b_L+2N_u])\right\}\Big\}\Bigg],\end{aligned} \tag{5.165}$$

where $b_L = \frac{1}{2}N_u\frac{\Delta\omega_L}{\omega_0}$ is a measure of the amplifier bandwidth (FWHM) relative to the coherent bandwidth of wiggler radiation, and is $O(1)$, while $\mathrm{Si}(x) = \int_0^x du\,\frac{\sin(u)}{u}$ is the sine integral. Note that $Si(-x) = -\mathrm{Si}(x)$, and for $|x| \lesssim \frac{\pi}{2}$, $\mathrm{Si}(x) \approx x - \frac{1}{18}x^3$, while for $x \gtrsim \frac{\pi}{2}$, $\mathrm{Si}(x)$ undergoes decaying oscillations, peaking at $\mathrm{Si}(\pi) \approx 1.852$ and asymptotically approaching $\mathrm{Si}(x) \to \frac{\pi}{2}$ as $x \to \infty$. The first term on the RHS of (5.165) is due entirely to self-fields, and is just the square of the first-order conditional average; the second term is the contribution from shot noise, reflecting heating from other particles within the sample, while the final terms reflect the minimal amplifier noise consistent with quantum mechanics.

In principle, the conditional two-particle correlations $\langle \Delta\gamma_j \Delta\gamma_k \,|\, \gamma_j, \gamma_0 \rangle$ are also needed to determine the evolution of the beam energy spread, but with good mixing, the effect of these cross-terms will be negligible when averaged over all the particles in the beam, so we need not actually calculate them explicitly.

5.12.3 From Single Particle Statistics to Beam Properties: Evolution of the RMS Beam Energy and Energy Spread

Because the signal-to-noise ratio is so high, the magnitude and even direction of any give particle's energy kick on any one pass are subject to large uncertainty, and cannot be estimated reliably. In conventional RF stochastic cooling, relatively small kicks delivered over many passes lead via a time-averaging or smoothing procedure to a continuos-time Fokker-Planck equation. Any one kick for any one particle cannot be predicted with much precision, but the cumulative effect of many kicks enjoys the usual advantages of the law of large numbers, and can be reliably determined.

In fast optical stochastic cooling, the beam is cooled with a relatively small number of relatively large kicks, so time-averaging is less justified and less effective at improving predictability. However, particles remain largely uncorrelated during the cooling, so while prediction of individual particle behavior is unreliable, averaging over all particles in the beam is expected to give $O\left(N_b^{-1/2}\right)$ improvement in predictive accuracy.

In analyzing the stochastic evolution of beam properties, we should be careful to distinguish between arithmetic means over all beam particles so as to define *intensive* beam properties such as the average velocity, energy, or energy spread, or (classical an/or quantum mechanical) statistical averages (expectation values) over our uncertainty in either individual particle or collective observables.

If t_- is some time just before the beam enters the cooling section, at which point the average beam energy is γ_0, and t_+ is a time just after it has passed through the cooler but before it reaches the next cooling section, then the net change in the mean particle energy can be predicted by further averaging the conditional averages (5.164) over the energy distribution for individual particles and over the particles in the bunch. The mean energy per particle of all beam particles beam at any time t is

$$\bar{\gamma}(t) = \frac{1}{N_b}\sum_{j=1}^{N_b}\gamma_j(t), \tag{5.166}$$

which remains a "random variable." Since all the beam particles are assumed identical, The expectation value of this mean per-particle energy after a cooling pass is

$$\langle\bar{\gamma}(t_+)\rangle = \langle\gamma_j(t_+)\rangle = \langle\gamma_j(t_-)\rangle + \langle\Delta\gamma_j\rangle\,, \tag{5.167}$$

where

$$\langle\Delta\gamma_j\rangle = \langle\langle\Delta\gamma_j\,|\,\gamma_j,\gamma_0\rangle\rangle \tag{5.168}$$

is the conditional average (5.164) averaged further with respect to the remaining random variables, i.e., with respect to the probability distribution for pre-kick particle energies. Ideally, the cooling will not change the mean beam energy but will only redistribute energy from more energetic to less energetic particles:

$$\langle\bar{\gamma}(t_+)\rangle = \langle\gamma_j(t_+)\rangle \approx \langle\gamma_j(t_-)\rangle = \gamma_0. \tag{5.169}$$

In order to determine either the expected decrease in energy spread due to cooling, or the uncertainty in our prediction (standard error) for the per-particle beam energy, second-order statistics for $(\Delta\gamma_j)^2$ are also needed, but these two quantities should not be conflated:

uncertainties and fluctuations are not the same thing. Since statistics for the different beam particles are approximately independent, we expect that the standard error of prediction for $\bar{\gamma}(t)$ will be $O(\frac{1}{\sqrt{N_b}})$ better than that for any one particle.

Evolution of the mean per-particle energy has nothing to do with actual cooling *per se*: beam slowing is not the same thing as beam cooling. The former damps the average eenrgy, while the latter damps deviations or fluctuations about the average energy. To assess the latter, we need to consider second moments.

First consider the unconditional variance of the jth particle's energy, defined by

$$\sigma_j^2(t) = \mathrm{Var}[\gamma_j(t)] = \left\langle [\gamma_j(t) - \langle \gamma_j(t) \rangle]^2 \right\rangle, \tag{5.170}$$

$$\begin{aligned} \sigma_j^2(t_+) = \sigma_j^2(t_-) &+ \left[\left\langle \left\langle (\Delta\gamma_j)^2 \,|\, \gamma_j, \gamma_0 \right\rangle \right\rangle - \langle \Delta\gamma_j \rangle^2 \right] \\ &+ 2 \left[\langle \langle \Delta\gamma_j \,|\, \gamma_j, \gamma_0 \rangle \, \delta\gamma_j(t_-) \rangle - \langle \Delta\gamma_j \rangle \, \langle \delta\gamma_j(t_-) \rangle \right] \end{aligned} \tag{5.171}$$

If $\langle \Delta\gamma_j \rangle \approx 0$, then this can be rearranged and simplified to yield:

$$\sigma_j^2(t_+) - \sigma_j^2(t_-) = 2 \langle \Delta\gamma_j \delta\gamma_j(t_-) \rangle + \left\langle (\Delta\gamma_j)^2 \right\rangle, \tag{5.172}$$

where $\left\langle (\Delta\gamma_j)^2 \right\rangle = \left\langle \left\langle (\Delta\gamma_j)^2 \,\middle|\, \gamma_j, \gamma_0 \right\rangle \right\rangle$ and $\langle \Delta\gamma_j \delta\gamma_j(t_-) \rangle = \langle \langle \Delta\gamma_j \,|\, \gamma_j, \gamma_0 \rangle \, \delta\gamma_j(t_-) \rangle$. Since for small energy deviations, the kick will satisfy $\Delta\gamma_j \sim -\delta\gamma_j$ by design, the first term on the right-hand side of (5.172) is negative and reflects the impact (on our uncertainty, not necessarily the particles) of cooling due to the coherent self-interaction, while the second term is always positive and represents the effects of heating. However, it is important to note that $\sigma_j(t_+)$ actually quantifies changes in uncertainty in the jth particle's energy after being subjected to the stochastic kick, and is not equal to the change in the particle's actual but unknown energy deviation $\delta\gamma_j(t_f)$, although it is equal in value to the RMS of the physical deviation.

But in the present context, cooling means reducing the actual extent of energy deviations amongst the beam particles, not reducing our uncertainty in what the particle energies might be, so (5.172) is not necessarily equal to the real quantity of interest for cooling.

Longitudinal cooling can be assessed by changes in some measure of the actual fluctuations about the mean energy, such as the RMS energy spread $\delta\gamma(t)$ of particles in the bunch at time t, defined by:

$$\delta\gamma(t) = \left[\frac{1}{N_b} \sum_{j=1}^{N_b} \left(\gamma_j(t) - \bar{\gamma}(t) \right)^2 \right]^{1/2}. \tag{5.173}$$

Because the kicks are stochastic, the $\delta\gamma_j(t_+)^2$, and hence $\delta\gamma(t)$ cannot be known exactly, but the expectation value of the latter can, and is approximately given by:

$$\left\langle [\delta\gamma(t)]^2 \right\rangle \approx \sigma_j^2(t). \tag{5.174}$$

That is, on average, the RMS energy spread of the bunch is approximately equal in value to the single-particle energy variance, as might be expected from a bunch comprised of particles which are assumed statistically independent. But while the magnitude and direction of any particular particle's kick are subject to large uncertainty and cannot be estimated reliably, averaging over all the beam particles ensures that (5.174) will be quite accurate as a point estimate for the overall RMS energy spread within the beam. The exact calculation require knowledge of still higher-order moments, but assuming a Gaussian moment closure, we can obtain the estimate $\mathrm{Var}\left[\delta\gamma(t_+)^2\right] \sim O\left(\frac{2}{N_b}\right)\sigma_j^2(t_+)$, so the relative uncertainty is expected to be $O\left(\frac{1}{\sqrt{N_b}}\right) \ll 1$.

Defining the instantaneous cooling rate τ_c^{-1} for energy spread as:

$$\tau_c^{-1}(t) = -\tfrac{1}{2}\frac{d}{dt}\ln\left[\left\langle (\delta\gamma(t))^2 \right\rangle\right], \tag{5.175}$$

a formal expression for the approximate energy cooling rate averaged over the current cooling pass can be found:

$$\begin{aligned} \tau_c^{-1} &\approx -f_0 \frac{\left\langle \delta\gamma(t_+)^2 \right\rangle - \left\langle \delta\gamma(t_-)^2 \right\rangle}{\left\langle \delta\gamma(t_-)^2 \right\rangle} \\ &= f_0 \left[\frac{\left\langle \left\langle \Delta\gamma_j \,|\, \gamma_j(t_-), \gamma_0 \right\rangle \left[-\delta\gamma_j(t_-)\right] \right\rangle}{\left\langle \delta\gamma(t_-)^2 \right\rangle} - \frac{1}{2}\frac{\left\langle (\Delta\gamma_j)^2 \right\rangle}{\left\langle \delta\gamma(t_-)^2 \right\rangle} \right], \end{aligned} \tag{5.176}$$

where again f_0 is the frequency of cooling kicks experienced by the bunch. Since $\Delta\gamma_j \sim \sqrt{G}$, Note that this is exactly of the form of equation (5.1).

If the particle energy distribution is assumed to remain Gaussian, and the time-of-flight delays are assumed to be linear functions of the energy deviations, then (5.176) can be calculated analytically. Assuming optimal choices for gain and time-of-flight delays, the expressions become essentially identical in form to (5.18), with $N_s \approx \frac{1}{3}\frac{N_b}{L_b}N_u\lambda_0$, and N_n assuming a quite complicated explicit form, but for which typically $N_n \sim O(\alpha^{-1})$. If the Gaussian approximation is not valid, then the distribution functions can be evolved numerically, but in this case higher-order moments may be needed, which are increasingly cumbersome to calculate.

The cooling rate is essentially that predicted classically, except for some unavoidable noise from the quantum amplifier which appears additively, roughly as the equivalent of an

extra $O(\alpha^{-1})$ or so particles in the sample, not as a multiplicative degradation as in the rate (5.19) in the Poissonian emission model, or worse, as in (5.21) with quantum jumps and Heisenberg phase noise.

To estimate the uncertainty in this predicted cooling rate, we would need to know fourth-order moments of the energy kicks and energy deviations, which can be laboriously worked out if needed, but again assuming Gaussian distributions, the estimated rate can be shown to be statistically reliable because of the now familiar $O\left(\frac{1}{\sqrt{N_b}}\right)$ scaling.

However, note that (5.176) should not necessarily be interpreted as the reciprocal of an exponential decay time, because of the continuum approximation. With traditional, slower stochastic cooling methods, typically the effects of many ($O(10^3)$) kicks must be accumulated before any appreciable reduction in phase space occurs whatsoever, so over such long times scales (compared to f_0^{-1}), cooling kicks can be assumed to be effectively smeared out more or less continuously in time, and the dynamics can be approximated by a Focker-Planck type differential equation. Here, each individual kick is much bigger, but fewer are imposed, and accurate cooling estimates really require evaluation of the difference equations.

5.12.4 Final Cooling in Muon Accelerator

With the classical model essentially confirmed apart from some lower bound on the additional additive amplifier noise that has its origins in quantum mechanics, rough estimates for cooling times achievable with given power, or power required to achieve given cooling rates can be determined. The demands of microsecond-scale cooling exceed present technology, but slower cooling would require more modest systems. Because our rough estimates agree with those presented in [165, 166], we just quote the results.

Cooling a $\gamma \sim O(10^3)$ beam with $N_b \sim O(10^9)$ particles per bunch, on a time-scale $\tau_c \sim O(10^{-7}$ s) in order to achieve an $O(10^{-3})$ relative decrease in longitudinal emittance would require stretching to a size where $N_s \sim O(10^2)$, and would require about $O(10)$ lasers amplifiers, each producing $O(10^2$ W) of average power at $O(10^2$ Hz) repetition rate, which is beyond the current capabilities of existing amplifiers, but perhaps achievable in the next few decades.

5.12.5 Some Limitations and Extensions

Transverse Cooling

Although for simplicity we do not treat transverse cooling here, we should point out that it is possible, and in fact can be performed simultaneously with longitudinal cooling, at the expense of a certain slow-down in the cooling rate if the same cooling sections are used for both longitudinal and transverse degrees-of-freedom.

As mentioned, the trick is to exploit dispersion. For sufficiently small deviations from the desired reference orbit, the full trajectory of a transverse coordinate (say the x coordinate) for a particular particle in the kicker is given by:

$$x_j(z) = \bar{x}_0(z) + x_{j\,\beta}(z) + \tfrac{\delta\gamma_j}{\gamma_0} D_x(z), \tag{5.177a}$$

$$x'_j(z) = \bar{x}'_0(z) + x'_{j\,\beta}(z) + \tfrac{\delta\gamma_j}{\gamma_0} D'_x(z), \tag{5.177b}$$

where $\bar{x}_0(z)$ is the reference orbit in the cooling section (including the expected quiver motion), $x_{j\,\beta}(z)$ represents the typically slowly-oscillating (compared to the quiver) betatron orbit (transverse deviations from the reference orbit in the absence of dispersion), determined by the beam optics of the external transport lattice and the particle initial conditions, $\frac{\delta\gamma_j}{\gamma_0}$ is the relative energy deviation from the reference orbit, and $D_x(z)$ is a dispersion function depending on the lattice and the insertion device (i.e., kicker wiggler) reflecting the fact that particles with different energies are bent differently in magnetic fields, while primes indicate derivatives with respect to the longitudinal position, which is just taken as z in the straight kicker section.

If the delay in the bypass between pickup and kicker is made proportional (in part) to the betatron amplitude, i.e., either $x_{j\,\beta}$ or $x'_{j\,\beta}$, then the resulting energy kick $\Delta\gamma_j$ will also depend on these errors in the transverse orbit. If the kicker is located in a region of appreciable dispersion, then as $\delta\gamma_j$ changes due to work done by the amplified radiation on the particle as it quivers in the kicker, $x_{j\,\beta}(z)$ and $x'_{j\,\beta}(z)$ must also change in compensation, since $x_0(z)$ and $x'_0(z)$ are fixed, and the overall transverse position $x_j(z)$ and angle $x'(z)$ must remain continuous. If the phasing is arranged properly, in this way the particle can be nudged onto a smaller-amplitude betatron orbit, that will be apparent when it emerges into a region of less dispersion.

Neglected Optical Effects

For the proof-of-principle argument here, we have assumed flat amplitude gain, and 1D radiation fields propagating in vacuum without diffraction or dispersion. Since the structure and phase of these fields is essential to the success of the transit-time OSC scheme, more careful calculations will eventually be needed to make quantitative predictions.

Dispersion

Since any cooling depends crucially on careful phasing, any dispersion in the amplifier or the optical transport system will diminish the effectiveness of the OSC. The times-of-flight can of course be adjusted to compensate for any overall average delay for the frequency band, linear group-velocity dispersion and any nonlinear phase-modulation within the amplifier are expected to be deleterious to some degree. Such dispersion can come from the usual linear index-of-refraction of the medium or the optics, or from the phase shifts that are proportional to the gain (amount of stimulated emission in the medium), which in practice is frequency-dependent.

Diffraction/Transverse Effects

We have here effectively assumed that the amplified wiggler radiation consists of transverse plane waves, but even for well-collimated particle beams where $\sigma_\perp \ll w_0$, certain paraxial effects may become quite important.

Recall that the fundamental Gaussian vacuum paraxial mode, propagating along the z direction with focus assumed to be located at $z = 0$, can be written as

$$\boldsymbol{E}(\boldsymbol{x}, t) = E_0 \tfrac{w_0}{w(z)} e^{-i[k_0 z - \omega_0 t - \eta(z)] - r^2 \left[\frac{1}{w(z)^2} + \frac{ik}{2R(z)}\right]}, \tag{5.178}$$

where E_0 is the overall complex amplitude, k_0 is the carrier number, $\omega_0 = ck_0$ is the carrier frequency, $r = \sqrt{x^2 + y^2}$ is the transverse distance off-axis, $w(z) = w_0 \left[1 + \left(\frac{z}{Z_\mathrm{R}}\right)^2\right]^{1/2}$ is the local waist, $R(z) = z\left[1 + \left(\frac{Z_\mathrm{R}}{z}\right)^2\right]$ is the radius of wavefront curvature, $\eta(z) = \arctan\left(\frac{z}{Z_\mathrm{R}}\right)$ is the Guoy phase, and $Z_\mathrm{R} = \frac{1}{2} k w_0^2$ is the Raleigh range, where w_0 is the waist size at focus.

Issues of coupling efficiency suggest that both kicker and pickup wigglers should have identical values for λ_u and nearly identical values for N_u, and the Raleigh range of the amplified radiation in the kicker wiggler should be comparable to that the radiation emitted

in the pickup, by proper choice of collection and transport optics. As the latter is approximately $Z_{\rm R} \sim \frac{1}{8\pi} N_u \lambda_u$, diffraction will probably become appreciable over the interaction distance $N_u \lambda_u$ in the kicker.

The direct effects of the wavefront curvature, are probably safely ignored, and the intensity fall-off due to the expansion of the transverse waist can be accommodated, although somewhat higher powers than estimated in the $1D$ theory might be needed. Of most concern is the Guoy phase shift. If the laser focus is placed near the center of the wiggler to help minimize power requirements, the Guoy phase will vary rapidly from close to $-\frac{\pi}{2}$ to $+\frac{\pi}{2}$ inside the kicker, and the despite best efforts to ensure careful relative timing, the energy kick may tend to average away. This needs to be studied carefully. If the focus needs to be moved off-center or, worse, perhaps placed outside the kicker, by a distance comparable to the Raleigh range, then power requirements for OSC will become even more severe.

Near-Field Effects

With quantum-mechanical fears allayed, an accurate quantitative calculation will require accurate knowledge of the classical field. The paraxial Gaussian form may not be sufficiently accurate. For example, the solenoidal (transverse) part of the classical electric field (and hence the expectation value of the hemi-classical field) radiated by a single beam particle in the pick-up can be expressed in the well-known Liénard-Wiechert form:

$$\begin{aligned} \boldsymbol{E}_j(\boldsymbol{x},t) = q \left[\frac{\hat{\boldsymbol{n}}_j - \boldsymbol{\beta}_j}{\gamma_j^2 (1-\boldsymbol{\beta}_j \hat{\boldsymbol{n}}_j) R_j^2} \right]_{t'_j} - q \left(\frac{\hat{\boldsymbol{n}}_j}{R_j^2} \right)_t \\ + \frac{q}{c} \left[\frac{\hat{\boldsymbol{n}}_j \times \left\{ (\hat{\boldsymbol{n}}_j - \boldsymbol{\beta}_j) \times \dot{\boldsymbol{\beta}}_j \right\}}{(1-\boldsymbol{\beta}_j \hat{\boldsymbol{n}}_j)^3 R_j} \right]_{t'_j} \end{aligned} , \tag{5.179}$$

where $\boldsymbol{R}_j = \boldsymbol{x} - \boldsymbol{r}_j(t)$, $R_j = |\boldsymbol{R}_j|$, $\hat{\boldsymbol{n}}_j = \frac{\boldsymbol{R}_j}{R}$, $\boldsymbol{r}_j(t)$ is the particle trajectory, $\boldsymbol{\beta}_j(t) = \frac{1}{c}\dot{\boldsymbol{r}}_j(t) = \frac{d}{dt}\boldsymbol{r}_j(t)$ is the normalized particle velocity, $\dot{\boldsymbol{\beta}}_j(t) = \frac{d}{dt}\boldsymbol{\beta}_j(t)$ is the normalized acceleration, and t'_j, defined implicitly by $t = t'_j + \frac{R_j}{c}$, is the retarded time.

The far-field radiation pattern is determined by the asymptotic form of the final term only, which is proportional to the particle acceleration, and falls of like $O\left(\frac{1}{R}\right)$ in distance from the source. The form of the far-field brightness function (i.e., optical Wigner function) of synchrotron radiation from relativistic particles moving n bending magnets, for which $\frac{a}{\gamma_0} \ll 1$, has been worked out[167] in the paraxial case for regimes in which $a_u \ll 1 ll \gamma_0$ or $1 \ll a_u \ll \gamma_0$ for essentially any $N_u > 1$, or for essentially any $a_u \ll \gamma_0$ if $N_u \sim 1$ or $N_u \gg 1$.

In the presumed OSC regime, where $N_u \sim O(10)$ while $a_u \gtrsim 1$, an exact analytic form

is lacking, and the velocity fields may not be negligible. Since the radiation is emitted over the length $N_u\lambda_u$, we would naively expect that a Fraunhofer-type approximation to the angular spectrum can be deduced from the acceleration fields and will be valid if the radiation is observed at some suitably large distance R from the wiggler, such that $R \gg \lambda_0$, $R \gg N_u\lambda_u$, and $R \gg \frac{(N_u\lambda_u)^2}{\lambda_0}$. While the first condition is entirely trivial, the second is unlikely to hold, and the third will certainly not hold. In order to have time and room to amplify the radiation and feed it back onto the particle beam, the wiggler radiation emitted over $O(5\text{ m})$ will be collected, collimated, and imaged into the amplifier at a distance only about $O(1\text{m})$ from the end of the pickup, only a few or maybe ten times λ_u downstream, where near-fields will still be present, and their effects are not entirely clear.

5.13 Discussion

Despite fears that quantum uncertainties would prove devastating for optical stochastic cooling based on the amplification of very weak pickup signals, the technique of OSC can in principle work despite these quantum uncertainties. What, then, went wrong with the naive reasoning?

5.13.1 Why "Naive" Quantum Treatments Fail

Particles do *not* emit wiggler radiation described by photon number states or any statistical mixtures of of such states. They do not radiate into any states resembling a number state, but rather a Glauber coherent state, or a classical statistical mixture of such states, in which photon number is not sharply defined, but phase information is partially available and quantum mechanical uncertainties are minimal. These states can have arbitrarily low expectation values for photon number or energy, but still be every bit as "real' and "present" as a number state.

Particles do not radiate in a series of stochastic quantum jumps, only once every $O(\alpha^{-1})$ passes through the pickup wiggler. The radiation from each particle is present on every pass, even though its average energy may correspond to much less than one photon, and is available to be amplified by a quantum mechanical interaction, provided no intervening measurement first projects the state by actually trying to count photons. Moreover the wiggler radiation possesses, on average, exactly the amplitude and phase it would possess classically, so the "coherent signal" information necessary for transit-time cooling is effec-

tively always present, only corrupted by the equivalent of some additional additive (not multiplicative) noise needed to satisfy the uncertainty principle, noise that physically can be traced back to amplified spontaneous emission within the gain medium.

A linear amplifier does *not* function by multiplying photon number, either coherently or by projective measurement, but rather acts unitarily on the full quantum mechanical state of the EM field, scaling the field amplitude and adding some extra noise. In the pickup, the particles do not radiate whole numbers of photons, and in the kicker the particles do not measure or sense photon number, but respond directly to the field. Nothing actually prepares, emits, multiplies, evolves towards, measures, or counts photon number states, so the "discrete" nature of photons is not terribly important.

A careful analysis reveals that the naive quantum mechanical fears were misguided: the cooling rate which accounts for quantum mechanics essentially agrees with the classical rate calculation, except that the amplifier noise, which can in principle be made arbitrarily small classically, is constrained by the quantum mechanics to some non-zero lower bound.

This lower bound reflects the fact that the total fluctuations of the amplified field include the amplified input noise, which just consists of minimum-uncertainty vacuum fluctuations for coherent states, as well as added amplifier noise consisting of an independent set of amplified vacuum fluctuations. Input noise is amplified along with the signal because of linearity: a linear amplifier cannot know what part of the input state will later be regarded as noise or as signal. The additional noise is added as a consequence of the uncertainty principle: a phase-insensitive linear amplifier amplifies both quadrature components of a field mode, which do not commute, so at least one extra mode must be involved.

Therefore quantum mechanical noise does not prevent fast optical cooling, but it does limit how small the incoherent heating terms can be made. The resulting expressions and estimates for the cooling times agree with classical calculations, provided the amplifier noise is accounted for through N_n.

5.13.2 Synchrotron Radiation Damping

If the naive "quantum jump" model of discrete Poissonian photon emission failed so spectacularly for the case of OSC, why does it work so well for synchrotron radiation damping (SRD) of electrons in storage rings? In such a damping scheme, relativistic electrons of normalized energy γ orbiting in a ring of radius r_0 spontaneously emit radiation in a continuum of frequencies up to about $\omega_c \approx \frac{3}{2}\gamma^3 \frac{c}{r_0}$, confined to an angular cone of about

$O(\frac{1}{\gamma})$ with respect to the electron orbit. Per revolution, each electron emits the equivalent of about $N_{\rm ph} \sim \frac{5\pi}{\sqrt{3}}\gamma\alpha$ per revolution, at an average energy $\langle\hbar\omega\rangle \sim \frac{8}{15\sqrt{3}}\hbar\omega_c$. For $O(\text{GeV})$ electron beams, corresponding to $\gamma \sim O(10^4)$, and circulating in rings with $r_0 \sim O(10^2$ m), each electron radiates the equivalent of about $N_{\rm ph} \sim O(6 \cdot 10^2)$ photons per revolution, at wavelengths down to about $\lambda_c = \frac{2\pi c}{\omega_c} \sim O(10^{-10}$ m), corresponding to $O(\text{keV})$ x-rays.

On average, each electron loses momentum along its direction of motion (parallel to its betatron orbit) due to the recoil form this emission. If they are then re-accelerated along the beam-line, transverse emittance is reduced.

Ultimately cooling is limited by quantum mechanical "randomness" in the emission, which, consistent with fluctuation-dissipation relations, leads to an incoherent heating term in addition to the damping term, which limits the cooling rate as well as the ultimate emittance that can be achieved in equilibrium. In a pioneering and influential study, M. Sands [168] successfully modeled these effects by using what we have called the "quantum jump" model, assuming that electrons emit radiation in the form of a whole number of quanta in random Poissonian fashion.

In hindsight, it is clear that the OSC and SRD occur in completely different regimes and rely on rather different physics. In the OSC, each particle emits on any pass the equivalent of very few photons of small momentum, so direct radiation reaction effects are completely negligible. In SDR, each electron emits the equivalent of many photons of high momentum, and recoil effects in the virtual Compton scattering are significant – in fact, the large cumulative effects of recoil over several turns are precisely what is of concern.

In OSC, the radiated field is collected, coherently amplified, and fed back onto the particles that emitted it, to induce an energy kick far larger than any direct losses due to emission, so the quantum state of the radiation matters greatly, while in SDR there is no feedback other than the individual electron recoils, once emitted radiation propagates away as far as the particles are concerned, and in fact one could in principle "trace out" the field degrees-of-freedom to obtain a non-Hamiltonian, diffusive evolution equation for the particles only.

In neither case do we actually expect the particles to radiate a whole number of quanta in discrete emission events, but in the case of SDR, whether the radiation is described by a number state or coherent state or some other state is largely irrelevant, as long as we get the low-order moments right. Many different microscopic stochastic models can lead to the same Forker-Planck equation. In effect, the Sands technique gets away with modeling

a diffusion equation by a particular choice of a random walk model for emission which does not faithfully describe the actual emission process or the resulting radiation phase space, but mimics the proper statistics for the electron phase space.

If simple scaling arguments and back-of-the-envelope calculations predict that the discrete nature of the *energy* eigenstates of the field might be important, then rather than justifying the quantum jump model, this only gives more reason to adopt a fully quantum treatment of the radiation, at least if radiation properties statistics are themselves needed, because quantum mechanics evolves the radiation modes or any system unitarily, not by a series of quantum jumps. If the radiation itself is never examined, but only the accumulated effects of the back-action on the particles during emission or absorption, then sometimes the Poissonian model will be useful, as with the Sands treatment of synchrotron radiation damping, or with the more sophisticated Poissonian emission model used by Friedland [221] to tease out the small-signal gain in FELs.

5.13.3 Is the Field Evolution Really Unitary Throughout?

The metaphor of corpuscular photons and quantum jumps may be alluring, but can be deeply misleading, in stochastic cooling as well as certain other applications of wiggler radiation. Benson and Madey[172] assert

> the wiggler is capable of reducing the wavefunction [of a particle] by causing it to emit a photon. The electrons position can therefore be deduced from subsequent measurements.

Reluctant as we are to disagree with pioneers of FEL physics, we are of the opinion that, as a reasonably accurate description of our original, naive opinions concerning quantum mechanical effects of the wiggler radiation, this statement is completely incorrect in the current context. Now we contend that the wiggler magnet itself does not in any way collapse or reduce any particles wavefunction – only a subsequent photon-counting or other optical measurement will do that, which never happens in the context of OSC. Until a reversible macroscopic measurement of particles or radiation is made, the EM fields and beam particles evolve unitarily. Besides, with many electrons in a sample length, observation of a photon can tell us very little about the position or momentum of any one electron, so subsequent measurement only partially project a particle wavefunction. Nor does the wiggler cause the particle to emit a photon, if by photon we mean either a number state of the electromagnetic field, or even just a state of definite energy and/or momentum. In most practical regimes

of interest, the particles in a wiggler will generally emit radiation described by statistical mixtures of Glauber coherent states, which behave far differently (more classically) than number states. If recoil or other effects become important, than the quantum state of the radiation may become highly entangled, or may exhibit other features associated with non-positive Glauber-Sudarshan quasi-distributions that cannot be reproduced classically, but there is no reason to think the states of the field will be close to eigenstates of either the number or energy operators or diagonal mixtures of these.

If photon-counting experiments are chosen to be performed, then whole numbers of photons will of course be measured, but this has little to do with what state described the radiation before measurement. Of course the calculations of the field dynamics and field statistics can be performed in number state basis, or any other basis, but the density matrices will not generally be sparse in any sense, so the P-function representation is particularly convenient.

Other fields learned these lessons earlier. In [213], Gordon, Louisell, and Walker conclude:

> We find that a classical description of the input fields and of the amplification process is completely valid provided we take correctly into account the response of the amplifier to the input zero-point fields. This result is valid for inputs of arbitrarily small power.

This is basically correct, apart perhaps from a possible factor of two clarified by Caves [195], arising, as we have seen, from the extra set of amplified vacuum fluctuations which must necessarily be added to the field for amplifiers with gain on non-commuting field observables, in addition to the original amplified vacuum fluctuations, which must be amplified along with the signal, because the amplifier cannot know what part of the input field is to be regarded as signal and what part is considered noise.

If the wiggler magnet is not collapsing the wave function of an electron and causing it to emit a photon, perhaps the amplifier is? For large gain, whether the action of the amplifier is construed to be a measurement or not depends on your favorite interpretation of the quantum mechanics. Some might argue that it can be regarded as an effective measurement because it amplifies the field to levels that "we can lay our classical hands on," i.e., that can be classically observed, measured, and manipulated without significant further disturbance, measurement back-action or degradation of the signal-to-noise ratio. Others might object, pointing out that no agent makes an irreversible, macroscopic record of any information in the radiation fields, at least not until after they are re-applied to the

particles. Either way, it makes little difference to our argument. Even if you construe the amplifier to be performing a measurement of the radiation, it is doing so in a coherent state basis, not a Fock state basis.

What about the environment at large? Even in a state-of-the art beam-line, with high vacuum, vibration control, etc., uncontrollable interactions with the numerous degrees-of-freedom in the environment will inevitably occur, leaving an imprint of information in various quantum entanglements. Without observational access to all of these environmental degrees-of-freedom, the reduced density matrix for the system alone appears impure, and subsequently the system can behave as if the environment has effectively performed a measurement and just not told us the answer.

This is the basic argument behind environmentally-induced decoherence[174, 175, 176, 222], which we have already invoked in arguing for the classical behavior in the electron beam itself. While quantum-mechanical decoherence in the radiation fields will undoubtedly proceed to some degree, it only helps our argument. For simplified but sensible models of "the environment", it is the coherent states, not the Fock states, that survive the "predictability sieve" and emerge as the preferred "pointer-basis" into which the decohered radiation states will appear to be projected[223, 176, 224, 225]. This should not be too surprising. Decoherence favors classical-looking states, and has in fact been suggested as a reason for emergence of a classical world. Conversely, a certain robustness or stability in the face of inevitable sources of decoherence is a key feature of classical-like states. Coherent states, or ignorance mixtures thereof, are according to the pioneering work of Glauber and Sudarshan and countless investigators since then, the closest thing to classical radiation fields allowed by quantum mechanics.

Conversely, If we want to see discrete photons, we have to perform photon-counting experiments, which are usually destructive, and do not leave the remaining field in a number state. If for some reason we really want the "amplifier" to multiply photon number states, rather than scale coherent states, then we must use photon-number-amplifiers (PNA)[220] rather than linear amplifiers, devices which have been proposed but which would be difficult in practice to operate with high fidelity. (Of course, using either unitary PNA or non-unitary photo-detection for OSC would be disastrous, as our naive model predicted). Fock states or other non-classical states with negative Glauber-Sudarshan quasi-distribution functions are just rather delicate and temperamental creatures, difficult to create and maintain, and fragile to the touch of the environment. The bane of many investigations, this fragility of quantum coherence works to our advantage here.

5.13.4 Can the Quantum Mechanical Noise Limits be Achieved?

We have found that quantum mechanics imposes a firm lower bound on the amount of noise that must be added by any linear amplifier in order to maintain consistency with the uncertainty principle and with unitarity, but have not addressed how difficult it may be to approach the ideal.

In practice, there are also also equilibrium and non-equilibrium thermodynamic limitations. The gain medium, as well as the various other optical elements are coupled to thermal baths, so the "initial" state of the radiation field prior to the arrival of the pickup signal is typically not a vacuum, but a thermal blackbody distribution, with an average of

$$\langle n_k \rangle = \frac{1}{e^{\frac{\hbar\omega_k}{k_{\rm B}T}} - 1} \tag{5.180}$$

photons in each mode k of frequency ω_k. At high temperatures, i.e., $k_{\rm B}T \gg \hbar\omega_k$, the average occupation $\langle n_k \rangle \to \frac{k_{\rm B}T}{\hbar\omega_k} \gg 1$, and the amplifier will have an effective noise at input far higher than the zero-point vacuum energy. Conversely, for low temperatures, where $k_{\rm B}T \ll \hbar\omega_k$, the average photon number becomes $\langle n_k \rangle \to e^{-\frac{\hbar\omega_k}{k_{\rm B}T}} \ll \frac{1}{2}$. For $\lambda \sim O(1\ \mu\text{m})$, or equivalently $\hbar\omega \sim O(1\text{ eV})$, and any temperature $T \lesssim O(10^4\text{K})$, the thermal occupation is negligible, and the blackbody state will be virtually indistinguishable from vacuum as input.

Along with large gain bandwidths, this is another tremendous advantage optical-frequency stochastic cooling might enjoy. Merely from purely thermal considerations, it would be extremely difficult to construct and operate amplifiers anywhere close to the quantum limit at radio or microwave frequencies, but this is possible at optical frequencies, thanks to the exponential dependence of the Boltzmann factor. However, with very strong pumping, some active or passive cooling may need to be supplied to ensure quiet and efficient operation.

However, while the surrounding radiation field may start out in thermodynamic equilibrium, the amplifier itself is not, but instead subject to various energetic flows from the pump as well as the input and output signals. In fact a non-equilibrium state is of course necessary to achieve gain. A finite pump strength leads to a finite population inversion which in turn leads not only to smaller gain, but to "excess" spontaneous emission beyond that needed to satisfy the uncertainty principle.

Using semi-classical arguments, it can be shown that the gain for narrow-band propagation along z in a typical inverted medium can be written

$$G = e^{\Gamma z} \gg 1, \tag{5.181}$$

where

$$\Gamma \approx \frac{n_2 \eta_2 \lambda^2}{4 P n_{\rm ir}^2 \tau_0 \Delta\omega}, \tag{5.182}$$

where $\Delta\omega$ is the bandwidth, $n_{\rm ir} \geq 1$ is the index of refraction of the gain medium, λ is the wavelength for the resonant lasing transition between levels 2 and 1, $\eta_{23} \leq 1$ is the quantum efficiency for pumping, τ_0 is the characteristic fluorescence lifetime, or time-scale for spontaneous emission, n_2 is the average occupation density of the upper level 2, and

$$P = \frac{n_2}{n_2 - \frac{g_1}{g_2} n_1} \geq 1 \tag{5.183}$$

is the population inversion parameter, in which n_1 is the average occupation density of the lower level 1 and g_1 and g_2 are degeneracy factors for these levels.

The effective number of noise photons then becomes

$$\bar{n}_{\rm noise} = \tfrac{1}{2} + (P - \tfrac{1}{2}) \left| 1 - \tfrac{1}{G} \right|, \tag{5.184}$$

which approaches the ideal lower bound as $G \to \infty$ and $P \to 1^+$ the latter implying total population inversion. A back-of-the-envelope calculation indicates that required pump power in a multi-stage Ti:sapphire solid state laser amplifier suitable for OSC might correspond to $G \sim O\left(10^4\right) \gg 1$ and $P \sim O\left(\frac{7}{4}\right)$, which corresponds to a noise power only $O(25\%)$ higher than the quantum limit.

Quantum mechanics neither requires nor prevents an actual amplifier from exhibiting some multiplicative noise in addition to the required additive noise, for example due to pump fluctuations, but it will not be of the drastic all-or-nothing character that proved so deleterious to cooling in the Poissonian emission model.

5.13.5 Comparison to the Heifets-Zolotorev Model

Independently, Heifeits and Zolotorev[226] developed a fully quantum 1D model of transit-time optical stochastic cooling.[19] They follow the combined von Neumann density matrix for the particle bunch and the radiation through the various components of the cooling system, including the pickup wiggler (still with classical fields), the bypass for the particles and the amplifier for the radiation, and then the classical fields of the kicker wiggler. While both particles and fields are treated quantum mechanically, a simplified 1D,

[19] M. Zolotorev should be credited with instigating both approaches to the problem. After he suggested this investigation to us, and our answers were not forthcoming with sufficient alacrity, he went in search of a theoretical mercenary with bigger mathematical weapons at his disposal.

single-mode model of the interaction inside the wiggler is used to calculate the initial emission by the particles and the final energy kick, and a simple model of the optical amplifier is employed, based on a completely-inverted population of two-level atoms.

To solve for the evolution, a number of expansions are made in various small parameters, particularly $\frac{N_u}{\gamma_0}\frac{\lambda_c}{\lambda_0} \ll 1$, $\mathscr{N}_{\rm ph} \ll 1$ (the mean number of coherent photons emitted per particle in the pickup), $\frac{N_s}{N_B}$, and $\frac{\hbar\omega_0}{mc^2\delta\gamma}$.

After a minor *tour de force* in the manipulation of special functions, they obtain results which appear to be very similar to those of our hemi-classical theory: namely, assuming the coherent field mode is initially in vacuum, the pickup radiation from a single particle will be described by a Glauber coherent state, with average photon number $\mathscr{N}_{\rm ph} \sim \alpha$. The reduced density matrix of the amplified radiation is that of a chaotic, or thermal-like state, with the equivalent of one photon of amplifier noise at input (half for the original "zero-point motion," and half again from an independent set of atomic fluctuations, because the amplifier provides gain for both non-commuting quadrature components), although combined particle-field density matrix contains the small but vital correlations (what classically we called the coherent component of the signal) needed to cool the particles. The cooling rates are close to the classical results except with the equivalent of $O(\alpha^{-1})$ extra particles in the sample.

Even for a simplified dynamical model, the fully quantum results are impressive, but only seem to confirm what we have suggested, that a hemi-classical formalism (which in the end really just amounts to a classical but stochastic model) is sufficient, and in principle can be more easily generalized to account for finite bandwidth, off-axis, or other neglected effects, which however is beyond the scope of our current investigations.

5.14 Conclusions: Summary and Future Directions

Assuming optimal gain, the rate of any stochastic cooling scheme can be improved by increasing the bandwidth of the pickup-amplifier- kicker system or by decreasing number neighboring particles (which actually cause heating) in a sample length, which varies inversely with the bandwidth. Large bandwidths available in solid-state or certain parametric amplifiers at optical wavelengths are very attractive for stochastic cooling, if various technological challenges can be met. In the transit-time OSC scheme, spontaneous undulator radiation from the pickup wiggler is amplified while particles are given appropriate phase

delays in bypass, then amplified radiation acts back on particles within the kicker wiggler to reduce momentum spread and/or betatron amplitude.

The pickup self-radiation of each particle (responsible for cooling) is very small – on average, only $O(\alpha)$ photons per particle in coherent mode, suggesting quantum effects may be important In a naive picture, akin to Sands treatment of synchrotron radiation damping, where particles emit whole photons in discrete Poissonian quantum jumps, it would seem particles rarely experience cooling self-radiation but almost always feel large heating cross-radiation, leading to drastically slower cooling or over-correction when the cooling photon finally appears.

A more careful analysis has indicates the feared quantum catastrophe is a red herring: particles radiate into Glauber coherent states, or statistical mixtures thereof, and amplifiers amplify coherent states, not Fock (photon number) states. The cooling signal is really present on every pass. Particles emit radiation in a quantum mechanical state in which photon number and field energy are not sharply defined, but in which the expectation values of amplitude and phase are given by the classical values, rather than in some sort of quasi-classical Poisson process where whole photons are emitted in a series of discrete "quantum jumps." A low-noise amplifier will operate quantum mechanically, transforming the quantum state of the fields according to Hamiltonian dynamics, and does not act to first projectively measure then multiply photon number. After amplification, the amplified radiation in the kicker is essentially classical, and particles respond linearly to fields rather than via discrete photon absorption. The amplified phase signal from each particle is present on each pass, albeit partially obscured by some amount of noise due to amplified vacuum fluctuations of the original radiation field and certain internal amplifier modes. Therefore, quantum mechanical effects do not destroy cooling, but just introduce additional (additive) noise, approximately equivalent (prior to amplification) to about $O(\alpha^{-1})$ additional particles in a sample, setting an ultimate upper limit on cooling rates and on the gains in efficiency that can be expected from further beam dilution

If sufficient care is given to the quantum mechanical aspects of the radiation and amplification, which in the end just amounts to adding at least the minimum amount of amplifier noise that can physically be traced back to spontaneous emission in the gain medium, then optical stochastic cooling rates can be calculated using an essentially classical treatment. For ultra-fast OSC, where the correction consists of a relatively small number of relatively-large kicks compared to traditional RF-based schemes, temporal smoothing or averaging is not really justified, but while individual particle behavior is subject to high uncertainty,

properties the beam as a whole (mean per-particle energy and energy spread) can be reliably estimated because of averaging over many particles in the beam, rather than averaging over many time steps.

Using the cooling rates as derived above for the simplified model of a cooling section, a very preliminary calculation indicates that micro-second cooling for muons is possible under somewhat optimistic expectations for the precision of the beam optics and an optimistic extrapolation for the power available in low-noise, high-bandwidth amplifiers. Furthermore, the bunch charge $N_b \sim O(10^{10})$ assumed in this study is smaller by an order of magnitude or two than what is discussed in current proposals for a muon collider, although because of the additional stochastic cooling, the final luminosity can be just as high.

Now that concerns about potentially devastating quantum mechanical have been assuaged, the possibility of ultra-fast cooling rates appears sufficiently promising to justify more careful investigation. Attention can turn to of the more prosaic classical effects that might limit performance, in order to include important physics so far neglected.

A more realistic model of the classical wiggler radiation should be used, with the effects of off-axis injection, diffraction included. A careful classical calculation will be somewhat complicated by the unusual parameter regime of OSC wigglers, involving high magnetic field strengths and a relatively small number of relatively long-wavelength poles. The effects on particle orbits of angular deviations and oscillating components of the longitudinal motion in a planar wiggler are not necessarily small, but have so far been neglected. Most critically, sensitivity to the errors in the bypass lattice optics must be investigated, the effects of only partial mixing between passes or unwanted mixing within passes addressed, the Guoy phase included, and the effects of non-uniform gain and optical dispersion in the amplifier realistically treated.

Many technological challenges need to be addressed and assessed: the beam stretching needed to reversibly increase bunch length and decrease energy spread to manageable levels before cooling must be accomplished on microsecond time-scales; as must he re-compression after cooling, which is even more difficult since there will by design be less energy spread to be exploited in the high-dispersion lattice. The bypass and kicker optics must be carefully engineered, sufficiently adjustable to calibrate particle orbits within a fraction of a micron tolerance, and presumably be dynamically controllable through active feedback. The optical amplifiers must be roust, stable, single-pass (non-regenerative), highly linear even at high power, variable gain, low-noise, and probably actively-cooled. Mixing (effective shifting longitudinal particle positions on the scale a radiation wavelength or more) must be very

thorough between passes the cooling sections, but negligible between the pickup and cooler of any given cooling section.

Now, at least, it is confirmed that such cooling appears in principle completely consistent with quantum mechanical features of radiation and amplification.

Acknowledgements

This problem was suggested to us by M. Zolotorev, who also offered many useful suggestions. We learned many interesting things about quantum optics while figuring out how to grade an ostensibly simple but profound homework problem assigned by Professor Eugene Commins for his graduate course in Statistical Mechanics.

Chapter 6

Electromagnetically-Induced Transparency in Magnetized Plasmas: Quantum Treatments and Atomic Analogies

Light seeking light doth light of light beguile:
So, ere you find where light in darkness lies...

WILLIAM SHAKESPEARE
Love's Labour's Lost, Act I, Scene 1

It makes all the difference whether one sees darkness
through the light, or brightness through the shadows.

DAVID LINDSAY
A Voyage to Arcturus (1920)

6.1 Introduction

Electromagnetically-Induced Transparency (EIT)[227, 228] is a technique, or phenomenon, by which an otherwise opaque medium is made transparent to certain electromagnetic (EM) radiation via the application of certain other EM fields – whereby light can modulate or manipulate the propagation of other light. Sometimes it is said that EIT seeks to eliminate the effects of the medium on a propagating beam of EM radiation, but typically

associated with the transparency is a significant modification of the refractive, or dispersive, properties compared to what would be observed in vacuum, leading to "slow light" and other effects. In recent years, EIT has been the subject of significant theoretical and experimental attention [227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242], both out of fundamental interest and for its many possible applications, involving the possible suppression of self-focusing, enhancement of other nonlinear optical effects which would otherwise be masked by absorption, ultra-sensitive magnetometry, precision low-energy tests of the standard model, lasing "without inversion;" optical communications and switching or new optoelectronic devices, quantum information storage or processing, the creation of a so-called phaseonium, or large extended population of coherently-phased atoms, or even quantum entanglement exchange between EM fields and matter[227, 243, 231, 244, 245, 246, 247, 236, 240, 248, 249, 250]. (For reviews of the state of the field as of about a decade ago, see [251] or [252]. For a more contemporary summary, see [253] or [254].)

EIT for probe lasers propagating in atomic vapors is now reasonably well-understood as a consequence of, and often held as a striking example of, quantum-mechanical superposition. In a Feynman sum-over histories, or transition amplitude, picture, transparency results from destructive interference between several excitation pathways which connect the ground and excited states in the otherwise opaque atomic medium. From a Schrödinger, or quantum state, picture, transparency may be described by the coherent mixing and re-splitting of the two atomic levels coupled by the pump field, as in textbook perturbation theory, in such a way that a probe field applied at the original resonant transition frequency is de-tuned above one pump-mixed level and below the other, so that the probability amplitudes for probe absorption cancel. From an interaction-picture perspective, where the pump and probe carrier oscillations are transformed away but field amplitudes may still be slowly modulated, EIT may be understood in terms of a "dressing" of the bare quantum levels of the medium by the couplings to the pump and probe fields, leading to a dressed "dark" eigenstate of the interaction Hamiltonian which has de-coupled from the probe and which does not contain any admixture of the excited state that would be a signature of probe absorption, and which furthermore can be adiabatically connected to the ground state and remain exclusively populated if the pump is established before the probe and certain slowness conditions on the field envelopes are maintained. From a Heisenberg, or quantum operator, picture, an analogous dressing emerges for Hilbert-space operators rather than state vectors, wherein a so-called Dark-State Polariton (DSP) mode operator

creates quanta of excitations combining field and material degrees-of-freedom (DOFs) in just such a manner as to allow the suitably "disguised" signal to "sneak" into and through the medium. Different descriptions may be preferable as a matter of personal taste or suitability to different applications or initial conditions or parameter regimes, but all are ultimately equivalent, and each relies essentially on quantum interference.

However, more recently [255, 256, 257, 258, 259], a similar phenomenon in a classical plasma, with numerous and obvious analogies to EIT in atomic vapors, has been and proposed and analyzed, involving electromagnetically-induced transparency of a microwave signal in an axially-magnetized electron plasma at a frequency which in the absence of the pump field would suffer resonant cyclotron absorption. Yet this magnetized-plasma EIT effect has ostensibly been adequately described and explained within a fully classical description of both matter and fields, where, from the perspective of plasma electrons oscillating in a longitudinal Langmuir wave ponderomotively excited by the beating between pump and probe, the effects of a side-band of the pump field can cancel the probe-induced currents which would otherwise lead to resonant absorption. This naturally suggests interesting questions of principle concerning whether and how the plasma and atomic versions of EIT are related, and to what extent quantum-mechanical interference is at bottom responsible or necessary for EIT.

Here we seek to answer these questions, by formulating analogous collective-mode quantum-mechanical formalisms for EIT in the cases of a cold magnetized plasma or atomic vapor, and whose classical limit in the plasma case is manifest.

6.2 EIT: Types and Comparisons

Electromagnetically-Induced Transparency (EIT) may be defined as the reduction (or ideally, elimination) of resonant absorption, or enhancement of transmission, of a probe (also called signal) EM field in an otherwise opaque medium, via the prior application of a suitably intense, suitably-tuned, suitably-enveloped pump (or drive, or control) EM field.

6.2.1 Atomic Systems

In atomic vapors, EIT is closely associated with the widely-studied phenomenon of "slow light" – a reduced effective group velocity – and procedures which go by the name of Adiabatic Passage or Coherent Population Transfer, techniques belonging to the growing

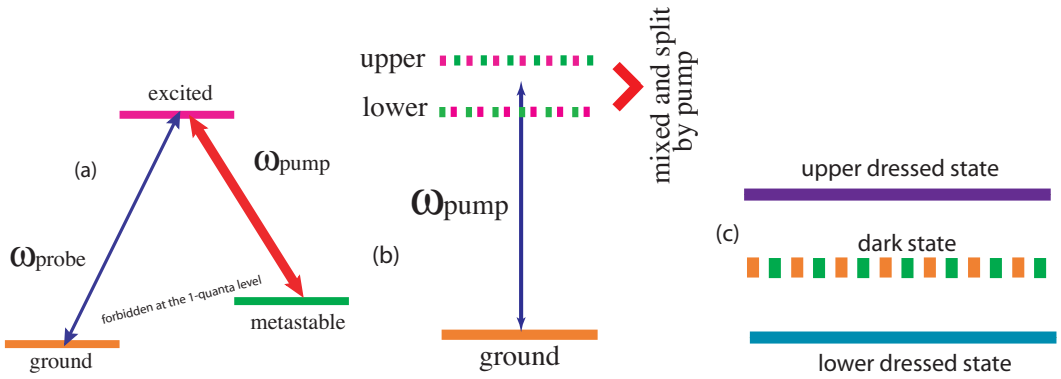

Figure 6.1. Schematic Λ-level diagram for relevant electronic states of an individual atom subject to EIT. In (a) the states are presented in the unperturbed bare basis, and the couplings to the probe and pump fields are indicated for the fully resonant case. In (b), the states coupled by the pump field are shown schematically in the pump-mixed basis, which diagonalizes only that part of the Hamiltonian involving these states and their pump-dependent coupling. The probe can experience vanishing absorption through cancellation between the probability amplitudes for transitions into the upper and lower pump-mixed states. In (c), the dressed states diagonalizing the complete interaction Hamiltonian are sketched. Each dressed state evolves independently. The dark state contains little or no admixture of the excited state that would indicate absorption of probe photons, yet is adiabatically accessible from the ground state under suitable pump and probe preparation.

field of quantum control theory in which atoms, molecules, or other quantum systems are coherently driven or manipulated. Although more elaborate quantum state configurations are also possible, as originally formulated ana analyzed, EIT in atomic systems occurs in gases of atoms in which the relevant electronic energy manifold can be modeled by certain three-level ("Lambda") structure, shown schematically in Fig. 6.1, with an intermediate metastable state lying in between an excited state and a "ground" state.

This lowest state may or may not be the actual ground state of the full system, but its lifetime, if finite (due to radiative or other losses not included in three-state dynamics) is assumed to be sufficiently long compared to other time-scales in the problem, as is the natural lifetime of the metastable state. A nearly-resonant probe EM field couples the ground and excited states, so would normally be absorbed if incident on a sufficiently dense gas of these atoms. A pump field, downshifted in frequency with respect to the probe, can also couple the excited and metastable states, but single-photon transitions between the metastable and ground state are assumed forbidden by selection rules. Actually, in order to induce EIT, the pump may be de-tuned from the one-photon transition between the metastable and excited state, as long as the pump and probe frequencies differ by the Bohr frequency associated with the forbidden transition between the ground and metastable

states, thereby satisfying the two-photon resonance condition for Raman transitions between the ground and metastable states that need involve the excited state only virtually. However, EIT generally works most efficiently when the pump frequency also satisfies the single-photon resonance.

Both the metastable state and excited state are assumed initially unpopulated, and by design the excited state should remain largely unpopulated, since it can be populated only at the expense of probe photons. Therefore the EIT effect can be immune, or at least resistant, to higher rates of spontaneous emission from the excited state into omitted states than from the ground or metastable states. In order to establish EIT, the pump must be turned on sufficiently slowly prior to the arrival of the probe, which must also be modulated sufficiently slowly to maintain adiabatic conditions and avoid absorption. Because the pump only appreciably connects the metastable and excited states states, both of which are initially unpopulated and one of which remains unpopulated, naively it would seem not to do anything, and therefore its required presence before the probe has been described as "counterintuitive." But the conditions for the destructive interference of the probability amplitudes for absorption must be established before that cancellation can in fact occur. Or, in different language, the dark state must be dressed before the system can be nudged into it, so this preparation is not really counterintuitive upon a bit more thought.

In summary, note that the transparency is attributed to quantum interference in the dipole excitation of the *internal* states of each atom.[1] A *downshifted* coupling field can make the atomic vapor transparent to a probe field which is otherwise *resonantly absorbed*.

The original phenomenon described as EIT in plasmas involves interactions between a propagating laser field and an an unmagnetized plasma. An *upshifted* pump EM wave can make the plasma transparent to a probe EM wave near the *cutoff* at the plasma frequency. It is attributed to classical interference and relativistic effects in the *collective* excitation of the medium.

In the more recently-discovered EIT in magnetized plasmas, a *downshifted* pump field can make a plasma with an axial magnetic field transparent to a right-handed, circularly-

[1] Experiments and analysis have primarily focused on cooled but non-degenerate (non-condensed) atomic vapors, where atoms may be treated as non-interacting (but not necessarily uncorrelated, due to possible Dicke superradiance effects), but applications of EIT ideas to Bose-Einstein Condensates (BECs) have now also been suggested[260, 261, 262, 263], where inter-atomic interactions and center-of mass dynamics may be important, while EIT in crystals[264, 265] could work with optical control fields acting on electronic levels as in the gaseous phase, but might also involve phonons, spin waves, or other collective modes. EIT in photonic bandgap materials[266], superconducting quantum circuits[267], and other exotic media have also been proposed.

polarized microwave probe near the cyclotron *resonance*, which would otherwise be strongly absorbed. It has been attributed to classical interference in both the *collective* and *single-particle* excitation of the plasma.

While salient similarities and differences are apparent between all three phenomena, because the unmagnetized plasma EIT effect involves an upshifted rather than downshifted pump frequency and enhancement of transmission near a cutoff rather than a resonance, we suspect that EIT in magnetized plasma will prove a much closer analogy to atomic EIT.

6.2.2 Classical EIT in Unmagnetized Plasmas: Cutoff Diminution and Periodic Tunneling

The original phenomenon described as classical EIT in plasmas[268, 269, 270, 271] involves interactions between a propagating laser field and an unmagnetized plasma. Just to avoid confusion with the magnetized-plasma version upon which we will focus, let us briefly summarize (and dismiss) this phenomenon. An intense pump wave at the frequency $\omega_2 = \omega_1 + \omega_p$, upshifted by the electron plasma frequency ω_p (defined below), allows propagation of a probe wave at the frequency $\omega_1 \lesssim \omega_p$, which would otherwise be reflected as it oscillates at a frequency somewhat below the ostensible cutoff for transverse EM modes in the plasma. This transparency is attributed to several classical effects: first, a classical interference, whereby the pump beats with the probe to ponderomotively drive a plasma wave; and the plasma wave then beats with the pump to partially cancel currents associated with the probe which would otherwise tend to cancel the transmission and contribute to reflection of the probe.[2] So far, this sounds somewhat reminiscent of atomic EIT, but a more direct mechanism is at work, involving a shift in the effective cutoff, primarily from an increase in the effective relativistic mass due to the transverse quiver motion of the plasma electrons in the pump field, and possibly also due to the presence of the induced plasma wave itself.

Specifically, in the presence of the pump field and possibly a ponderomotively-driven plasma wave (with amplitude below the cold wave-breaking limit), using a simple one-dimensional cold fluid theory describing electrons of rest mass m_e, charge $q_e = -e < 0$, unperturbed number density n_0, and temperature $T_e = 0$, and a neutralizing background of massive immobile ions of density $n_i = n_0$, charge $q_i = +e$, and rest mass $m_i \gg m_e$, the effective local dispersion relation for an EM probe wave of frequency ω and wavenumber k

[2] To the alert reader there might seem to be a bit of a chicken-or-egg puzzle here, in which it is unclear how the probe pulse can initially penetrate the plasma to beat with the probe pulse and set up the plasma wave, of which more will be said later.

in a (reasonably cold) unmagnetized plasma then becomes

$$\omega^2 \approx \omega_p^2 \frac{1+\delta n/n_0}{\gamma^2} + c^2 k^2, \tag{6.1}$$

where: c is the speed of light in vacuum; in Gaussian units, the linear electron plasma frequency ω_p is given by

$$\omega_p = \sqrt{\frac{4\pi n_0 e^2}{m_e}}; \tag{6.2}$$

$\delta n = n_e - n_0$ is the local density perturbation associated with the plasma wave; and γ is the relativistic kinematic factor, which in one-dimensional geometry can be written as

$$\gamma = \frac{\sqrt{1+a_{\text{pump}}^2}}{\sqrt{1-\beta_z^2}} \tag{6.3}$$

in which $a_{\text{pump}} = \frac{e}{m_e c^2} \|\boldsymbol{A}_{\text{pump}}\|$ is the RMS magnitude of the normalized Coulomb-gauge vector potential for the pump, and determines the rapid quiver momentum of the electrons by way of transverse canonical momentum conservation; and finally $\beta_z = v_z/c$ is the scaled longitudinal electron fluid velocity associated with the plasma wave. As γ increases due to either transverse quiver or (what is typically a much smaller effect) longitudinal motion in the plasma wave, the *effective* cutoff frequency downshifts.

Furthermore, the density perturbation associated with the plasma wave can then periodically modulate the cutoff frequency, leading to a "dynamic etalon" effect in which over some range of plasma, pump, and probe parameters, the effective cutoff frequency is such that the probe wave can propagate above the effective cutoff through the rarefied regions and tunnel through the denser regions, resulting in net transmission.

In summary, despite a few similarities, this effect differs form atomic EIT in significant ways. Here an *upshifted* pump EM wave can make the medium transparent to a probe EM wave just below the *cutoff* frequency. It is attributed to classical interference and relativistic effects in the *collective* excitation of the medium. Because the pump wave is upshifted relative to the probe by wave by the plasma frequency, in practice it might be difficult to distinguish actual transmission of the probe from mere Raman scattering of the pump.[3]

6.2.3 EIT in a Magnetized Plasma: Resonance Suppression

While the induced transparency in unmagnetized plasmas may be interesting, on closer inspection it does not, in our opinion, actually have a great deal in common with the case

[3] In fact, just such Raman scattering might initially induce the plasma wave and resolve our chicken-or-egg dilemma.

of atomic EIT, at least compared to the more-recently suggested phenomena of EIT in a magnetized plasma[255, 256, 257, 258], to which we now turn.

Longitudinally-Propagating Mode Structure in Magnetized Plasma

In the presence of a static, spatially-homogeneous, magnetic field of an ideal solenoid oriented longitudinally (taken for definiteness to be the $\hat{\boldsymbol{z}}$ direction), i.e., $\boldsymbol{B} = B_0\hat{\boldsymbol{z}}$ for some constant field strength $B_0 \geq 0$, and ignoring relativistic effects, the cold fluid theory mentioned above predicts the well-known linear dispersion relation[272] for a weak EM wave of frequency ω and wavevector $\boldsymbol{k} = \hat{\boldsymbol{z}}$ traveling parallel (or anti-parallel) to the magnetic field:

$$\eta_{pm}^2 = \frac{c^2k^2}{\omega^2} = 1 - \frac{\omega_p^2/\omega^2}{1 \mp \Omega_c/\omega}, \tag{6.4}$$

where η_{pm} is the index of refraction for waves of circular polarization (the conventional notation involving n being avoided in order to minimize confusion with various densities, indices, and other uses for this over-taxed letter), and

$$\Omega_c = \left|\frac{q_eB_0}{m_ec}\right| \geq 0 \tag{6.5}$$

is defined as the positive linear electron cyclotron frequency (also called the gyrofrequency, or Larmor frequency). For waves traveling with $k > 0$ in the $+\hat{\boldsymbol{z}}$ direction (and still supposing $B_0 > 0$ without loss of generality), the upper sign corresponds to the right-handed, or so-called R-wave, where the electric field of the wave rotates in the same sense as the negatively-charged gyrating electrons, leading to a resonance exactly at $\omega = \Omega_c$. Two distinct branches with this polarization are supported. Between this resonance at Ω_c for what turns out to be the lower branch of the R-wave (sometimes called the whistler mode or the electron-cyclotron wave) and the cutoff of the upper (or optical) R-wave branch located at

$$\omega_r = \tfrac{1}{2}\Omega_c + \sqrt{\omega_p^2 + \tfrac{1}{4}\Omega_c^2} \geq \Omega_c \tag{6.6}$$

lies a stop band, whereby no traveling R-waves within the frequency interval

$$\Omega_c \leq \omega \leq \omega_r \equiv \Omega_c + \Delta_c \tag{6.7}$$

can be transmitted through the plasma, but will rather be absorbed and/or reflected. In contrast, the counter-rotating, left-handed, or L-wave consists of a single optical branch with a cutoff at

$$\omega_\ell = -\tfrac{1}{2}\Omega_c + \sqrt{\omega_p^2 + \tfrac{1}{4}\Omega_c^2} \leq \omega_{\rm R} \tag{6.8}$$

but has no resonances. While $\omega_\ell \leq \omega_p \leq \omega_r \leq \omega_p + \Omega_c$ in all cases, ω_ℓ can lie above or below Ω_c, so the L-wave can propagate through at least part and possibly all of the R-wave's stop band. The dispersion relations for all three branches are shown below in Fig. 6.5 (along with other approximate forms), clearly showing the stop-band.

Classical Model of EIT Phenomenon

However, if an intense pump at frequency $\omega_2 \approx \omega_1 - \omega_p$ is slowly turned on before the arrival of the probe, it can actually make the medium transparent to a probe of frequency $\omega_1 \gtrsim Omega_c$ in what would otherwise be the stop-band. This is a beautiful example of classical EIT in magnetized plasma, and has a completely classical explanation within the simplified $1D$, cold fluid model (See [255, 257] for more details.) The beating between a weak, R-polarized probe, at a near-resonant frequency $\omega_1 \gtrsim \Omega_c$ and an intense pump at downshifted frequency $\omega_2 < \omega_1$ will tend to ponderomotively excite a plasma wave as long as $\omega_1 - \omega_2 \approx \omega_p$. Because $\omega_r \leq \omega_p + \Omega_c$, the pump frequency $\omega_2 \approx \omega_1 - \omega_p$ will lie on the lower (cyclotron) branch of the R-wave if the probe frequency ω_1 is within the stop band.

In the absence of a suitable pump field, particle gyration due to the axial magnetic field is in resonance or nearly in resonance with the probe, leading to cyclotron absorption, but motion in the presence of the pump is more interesting. For given values of the longitudinal magnetic field, transverse electromagnetic (EM) fields, and longitudinal electrostatic (i.e., plasma wave, or Langmuir wave) field, the longitudinal degrees-of-freedom (DOF) of each electron are coupled both by the Lorentz force and by the longitudinal variation of the transverse EM fields, raising the possibility of cancellation under the right circumstances . That is, at the time t, and assuming small-amplitude longitudinal motion, the jth electron experiences a transverse force given approximately by

$$\boldsymbol{F}_{j\perp}(t) \approx -\tfrac{e}{c}\dot{\xi}_j \hat{\boldsymbol{z}} \times \boldsymbol{B}(\bar{z}_j, t) - e\left[\boldsymbol{E}(\bar{z}_j, t) + \xi_j \tfrac{\partial}{\partial \bar{z}_j}\boldsymbol{E}(\bar{z}_j, t)\right], \tag{6.9}$$

where $\bar{z}_j$ is the initial equilibrium position of the electron, $\xi_j = z_j(t) - \bar{z}_j$ is the Lagrangian displacement of the electron from its initial position, and $\dot{\xi}_j = \frac{d}{dt}\xi_j$ is the longitudinal velocity, and so the transverse force contains two distinct terms proportional at linear order to the longitudinal response.

Due to their longitudinal motion in the plasma wave and/or ponderomotive potential of the EM waves, electrons gyrating transversely and oscillating longitudinally will (from their point-of-view) experience the original waves as well as doppler-shifted side-bands of the

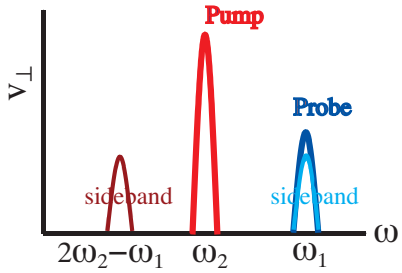

Figure 6.2. Schematic representation of the frequency spectrum of the transverse electron velocity response in presence of the probe and pump. Oscillating longitudinally in the ponderomotive potential and/or Langmuir wave, plasma electrons see Doppler-shifted sidebands of the pump which can cancel the effects of the probe acting alone.

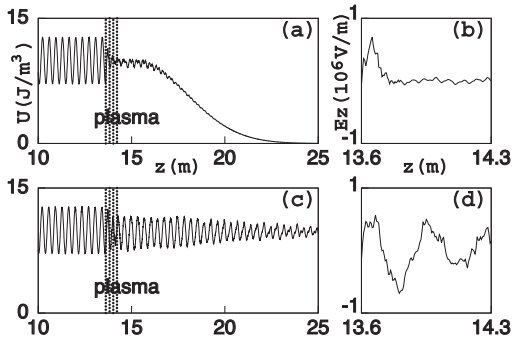

Figure 6.3. Numerical simulations using XOOPIC of induced transparency at the cyclotron resonance, showing (a) EM wave energy and (b) longitudinal electric field at an early time, when the probe pulse has just begun to penetrate the plasma, and EM wave energy (c) and longitudinal electric field (d) at a later time, after the leading edge has passed through the plasma.

pump, shown schematically in Figure 6.2. For suitable initial conditions and adiabatically-rising pulse envelopes, the transverse velocity response of each electron due to the probe and due to the upper sideband of the pump can destructively interfere, suppressing what would be the near-resonant effect of the probe on the electron gyration.

Numerical Simulations

Some typical results (in MKS units) of 1D particle-In-Cell (PIC) simulations using the XOOPIC code developed are shown in Figure 6.3 - 6.4.

These and other simulations have confirmed that the pump induces transmission in what was otherwise a stop-band, but at a slow effective group velocity, in support of the

theoretical predictions of classical EIT and classical "slow light." Appreciable transmission at the probe frequency requires the prior presence of a suitably down-shifted pump. The simulations also clearly indicate that a plasma wave must be established in a region of the plasma before the probe can propagate into that region as a traveling wave, and resolve the minor mystery as to how the plasma wave is established if, after all, the near-resonant probe wave should be absorbed or reflected before it can beat with the pump. Initially, in order to produce transparency locally, it sufficient that the beating between pump and probe occur in *time*, so the leading edge of an incoming probe can still tunnel a short distance into the remaining non-transparent part of the plasma as an evanescent wave with exponential decay in space and damped oscillations in time, then beat with the pump to generate a local plasma wave, which then induces transparency and allows trailing portions of the probe pulse to propagate as a traveling wave up to that point, while the leading edge burrows further into the plasma.

6.2.4 Preliminary Comparison of Atomic and Magnetized-Plasma EIT

At realistic magnetic field strengths and plasma densities, this EIT effect is possible in magnetized plasmas for pump and probe fields in the microwave region. So although it is not relevant for laser-plasma interactions, this effect might find possible applications in ion acceleration, "optical" storage and information processing, plasma-based switching or routing, and plasma-heating for fusion or other studies.

Plasma-based EIT is also of fundamental theoretical interest, because of its many parallels with atomic EIT, but nevertheless would seem to have a completely classical description and explanation. A preliminary comparison shows striking similarities; in both:

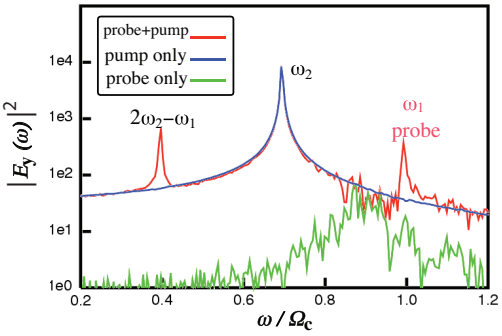

Figure 6.4. FFT power spectrum of numerically simulated transverse EM fields, plotted on a logarithmic scale. The enhancement of the probe transmission in the presence of the pump is apparent.

- absorption near a resonance is eliminated or reduced, or at least split and moved away from the bare resonance;
- the frequency of the required pump field is de-tuned downward from the probe frequency by a frequency corresponding (approximately) to a Raman (non-allowed) transition;
- require adiabatic pulse envelopes and the so-called "counterintuitive turn-on" of the pump prior to the probe; and
- transmission exemplifies "slow light," involving a reduced effective group velocity, which increases with the coupling strength.

But differences are obvious as well. In the atomic case,

- transparency relies on quantum mechanical interference of excitation pathways; and
- this excitation involves dielectric (dipolar) response of individual atoms, with quantized energy levels, selections rules, etc., in a manifestly quantum treatment;

while in the plasma case,

- the transparency can be described by classical physics, and relies on classical interference, namely the cancellation of additive forces; and
- the excitation involves the collective response of (unbound) classical electrons.

Such comparison and contrast raises the questions as to whether these two EIT mechanisms are truly analogous, and, if so, how and to what degree, and what are the implications for issues of classical/quantum correspondence in these systems?

6.3 Simplified Quantum Description of Plasma EIT

Such questions motivate the current work. In order to make a proper comparison, we need to find a common formalism capable of describing both types of systems. Because quantum mechanics can mimic features of classical physics in the correspondence-principle limit but not conversely, this means that the plasma-based EIT must be modeled (mostly) quantum mechanically – but in a manner in which the classical limit can be obtained easily.

Because the plasma EIT is naturally described in terms of a continuum fluid model where at least some of the excitations involve collective degrees-of-freedom of many particles, the atomic case should be re-formulated from this collective point-of-view as well.

We stress that this quantum description of plasma-based EIT is here developed out of theoretical rather than practical interest, in order to clarify the fundamental physics and effect clear qualitative comparisons, and certainly not in anticipation of any actual deficiencies in the classical description of plasma EIT in any practically accessible parameter regime.

Because we are more concerned with making an "in-principle" comparison and not reproducing or extending quantitative predictions that can be derived classically, it is adequate to focus on the essential features of the phenomenon, and make use of a highly-idealized model in which analytic progress can be made without getting too encumbered with complications and calculations. In particular, we will adopt a conservative Hamiltonian formalism, neglecting any form of dissipation such as collisional or cyclotron damping of the transverse waves and collisional or Landau damping of longitudinal waves. After the model is developed and equations of motion derived , damping could be added phenomenologically, if desired. The plasma electrons are described as a cold, homogeneous, fluid which is assumed quiescent initially (before the arrival of the pump or probe). By this we mean that the plasma is cold and quiet enough to neglect thermal and hydrodynamic effects, but warm enough to neglect collisional damping and Fermi degeneracy effects. Because the ions are much more massive than the electrons, and the characteristic time-scales for motion are correspondingly larger, we suppose, as we did above, that the ions remain motionless and merely provide a stationary, neutralizing background. For simplicity, non-relativistic particle dynamics are assumed throughout, although relativistic detuning of plasma, cyclotron, or EM waves is probably not negligible in the actual dynamics. These effects could be reintroduced perturbatively, at the expense of more complicated nonlinearities that would obscure the essential similarities between the atomic and plasma cases. We also consider one-dimensional geometry, in which all field and fluid quantities vary only in time t and longitudinal position z. Two-dimensional or three-dimensional generalizations may be possible, if quite cumbersome. We will see that the cold one-dimensional fluid theory is particularly convenient because the electric field remains a linear function of the Lagrangian fluid displacement, and the classical dynamics are derivable from a simple Hamiltonian, which can then be directly quantized using the Dirac-Weyl canonical prescription. All collective degrees-of-freedom will mostly be decomposed into a discrete set of modes inside a finite quantization volume

$\mathcal{V}$, although it will also sometimes prove convenient to consider the $\mathcal{V} \to \infty$ continuum limit in the usual way, as well as a slowly-varying envelope approximation (SVEA) consisting of narrow-but-finite bandwidths around some carrier oscillations. After verifying accuracy, we will mostly work within the Rotating-Wave-Approximation (RWA), where only action-conserving terms are retained,[4] and off-resonant terms are dropped. The non-RWA 2-wave (i.e., bilinear terms) could be retained instead if desired, but keeping 3-wave non-RWA terms would make the model intractable, and hence impractical. Fortunately, the RWA will be reasonably accurate for modes in the primary region of interest, near the resonant frequency Ω_c. Further specialized assumptions will be introduced as used in order to make further analytic progress.

At the end of our investigation we make a comparison of EIT informed by the common Hamiltonian description, and will return to these simplifications and limitations, assess their impact on our conclusions, and discuss the potential for relaxing certain assumptions in more complete models.

6.3.1 Coupled Plasma-EM Field Hamiltonian

Our strategy will consist in deriving, within the chosen geometry, and with suitable approximations, a reduced non-relativistic Hamiltonian description of the coupled EM field and material (plasma) DOFs. One may translate between classical and quantum language by means of the Dirac-Weyl canonical prescription, replacing Poisson brackets with commutators and c-number classical phase space functions with q-number observables, with careful attention to ordering. Throughout our derivation, it will also prove very convenient to be able to shift between discrete representations for the particles and/or fields and various continuum limits, as well as between real-space (or position space) and Fourier space (or wavenumber, reciprocal, or momentum space) representations of either the discrete or continuous variety. We point out that, although we will eventually employ what is essentially a second-quantized formalism, we do not ever need to use the Dirac equation for the plasma electrons, in part because we are concerned with low-energy, non-relativistic motion, and more fundamentally because we really only quantize collective Bosonic DOF carried by the continuum non-degenerate electron fluid, not the particle degrees-of-freedom of each individual electron. Unless otherwise specified, dynamics will be described in the Heisenberg picture, where the quantum state remains constant and the time dependence resides implicitly in the operators (but said time dependence is generally omitted in the notation

[4] For some assessments of the accuracy of the RWA in similar settings, see, for example, [273, 274].

unless the operator also has explicit time dependence not generated by the Hamiltonian), but we will also work in an interaction picture where part of the time dependence is explicitly transformed away, and occasionally reference the Schrödinger picture, where the states take center stage and evolve instead of the operators.

Preliminaries: Formalisms and Fourier Transforms, Conventions and Other Considerations

Because we will frequently switch between different mathematical representations for the physical observables, we spell out these transformations in some detail for future reference. We distinguish representations where longitudinal positions are taken to be discrete (lattice model) or continuous, the longitudinal wavenumbers are discrete (reciprocal lattice) or continuous, and other electron observables are discrete (particle-like) or continuous (fluid-like). To avoid doubling of an already expansive notation, we sometimes use the same symbol for q-number (operator-valued) and corresponding c-number (complex-valued) quantities, i.e., for quantum observables and their eigenvalues of expectation values. The distinction will (hopefully) be pointed out where it is not obvious from context.

We work throughout in one-dimensional geometry, where all physical field and particle observables *vary* only with time t and longitudinal position z, although of course some vector quantities such as velocity or momentum may have non-vanishing components in transverse directions.

We begin with a fully discrete representation, where the transverse EM field may be assumed to be decomposable into countable set of modes, and the plasma particles are taken as discrete electrons moving in a dynamically-inert neutralizing ionic background. Actually, we imagine an even "more" discrete description, where *equilibrium* particle positions are assumed to fall on the points of a lattice, which is of course physically unrealistic for non-crystalline media but offers a convenient mathematical starting point. Various limits then can be taken as convenient to recover continuous dependence on position and/or momentum, or to course-grain and smooth observables to obtain a fluid description.

We begin by formally restricting attention to particles and fields confined within some long quantization length $Z_0 \leq z \leq Z_0 + L$, and impose periodic (Born von-Karman) boundary conditions, which lead to simpler mathematical results than the more physically realistic hard-wall boundary conditions, although the differences will vanish in the $L \to \infty$ limit. The simplest way to ensure that all scalings and coupling constants remain physically mean-

ingful and dimensionally correct is to first also assume transverse confinement within some cylinder of radius R and cross-sectional area $\mathcal{A} = \pi R^2$ throughout the quantization volume $\mathcal{V} = \mathcal{A}L$, and then at the very end of any calculation take the $R \to \infty$ limit. Actually, given that we will always have this limit in mind, we can consistently take R to be large but finite for the purposes of normalization, yet implicitly assume $R \to \infty$ from the start by neglecting transverse variation in, and longitudinal components of, the pump and probe fields (arising from hollow waveguide boundary conditions), transverse components of space-charge fields (arising from finite plasma cross sections), and any transverse variation of particle response due to transverse structure in the fields.

We define a set of $N_z \gg 1$ discrete lattice positions over the quantization length:

$$\bar{z}_j = z_0 + (j-1)\Delta z, \quad \Delta z = \tfrac{L}{N_z}, \quad j = 1, \ldots, N_z; \quad \bar{z}_{j+N_z} = \bar{z}_j. \tag{6.10a}$$

In Fourier space, the $N_k \gg 1$ discrete wavenumbers are taken to be

$$k = k_n = n\Delta k, \quad \Delta k = \tfrac{2\pi}{N_k \Delta z} = \tfrac{N_z}{N_k}\tfrac{2\pi}{L}, \quad n = -\tfrac{N_k}{2} + 1, \ldots, -1, 0, +1, \ldots, +\tfrac{N_k}{2}, \tag{6.11}$$

where for simplicity we assume N_k is chosen to be an even integer. Of course, these wavenumbers will constitute a reciprocal lattice conjugate to the position lattice if and only if $N_k = N_z$, but we can allow for more generality.

As for the plasma particles, we assume the position the ith electron is given by

$$\boldsymbol{x}_i = \boldsymbol{x}_i(t) = \boldsymbol{x}_\perp + \hat{\boldsymbol{z}} z_i = \hat{\boldsymbol{x}} x_i + \hat{\boldsymbol{y}} y_i + \hat{\boldsymbol{z}}(\bar{z}_i + \xi_i), \tag{6.12}$$

where the longitudinal component can be expressed either in Eulerian or Lagrangian form as convenient. The plasma is assumed initially quiescent, and the unperturbed plasma density $\bar{n}_0$ is assumed uniform, so over the length L we assume $N_e = \mathcal{V}\bar{n}_0 \gg 1$ electrons uniformly distributed at discrete intervals between $z = Z_0$ and $z = Z_0 + L$. For simplicity , we can choose the number of electrons to be commensurate with the number of grid lattice points, so that the inverse filing fraction $N_z/N_e = s$ is an integer, and we have

$$z_i(t=0) = \bar{z}_{si}, \quad i = 1, \ldots, N_e. \tag{6.13}$$

For the plasma particles, these equilibrium positions will either be considered classical equilibrium positions in the correspondence-principle limit, or regarded quantum mechanical averages of the initial positions, so are taken as fixed c-numbers in either case. That is, all the longitudinal dynamics (q-number dependence) is contained in the Lagrangian displacements ξ_i. Later, when we turn to the atomic case, the quantities analogous to the $\bar{z}_j$ will be the

center-of mass (COM) coordinates, which are dynamical q-number DOFs independent of the relative q-number electron positions analogous to the xi_i. If the fluid limit is ultimately desired, we may conveniently take $s = 1$, or $N_e = N_z$ without any real loss of generality.

For q-number (operator-valued) or c-number (complex-valued) quantities defined on the lattice, the discrete Fourier transforms effecting changes between the equilibrium discrete position-space representation (indexed either by j, or sometimes just by z if the meaning is clear) and the reciprocal-space representation (indexed either by n, or sometimes for simplicity directly by $k = k_n$) take the form:

$$f_n \equiv f_k = \frac{1}{\sqrt{N_z}} \sum_{j=1}^{N_z} f_j e^{-ik_n z_j} \equiv \frac{1}{\sqrt{N_z}} \sum_{z} f_z e^{-ikz}; \tag{6.14a}$$

$$f_j \equiv f_z = \frac{1}{\sqrt{N_k}} \sum_{n=-\frac{N_k}{2}-1}^{\frac{N_k}{2}} f_n e^{+ik_n z_j} \equiv \frac{1}{\sqrt{N_k}} \sum_{k} f_k e^{+ikz}. \tag{6.14b}$$

These two representations contain exactly the same information,and in fact are unitary transformations of each other, when evaluated on conjugate lattices where $N_k = N_z$, which ensures that canonical commutation relations (CCRs) will be preserved. That is, for $N_z = N_k$, $\left[f_j, f_{j'}^\dagger\right] = \delta_{jj'}$ if and only if $\left[f_n, f_{n'}^\dagger\right] = \delta_{nn'}$, where $\delta_{jj'}$ is the usual Kronecker delta symbol.

Now, especially in the context of plasmas (or atomic vapors), it may seem more than a little odd to describe such an amorphous media on a finite discrete lattice of positions and a finite discrete reciprocal lattice for conjugate momentum, but this is really just a convenient mathematical artifice used along the way to the continuum expressions, which can then be re-discretized in momentum for sufficiently narrow-band fields and/or periodically confined plasmas.

To ease doubts, we next explicitly transform to a hybrid description, where equilibrium particle positions for a finite number of electrons are allowed to take on continuous values, but the particles are still assumed confined to a region of length L, so momenta remain discretized, but can now assume a countably-infinite number of different values. Retaining the mathematically-convenient periodic boundary conditions rather than more realistic hard-wall boundary conditions, we can easily relate the continuous (but periodic) position-space quantity $f(z)$ to the countable momentum-space representation f_k.

Specifically, we consider the limit allowing $N_z = N_k \to \infty$ while fixing N_e and L, so that $\Delta z \to 0$ whereas $\Delta k = \frac{2\pi}{L}$ remains constant. Now, taking any full or partial continuum

limit is just equivalent to making certain substitutions of the form:

$$z_j \to z, \tag{6.15a}$$

$$b_j \to \sqrt{\tfrac{L}{N_z}} b(z); \tag{6.15b}$$

$$\delta_{jj'} \to \tfrac{L}{N_z} \delta(z - z'), \tag{6.15c}$$

$$\sum_j \to \tfrac{N_z}{L} \int dz \ , \tag{6.15d}$$

$$k_n \to k, \tag{6.15e}$$

$$a_n \to \sqrt{\tfrac{N_z}{N_k}} \sqrt{\tfrac{2\pi}{L}} a(k), \tag{6.15f}$$

$$\delta_{nn'} \to \tfrac{N_z}{N_k} \tfrac{2\pi}{L} \delta(k - k'), \tag{6.15g}$$

$$\sum_n \to \tfrac{N_k}{N_z} \tfrac{L}{2\pi} \int dk, \tag{6.15h}$$

everywhere needed, and canceling divergent parameters. Here $\delta(s)$ is the one-dimensional Dirac delta function, b_j can be any quantity which either is *linear* in the fields values (evaluated at position $\bar{z}_j$) or linear in the jth particle's position or velocity. Higher-order observables will be constituted from these basic linear ones.

Specifically, we use the position-dependent half of these correspondences, obtaining the usual Fourier series relations

$$f_n = \tfrac{1}{\sqrt{L}} \int_{z=Z_0}^{Z_0+L} dz \, f(z) e^{-ik_n z} \quad \text{for all } n \in \mathbb{Z}; \tag{6.16a}$$

$$f(z) = \tfrac{1}{\sqrt{L}} \sum_{n=-\infty}^{\infty} f_n e^{+ik_n z} \quad \text{for all } z \in [Z_0, Z_0 + L) \, . \tag{6.16b}$$

These relations are unitary over one a spatial period L, such that $\left[f(z), f(z')^\dagger\right] = \sum_j \delta(z - z' + jL)$ (i.e., a Dirac comb) if and only if $\left[f_n, f_{n'}^\dagger\right] = \delta_{nn'}$. Again, they are appropriate to a finite, discrete number of electrons and a discrete set of EM field modes all confined to a periodic box.

Taking the $L \to \infty$ limit of these results by using the wavenumber-dependent correspondences in (6.15), we obtain the usual description of a finite number N_e electrons in a continuum phase space. We can either track individual particle observables defined on this continuous phase space, or adopt a Klimontovitch-type representation, where electron densities (number, mass, charge, current, etc.) are written as sums over impulsive contributions

at the discrete electron positions. For example, the charge density is just

$$\rho_e(\boldsymbol{x}) = -e \sum_i \delta(\boldsymbol{x} - \boldsymbol{x}_i) \tag{6.17}$$

in $3D$, where $\delta(\boldsymbol{x}) = \delta^{(3)}(\boldsymbol{x})$ is the three-dimensional Dirac delta function, or

$$\rho_e(z) = -\tfrac{e}{\mathcal{A}} \sum_i \delta(z - z_i) \tag{6.18}$$

in $1D$ geometry. This is in principle an exact description of the plasma and fields. In this fully continuum limit, the Fourier transform pairs

$$f(k) = \tfrac{1}{\sqrt{2\pi}} \int dz\, f(z) e^{-ikz} \tag{6.19a}$$

$$f(z) = \tfrac{1}{\sqrt{2\pi}} \int dk\, f(k) e^{+ikz} \tag{6.19b}$$

take unitary form, so that $\left[f(z), f(z')^\dagger\right] = \delta(z - z')$ if and only if $\left[f(k), f(k')^\dagger\right] = \delta(k - k')$.

Finally, we can if desired consider the fluid limit, in which we replace the actual electrons with fluid elements, which have the same charge-to-mass ratio (and spin-to-charge-ratio) as ordinary electrons, but at least in our imaginations may be arbitrarily subdivided to approach a continuum limit. That is, we take particle number $N_e \to \infty$ and allow $q_e \to 0$ and $m_e \to 0$, so as to keep fixed the charge density $q_e \bar{n}_0$, mass density $m_e \bar{n}_e$, charge-to-mass ratio $\frac{q_e}{m_e}$, and therefore also the linear plasma frequency ω_p, linear cyclotron frequency Ω_c, and Bohr magneton $\mu_{\rm B} = \frac{e\hbar}{m_e c}$, where $\hbar$ is the usual reduced Planck's constant. The electron fluid is described in terms of continuous scalar or vector fields (e.g., velocity or momentum) and densities (e.g., mass or charge). The lattice model erred by making space un-physically discrete. The fluid model in effect errs by making particles un-physically continuous, but both limits have their uses. In this fully continuum limit, the Fourier transform pairs retain the unitary form (6.19).

Full (Non-Relativistic) Hamiltonian

With these conventions, the full non-relativistic Hamiltonian for EM fields and discrete plasma electrons can be written (in Gaussian units) as

$$\begin{aligned}\mathscr{H}_{\rm tot} = {} & \tfrac{\mathcal{A}}{8\pi} \int dz\, \left[\|\boldsymbol{E}_\perp\|^2 + \|\boldsymbol{B}_\perp\|^2 + \|\boldsymbol{B}_0\|^2\right] \\ & + \sum_j \left[\tfrac{1}{2m_e} \|\boldsymbol{P}_j - \tfrac{q_e}{c} \boldsymbol{A}_0(\boldsymbol{x}_j) - \tfrac{q_e}{c} \boldsymbol{A}_\perp(z_j, t)\|^2 + q_e \Phi(z_j, t) + \tfrac{e\hbar}{2m_e c} \boldsymbol{\sigma}_j \cdot (\boldsymbol{B}_\perp(z_j, t) + \boldsymbol{B}_0)\right],\end{aligned} \tag{6.20}$$

where: $\boldsymbol{A}_0(\boldsymbol{x})$ is the (time-independent) vector potential associated with the constant longitudinal magnetic field $\boldsymbol{B}_0 = \boldsymbol{\nabla} \times \boldsymbol{A}_0 = B_0\hat{\boldsymbol{z}}$; $\boldsymbol{A}_\perp(z,t)$ is the vector potential associated with the transverse electric fields $\boldsymbol{E}_\perp(z,t) = -\frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{A}_\perp$ and transverse magnetic fields $\boldsymbol{B}_\perp(z,t) = \boldsymbol{\nabla} \times \boldsymbol{A}_\perp$, and can be further decomposed as $\boldsymbol{A}_\perp = \boldsymbol{A}_1 + \boldsymbol{A}_2$ into probe, or signal (wave 1) and pump, or control (wave 2) contributions (with corresponding fields decomposed similarly); $\boldsymbol{P}_j = \boldsymbol{p}_j + \frac{q_e}{c}\boldsymbol{A}_0(\boldsymbol{x}_j) + \frac{q_e}{c}\boldsymbol{A}_\perp(z_j,t)$ is the *canonical* momentum of the jth particle, expressed as a function of the total vector potential at the particle's position and of the *kinetic* momentum $\boldsymbol{p}_j = m_e\boldsymbol{v}_j$, where $\boldsymbol{v}_j = \frac{d}{dt}\boldsymbol{x}_j$ is the particle velocity; $\Phi(z,t)$ is the electrostatic scalar potential; $\boldsymbol{\sigma}_j$ is the electron spin operator; and $\hbar = \frac{h}{2\pi}$ is the usual reduced Planck's constant.

In the $1D$ Coulomb gauge, the vector potential $\boldsymbol{A}(z,t)$ is both geometrically and functionally transverse, i.e.,

$$\hat{\boldsymbol{z}}\cdot\boldsymbol{A} = A_z = 0, \tag{6.21a}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{A} = \frac{\partial}{\partial z}A_z + \boldsymbol{\nabla}_\perp\cdot\boldsymbol{A}_\perp = 0, \tag{6.21b}$$

which means that the longitudinal ($\hat{\boldsymbol{z}}$) components of canonical and kinetic particle momenta will coincide. The scalar potential $\Phi(z,t)$ is associated with the unretarded Coulomb fields tied to the instantaneous positions of the electrons in the ionic background:

$$-\nabla^2\Phi = -\frac{\partial^2}{\partial z^2}\Phi(z,t) = 4\pi\rho(z,t) = 4\pi\bar{n}_0 e - 4\pi\frac{e}{\mathcal{A}}\sum_j \delta\big(z - z_j(t)\big), \tag{6.22}$$

and as such does not represent independent field degrees-of-freedom, so will be included in the particle Hamiltonian.

In this gauge, the vector potential generating the constant longitudinal magnetic field $\boldsymbol{B}_0 = B_0\hat{\boldsymbol{z}}$ is given by $\boldsymbol{A}_0 = \frac{1}{2}B_0(x\,\hat{\boldsymbol{y}} - y\,\hat{\boldsymbol{x}})$. Note that the vector potential has transverse position dependence, but the associated magnetic field does not. Because this axial field is constant, homogeneous, and in regimes of interest much larger than the pump or probe fields, it will be taken as classical throughout our analysis, i.e., $B_0 > 0$ is everywhere just regarded as a c-number.

The Hamiltonian can then be re-written as:

$$\mathscr{H}_{\text{tot}} = \mathcal{H}_{\text{EM}} + \mathcal{H}_p + \mathcal{H}_g + \mathcal{H}_{dp} + \mathcal{H}_{A^2} + \mathcal{H}_{\text{spin}} \tag{6.23}$$

where:

$$\mathcal{H}_{\text{EM}} = \frac{\mathcal{A}}{8\pi}\int dz\big[\|\boldsymbol{E}_\perp\|^2 + \|\boldsymbol{B}_\perp\|^2\big] + \frac{\mathcal{V}B_0^2}{8\pi} + \frac{e^2}{2m_e c^2}\sum_j \|\boldsymbol{A}_\perp(\bar{z}_j,t)\|^2, \tag{6.24}$$

is the Hamiltonian for the electromagnetic fields including the dielectric effects of the unperturbed background plasma;

$$\mathscr{H}_p = \sum_j \left[\tfrac{1}{2m} p_{z\,j}^2 - e\Phi(z_j,t)\right] \tag{6.25}$$

describes the longitudinal dynamics associated with the free plasma, that is to say any plasma or Langmuir wave;

$$\mathscr{H}_g = \sum_j \tfrac{1}{2m_e}\|\boldsymbol{P}_j - \tfrac{q_e}{c}\boldsymbol{A}_0(\boldsymbol{x}_j)\|^2 = \sum_j \tfrac{1}{2m_e}\left[\left(P_{x\,j}^2 - \tfrac{1}{2}m_e\Omega_c y_j\right)^2 + \left(P_{y\,j}^2 + \tfrac{1}{2}m_e\Omega_c y_j\right)^2\right] \tag{6.26}$$

generates the transverse gyration of electrons in the longitudinal magnetic field;

$$\mathcal{H}_{dp} = \tfrac{e}{2m_e c}\sum_j \left[\left(\boldsymbol{P}_j - \tfrac{q}{c}\boldsymbol{A}_0(\boldsymbol{x}_j)\right)\cdot\boldsymbol{A}_\perp(z_j,t) + \boldsymbol{A}_\perp(z_j,t)\cdot\left(\boldsymbol{P}_j - \tfrac{q}{c}\boldsymbol{A}_0(\boldsymbol{x}_j)\right)\right] \tag{6.27}$$

represents the (predominately dipolar) interaction between electrons and EM fields which is first-order in the pump or probe amplitudes;

$$\mathcal{H}_{A^2} = \tfrac{e^2}{2m_e c^2}\sum_j \left[\|\boldsymbol{A}_\perp(z_j,t)\|^2 - \|\boldsymbol{A}_\perp(\bar{z}_j,t)\|^2\right] \tag{6.28}$$

describes the coupling between electron motion and fields which is second-order in the transverse EM fields; and finally

$$\mathcal{H}_{\text{spin}} = \tfrac{1}{2}\hbar\Omega_c \sum_j \left[\sigma_{z\,j} + \boldsymbol{\sigma}_{\perp\,j}\cdot\tfrac{\boldsymbol{B}_\perp(z_j,t)}{B_0}\right] \tag{6.29}$$

is the spin contribution. We will address (and simplify) each of these terms in turn.

Transverse EM Hamiltonian: Photons

In the quantization region with periodic boundary conditions, we can decompose the transverse vector potential and fields into a countable set of Fourier modes. Although more general choices are possible, for polarization vectors we adopt the circular basis

$$\hat{\boldsymbol{\epsilon}}_{k\pm} = \hat{\boldsymbol{\epsilon}}_{|k|\pm} = \hat{\boldsymbol{e}}_\pm \equiv \tfrac{1}{\sqrt{2}}\left(\hat{\boldsymbol{x}} \pm \hat{\boldsymbol{y}}\right), \tag{6.30}$$

which is well adapted to our magnetized medium, satisfying:

$$\hat{\boldsymbol{\epsilon}}_{k\pm}^* = \hat{\boldsymbol{\epsilon}}_{k\mp}; \tag{6.31a}$$

$$\hat{\boldsymbol{\epsilon}}_{k+}^*\cdot\hat{\boldsymbol{\epsilon}}_{k+} = \hat{\boldsymbol{\epsilon}}_{k-}^*\cdot\hat{\boldsymbol{\epsilon}}_{k-} = 1; \tag{6.31b}$$

$$\hat{\boldsymbol{\epsilon}}_{k+}\cdot\hat{\boldsymbol{\epsilon}}_{k+} = \hat{\boldsymbol{\epsilon}}_{k-}\cdot\hat{\boldsymbol{\epsilon}}_{k-} = 0; \tag{6.31c}$$

$$\hat{\boldsymbol{\epsilon}}_{k+}^*\cdot\hat{\boldsymbol{\epsilon}}_{k+}^* = \hat{\boldsymbol{\epsilon}}_{k-}^*\cdot\hat{\boldsymbol{\epsilon}}_{k-}^* = 0; \tag{6.31d}$$

$$\hat{\boldsymbol{z}}\cdot\hat{\boldsymbol{\epsilon}}_{k\pm} = 0; \tag{6.31e}$$

$$i\hat{\boldsymbol{z}}\times\hat{\boldsymbol{\epsilon}}_{k\pm} = \pm\hat{\boldsymbol{\epsilon}}_{k\pm}. \tag{6.31f}$$

This coincides with the helicity eigenbasis for right-moving waves ($k > 0$), but to the conjugate of the helicity basis for left-moving waves ($k < 0$). We can then write

$$\boldsymbol{A}_\perp = \sum_k \sum_{\mu=\pm} \sqrt{\tfrac{2\pi}{\mathcal{A}L}} \sqrt{\tfrac{\hbar c^2}{\omega_{k\mu}}} \left[\hat{\boldsymbol{\epsilon}}_{k\mu} a_{k\mu} e^{+ikz} + \hat{\boldsymbol{\epsilon}}^*_{k\mu} a^*_{k\mu} e^{-ikz} \right]; \tag{6.32a}$$

$$\boldsymbol{B}_\perp = \boldsymbol{\nabla} \times \boldsymbol{A}_\perp = \sum_k \sum_{\mu=\pm} \sqrt{\tfrac{2\pi}{\mathcal{A}L}} \sqrt{\tfrac{\hbar c^2 k^2}{\omega_{k\mu}}} \mu \left[\hat{\boldsymbol{\epsilon}}_{k\mu} a_{k\mu} e^{+ikz} + \hat{\boldsymbol{\epsilon}}^*_{k\mu} a^*_{k\mu} e^{-ikz} \right]; \tag{6.32b}$$

$$\boldsymbol{E}_\perp = -\tfrac{1}{c} \tfrac{\partial}{\partial t} \boldsymbol{A}_\perp = \sum_k \sum_{\mu=\pm} i \sqrt{\tfrac{2\pi}{\mathcal{A}L}} \sqrt{\hbar \omega_{k\mu}} \left[\hat{\boldsymbol{\epsilon}}_{k\mu} a_{k\mu} e^{+ikz} - \hat{\boldsymbol{\epsilon}}^*_{k\mu} a^*_{k\mu} e^{-ikz} \right]; \tag{6.32c}$$

$$\tag{6.32d}$$

in terms of modal eigenfrequencies $\omega_{k\,\mu} \geq 0$ which are left unspecified for the moment, and annihilation and creation operators for photons satisfying the usual canonical commutation relations (CCRs):

$$\left[a_{k\mu}, a^\dagger_{k'\mu'} \right] = \delta_{\mu\mu'} \delta_{kk'}; \tag{6.33a}$$

$$\left[a_{k\mu}, a_{k'\mu'} \right] = \left[a^\dagger_{k\mu}, a^\dagger_{k'\mu'} \right] = 0. \tag{6.33b}$$

Also note that in these expressions, the sum is over all (positive and negative) wavenumbers, not just the positive half-space as in some field-theoretic expansions. Observing that $\frac{4\pi e^2 N}{m_e \mathcal{A}L} = \omega_p^2$ by our previous definitions, note that If these expansions for the fields are substituted into the final term in $\mathcal{H}_{\text{EM}}$, we find after some algebra that:

$$\tfrac{e^2}{2m_e c^2} \sum_j \| \boldsymbol{A}_\perp(\bar{z}_j, t) \|^2 = \tfrac{1}{4} \hbar \omega_p \sum_{k\,\mu} \tfrac{\omega_p}{\omega_{k\mu}} \left[2 a^\dagger_{k\mu} a_{k\mu} + a_{k\mu} a_{-k\mu} + a^\dagger_{k\mu} a^\dagger_{-k\mu} + \left(\tfrac{\omega_p}{c|k|} + \tfrac{c|k|}{\omega_p} \right) \right], \tag{6.34}$$

and then after some more algebra, obtain

$$\begin{aligned} \mathcal{H}_{\text{EM}} &= \tfrac{1}{2} \sum_{k\,\mu} \hbar \omega_{k\mu} \left(\tfrac{\omega_p^2 + c^2 k^2 + \omega_{k\mu}^2}{\omega_{k\mu}^2} \right) a^\dagger_{k\mu} a_{k\mu} \\ &\quad + \tfrac{1}{4} \sum_{k\,\mu} \hbar \omega_{k\mu} \left(\tfrac{\omega_p^2 + c^2 k^2 - \omega_{k\mu}^2}{\omega_{k\mu}^2} \right) \left[a_{k\mu} a_{-k\mu} + a^\dagger_{k\mu} a^\dagger_{-k\mu} \right] + constant \end{aligned} \tag{6.35}$$

Through cancellation of cross-terms, this will assume the familiar form of a bilinear, positive-definite harmonic oscillator Hamiltonian if and only if the eigenfrequencies satisfy

$$\omega_{k\mu} = \omega_{|k|} = +\sqrt{\omega_p^2 + c^2 k^2} \equiv \omega_0(k), \tag{6.36}$$

which, reassuringly, is the usual classical dispersion relation for EM waves in a cold non-relativistic plasma. So, after dropping irrelevant constants, the "free" EM field Hamiltonian can at last be written in discrete or continuum form, as

$$\mathcal{H}_{\text{EM}} = \sum_{k\,\mu} \hbar \omega_k a^\dagger_{k\mu} a_{k\mu} \;\leftrightarrow\; \sum_\mu \int dk\, \hbar \omega_0(k)\, a_\mu(k)^\dagger a_\mu(k). \tag{6.37}$$

In the first of what will be several stages of operator "dressing," we observe that the photons of energy $\hbar\omega_0(k)$ created by the $a^\dagger_{k\mu}$ are not true vacuum photons, but already include the dielectric effects, or "virtual quiver" response, of the background plasma. We could have started with vacuum photons satisfying $\omega(k) = ck$ but this would have entailed the additional complication of another layer of dressing to transform from vacuum photons to photons in the unmagnetized plasma, and is best avoided. As it stands, this formalism will work with either co-propagating or counter-propagating pump and probe fields, but will not work in the recently-discussed case of undulator-induced-transparency (UIT)[257, 258, 259], where the traveling-wave pump field is replaced with a static undulator field, which must be described quantum-mechanically by zero-frequency virtual, or "off-shell", photons which do not satisfy the unmagnetized plasma dispersion relation. However, a classical static undulator field could be added if desired. Nor can our $1D$ formalism capture the cross-propagating geometry sometimes assumed in the atomic case, which allows for time-dependent but spatially-independent control profiles. In the case of atomic vapors, crossed, linearly-polarized beams can still couple to dipole transitions into the same excited atomic state, but are of less interest here where helicity relative to the axial magnetic field is what matters.

Beware that two different scalings for a normalized, or dimensionless, vector potential may appear. In the classical convention, the transverse vector potential $\boldsymbol{a} = \frac{e}{m_e c^2}\boldsymbol{A}$ is normalized so as to yield the relativistically normalized kinetic quiver momentum $\gamma\beta_\perp = \boldsymbol{a}$ for the electrons, at least in the $1D$ case in which transverse canonical momentum is conserved. In the usual quantum convention, $a_{k\mu}$ is normalized such that in the case of discrete modes, $a^\dagger_{k\mu}a_{k\mu}$ yields the number of photons of momentum $\hbar k$ and polarization μ in the quantization volume $\mathcal{V}$. For the pump at least, it will be useful to translate between these conventions, using

$$\frac{e}{m_e c^2}\hat{\boldsymbol{\epsilon}}^*_+ \cdot \boldsymbol{A}_{\rm pump} = a_{\rm pump} = \frac{1}{\sqrt{N}}\frac{\sqrt{\hbar\omega_p}}{\sqrt{m_e c^2}}\frac{\sqrt{\omega_p}}{\sqrt{\omega_0(k_2)}}a_{k_2+} \tag{6.38}$$

in the case of discrete modes. Although one always hears that $\hbar$ cannot appear in any classical theory, there is no real reason that e, c, and the pure number we call the fine structure constant, $\alpha_e = \frac{e^2}{\hbar c} \approx \frac{1}{137}$ cannot appear, so either the "classical" or "quantum" normalizing conventions can be used as convenient in classical or quantum treatments, as long as we remain consistent.

Longitudinal Electron Contribution: Plasmons

In the Coulomb gauge, the scalar potential $\Phi(z,t)$ is a function only of instantaneous particle positions, and for the purpose of determining the potential or the corresponding longitudinal space-charge field $E_z(z,t) = -\frac{\partial}{\partial} z\Phi$, in $1D$ geometry the electrons act as transverse sheets of uniform surface charge density $\frac{q_e}{\mathcal{A}}$, and the immobile ions as sheets of opposite surface charge density. In a cold, non-relativistic $1D$ plasma, coherent Langmuir oscillations of this electric field (or the corresponding potential, or electron density) at frequency ω_p and any amplitude $|E|_z < E_0$ can be supported, where $E_0 = \frac{m_e c \omega_p}{e}$ is known as the cold linear wave-breaking limit. In the Eulerian fluid picture, this upper limit will correspond to the breakdown of velocity single-valuedness and the appearance of density singularities , while in the Lagrangian framework, it corresponds to crossings of the electron sheets.

Above wave-breaking, the coherent plasma wave required for EIT cannot persist, so we really only care about the plasma waves below wave-breaking, which take on a remarkably simple form in $1D$ when expressed in terms of the Lagrangian particle displacements. In the quiescent, neutral, homogeneous plasma, where al $\xi_j = 0$, the longitudinal electric field vanishes: $E_z(z, t = 0) = 0$. Since in $1D$ each charge sheet just produces a constant electric field of magnitude $2\pi |\sigma_e|$ on either side of the sheet, then as long as the electron charge sheets never cross, the longitudinal electric field experienced by the jth electron is just directly proportional to the net imbalance of ion charge sheets due to the electrons displacement:

$$E_z(z_j, t) = \tfrac{m_e \omega_p^2}{e} \xi_j(t) = E_0 \tfrac{\xi_j}{\xi_0}, \tag{6.39}$$

where $\xi_0 = c/\omega_p$ is known as the collisionless skin depth. Note that for the negatively-charged electrons, this always corresponds to a *linear restoring force*, regardless of how complicated or nonlinear the actual trajectory $\xi_j(t)$ might be. The corresponding scalar potential can then be taken to be:

$$\Phi(z_j, t) = -\tfrac{1}{2} \tfrac{m_e \omega_p^2}{e} \xi_j(t)^2 = -\tfrac{1}{2} \tfrac{m_e \omega_p^2}{e} \left[z_j(t) - \bar{z}_j \right]^2 \tag{6.40}$$

Now defining

$$f_j = \sqrt{\tfrac{m_e \omega_p}{2\hbar}} \xi_j + i \tfrac{1}{\sqrt{2 m_e \hbar \omega_p}} p_{z\,j} \quad \text{and} \tag{6.41a}$$

$$f_j^\dagger = \sqrt{\tfrac{m_e \omega_p}{2\hbar}} \xi_j - i \tfrac{1}{\sqrt{2 m_e \hbar \omega_p}} p_{z\,j} \tag{6.41b}$$

for each electron, and using the fact that, interpreted quantum-mechanically, ξ_j and $p_{z\,j}$ will satisfy the usual CCRs for particle positions and momentum:

$$\left[\xi_j, p_{z\,j'} \right] = \left[\bar{z}_j + \xi_{j'}, P_{z\,j'} \right] = \left[z_j, P_{z\,j'} \right] = i\hbar\, \delta_{jj'}, \tag{6.42}$$

then the $b_j^\dagger$ and b_j satisfy the CCRs for creation and annihilation operators:

$$\left[f_j\,,f_{j'}^\dagger\right]=\delta_{jj'};\tag{6.43a}$$

$$\left[f_j\,,f_{j'}\right]=\left[f_j^\dagger,f_{j'}^\dagger\right]=0;\tag{6.43b}$$

and this part of the Hamiltonian can, after a bit of manipulation, be written in standard Harmonic oscillator form:

$$\mathcal{H}_p=\sum_j\left[\tfrac{1}{2m}p_{z\,j}^2-e\Phi(z_j,t)\right]=\sum_j\left[\tfrac{1}{2m}p_{z\,j}^2+\tfrac{1}{2}\tfrac{m_e\omega_p^2}{e}\xi_j(t)^2\right]=\sum_j\hbar\omega_p f_j^\dagger b_j+constant,\tag{6.44}$$

Taking continuum limits and Fourier transforms, and dropping constant terms, this becomes

$$\mathcal{H}_p=\sum_j\hbar\omega_p f_j^\dagger f_j\;\leftrightarrow\;\int dz\,\hbar\omega_p f(z)^\dagger f(z)=\int dk\,\hbar\omega_p f(k)^\dagger f(k)\;\leftrightarrow\;\sum_k\hbar\omega_p f_k^\dagger f_k,\tag{6.45}$$

representing plasmons, or quanta of the longitudinal plasma wave oscillation, each of energy $\hbar\omega_p$. In the Schrödinger state picture, we will call the corresponding harmonic-oscillator-like quantum states Langmuir levels. Because particle DOFs are independent of the transverse field DOFs, these plasmon operators commute with the photon operators introduced above:

$$\left[a_{k\mu},f_{k'}\right]=\left[a_{k\mu}^\dagger,f_{k'}\right]=0;\tag{6.46a}$$

$$\left[a_{k\mu},f_{k'}^\dagger\right]=\left[a_{k\mu}^\dagger,f_{k'}^\dagger\right]=0.\tag{6.46b}$$

In the cold, non-relativistic limit, the plasmon frequency remains equal to ω_p, independent of the wavenumber, so plasmon energy is independent of momentum. Thermal corrections would contribute dispersion, and relativistic effects would additionally contribute nonlinear detunings and couplings to other DOFs.

Also, note carefully that the plasmon operator $f(z)$ is something of a odd creature, since its position argument is (the fine-grained limit of) the *equilibrium* position of the Lagrangian electron sheet which may be instantaneously displaced to some other position. While $\hbar\omega_p f(z)^\dagger f(z)$ may look like the continuum Eulerian energy density for the electron fluid element currently located at position z, rather it is the energy density of the infinitesimal Lagrangian fluid element whose equilibrium position is z.

This might appear awkward, but for our purposes these continuum Lagrangian operators are more convenient than the Eulerian versions, in which the constancy of the plasma

frequency emerges as a very non-obvious consequence of the intrinsically nonlinear equations of motion. One can actually derive physically-equivalent results using an Eulerian formalism where fluid quantities are decomposed by ansatz into plasmon normal coordinates in a manner analogous to our development of the photon dynamics above.[5] In particular, if we decompose the potential as

$$\Phi(z,t) = \sqrt{\tfrac{2\pi}{\mathcal{A}L}}\sqrt{\hbar\omega_p}\sum_k \left[\tfrac{i}{k}\, b_k e^{+ikz} - \tfrac{i}{k}\, b_k^\dagger e^{-ikz}\right] \tag{6.47}$$

for plasmon operators postulated to satisfy the CCRs, then the corresponding electric field must be decomposed as

$$E_z(z,t) = -\tfrac{d}{dx}\Phi = \sqrt{\tfrac{2\pi}{\mathcal{A}L}}\sqrt{\hbar\omega_p}\sum_k \left[b_k e^{+ikz} + b_k^\dagger e^{-ikz}\right], \tag{6.48}$$

and from Gauss's law, $\frac{d}{dz}E = 4\pi e(\bar{n}_0 - n_e)$, we deduce

$$n_e = \bar{n}_0 + \tfrac{1}{4\pi e}\sqrt{\tfrac{2\pi}{\mathcal{A}L}}\sqrt{\hbar\omega_p}\sum_k \left[-ik\, b_k e^{+ikz} + ik\, b_k^\dagger e^{-ikz}\right]; \tag{6.49}$$

and finally determine the current density decomposition consistent with the continuity equation (or equivalently, with Ampere's law):

$$J_e = q_e\, n_e v_e = \tfrac{1}{4\pi}\sqrt{\tfrac{2\pi}{\mathcal{A}L}}\sqrt{\hbar\omega_p}\,\omega_p \sum_k \left[i\, b_k e^{+ikz} - i\, b_k^\dagger e^{-ikz}\right]. \tag{6.50}$$

After taking the continuum limit and Fourier transforms, we obtain Eulerian plasmon operators $b(z)$ and $b(z)^\dagger$ which collectively contain the same information as the continuum Lagrangian operators $f(z)$ and $f(z)^\dagger$, but the corresponding plasma-wave contribution to the Eulerian Hamiltonian, just equal to the sum of kinetic and potential energy, i.e.

$$\mathcal{H}_p^{(\mathrm{E})} = \tfrac{1}{2}\int dz\, m_e \tfrac{J_e(z)^2}{q_e^2 n_e(z)} + \tfrac{1}{2}\int dz\, q_e n_e(z)\Phi(z) \tag{6.51}$$

is then no longer simply quadratic in the plasmon operators.

Transverse Electron Contribution: Gyrons

To simplify the form of the transverse plasma dynamics, we first define Hermitian transverse velocity operators

$$V_{x\,j} = \tfrac{1}{m_e}P_{x\,j} - \tfrac{1}{2}\Omega_c y_j; \tag{6.52a}$$

$$V_{y\,j} = \tfrac{1}{m_e}P_{y\,j} + \tfrac{1}{2}\Omega_c x_j; \tag{6.52b}$$

$$\tag{6.52c}$$

[5] The Hamiltonian aspects of (warm and cold) relativistic and non-relativistic $1D$ Eulerian fluid models are discussed more fully in the thesis of Ryan Lindberg.

and Hermitian transverse guiding-center coordinate variables

$$X_j = x_j - \Omega_c^{-1} V_{y\,j}; \tag{6.53a}$$

$$Y_j = y_j + \Omega_c^{-1} V_{x\,j}. \tag{6.53b}$$

Then using the CCRs for the transverse particle position $\boldsymbol{x}_{\perp\,j}$ and canonical momentum $\boldsymbol{P}_{\perp\,j}$, i.e.,

$$\left[x_{j'}, P_{x\,j'}\right] = \left[y_{j'}, P_{y\,j'}\right] = i\hbar\,\delta_{jj'}, \tag{6.54a}$$

$$\left[x_{j'}, P_{y\,j'}\right] = \left[y_{j'}, P_{x\,j'}\right] = 0, \tag{6.54b}$$

$$\left[x_{j'}, y_{j'}\right] = \left[P_{x\,j}, P_{y\,j'}\right] = 0, \tag{6.54c}$$

$$\tag{6.54d}$$

we immediately obtain

$$\left[X_j, Y_{j'}\right] = \ \ \tfrac{i\hbar}{m_e\Omega_c}\delta_{jj'}, \tag{6.55a}$$

$$\left[V_{x\,j}, V_{y\,j'}\right] = -\tfrac{i\hbar\Omega_c}{m_e}\delta_{jj'}. \tag{6.55b}$$

Next we define discrete creation and annihilation operators associated with the quanta of gyration, or "gyrons:"

$$g_j = \sqrt{\tfrac{m_e}{2\hbar\Omega_c}}\left[V_{x\,j} - iV_{y\,j}\right], \tag{6.56a}$$

$$g_j^\dagger = \sqrt{\tfrac{m_e}{2\hbar\Omega_c}}\left[V_{x\,j} + iV_{y\,j}\right], \tag{6.56b}$$

and discrete creation and annihilation operators associated with the quantized guiding center motion, whose excitations we we will call "centrons:"

$$c_j = \sqrt{\tfrac{m_e\Omega_c}{2\hbar}}\left[X_j + iY_j\right], \tag{6.57a}$$

$$c_j = \sqrt{\tfrac{m_e\Omega_c}{2\hbar}}\left[X_j - iY_j\right]. \tag{6.57b}$$

It is straightforward to verify that these satisfy the CCRs:

$$\left[g_j, g_{j'}^\dagger\right] = \left[c_j, c_{j'}^\dagger\right] = \delta_{jj'}, \tag{6.58a}$$

$$\left[g_j, c_{j'}^\dagger\right] = \left[g_j, c_{j'}\right] = 0, \tag{6.58b}$$

$$\left[g_j^\dagger, c_{j'}^\dagger\right] = \left[g_j^\dagger, c_{j'}\right] = 0. \tag{6.58c}$$

Because they are constructed from independent degrees-of-freedom, they also commute with all photon and plasmon operators, but as the menagerie of modes grows we will now adopt

the convention that any commutators not explicitly listed will be assumed to vanish unless otherwise specified.

Apart from an overall constant, which can be neglected, the transverse plasma Hamiltonian can be written as:

$$\mathcal{H}_g = \sum_j \hbar\Omega_c \, g_j^\dagger g_j \;\leftrightarrow\; \int dz \, \hbar\Omega_c \, g(z)^\dagger g(z) = \int dk \, \hbar\Omega_c \, g(k)^\dagger g(k). \tag{6.59}$$

Note that the Hamiltonian is completely independent of the guiding-center coordinates, as would be expected in a uniform field. In the single-particle picture, g_j just corresponds to a lowering operator for the equally-spaced Landau levels of the jth electron, analogous to a harmonic oscillator with frequency Ω_c. After continuum limits and Fourier transforms, the operator $g(k)$ may be interpreted as the annihilation operator for a collective gyron excitation of momentum $\hbar k$ and energy $\hbar\Omega_c$. It will prove convenient to express the Hamiltonian in terms of these collective-looking modes even though the dynamics arise exclusively from single-particle gyration effects; i.e., we might say that these are collective but not cooperative modes, unlike the plasmon degrees-of-freedom, which were both collective and cooperative in the sense that the longitudinal electric field arises as an intrinsically many-body effect (although we used the self-consistent Lagrangian single-particle equations of motion to deduce the dynamics). As with the plasmons, the continuum gyron operators are Lagrangian fluid quantities, such that operator $\hbar\Omega_c g(z)^\dagger g(z)$ represents the energy of gyration for the electron fluid elements whose equilibrium, not instantaneous, longitudinal position is given by z. These positions of course coincide if only gyrons and no plasmons are excited, but in general both gyrons (transverse motion) and plasmon (longitudinal motion) are present simultaneously.

Dipole Interactions: Gyron-Photon Coupling

Next we consider the interaction terms involving the transverse electron currents and the transverse EM field, *evaluated at the unperturbed longitudinal particle positions.* These are essentially dipole interactions, and will lead to the resonant coupling between the EM modes and the particle gyration, responsible for the absorption whose suppression remains the goal of EIT. Using the above definitions, we can write

$$\big(\boldsymbol{P}_{\perp j} - \tfrac{q_e}{c}\boldsymbol{A}_0(\boldsymbol{x}_{\perp j})\big) = \sqrt{m_e\hbar\Omega_c}\big(g_j\,\hat{\boldsymbol{\epsilon}}_+ + g_j^\dagger\,\hat{\boldsymbol{\epsilon}}_-\big). \tag{6.60}$$

Because this contains only transverse particle coordinates, it will commute with $\boldsymbol{A}_\perp(\bar{z}_j, t)$ (and with $\boldsymbol{A}_\perp(z_j, t)$ as well). We substitute in the modal decompositions, and simplify using

the Fourier transforms and commutation relations, and when the algebraic dust settles we find

$$\mathcal{H}_{dp} = \tfrac{1}{\sqrt{2}}\hbar\omega_p \int dk\, \tfrac{\sqrt{\Omega_c}}{\sqrt{\omega_0(k)}} \left[a_+(k) g(k)^\dagger + a_+(k)^\dagger g(k) + a_-(k) g(-k) + a_-(k)^\dagger g(-k)^\dagger \right]. \tag{6.61}$$

Contributions of this form are known generically as "2-wave interactions" in plasma physics. The interpretation of at least of the RWA terms (involving one annihilation and one creation operator) is clear: they effect absorption or emission of a photon within the plasma by gyrating electrons, or equivalently a conversion between photons and gyrons at the same wavenumber, and thus specifically represent cyclotron resonance effects. The anti-resonant, or non-RWA terms (involving two creation or two annihilation operators) do not conserve the total number of quanta, but their effects tend to oscillate and average away, so later, we will simplify matters further by mostly dropping the non-RWA terms, but we will retain all terms for now until we can better assess the accuracy of the RWA.

Spin Effects

The electron spin will couple to both the longitudinal and transverse magnetic fields. In realistic parameter regimes, the (classical) longitudinal magnetic field will be much stronger than any pump or probe fields, i.e., $B_0 \gg \|\boldsymbol{B}_\perp\|$, so the interaction energy per plasma electron associated with the transverse spin-coupling is expected to be very small compared to just one probe quanta,

$$\tfrac{\mathcal{H}_{\text{spin}\,\perp}}{N} \sim \tfrac{1}{2}\tfrac{\|\boldsymbol{B}_\perp\|}{B_0}\hbar\Omega_c \ll \hbar\Omega_c \sim \hbar\omega_1, \tag{6.62}$$

and can be safely neglected in most circumstances. The longitudinal contribution $\mathcal{H}_{\text{spin}\,z} = \frac{1}{2}\sum_j \hbar\Omega_c \sigma_{j\,z}$ involves only the static longitudinal field and not any dynamical degrees-of-freedom of relevance here, so for our purposes can be disregarded regardless of its magnitude.

3-Wave Interactions: Raman Scattering and Other Effects

So far, we have actually included only the effects of plasma-EM field interactions evaluated at the equilibrium electron positions $\bar{z}_j$, but now we evaluate the leading-order corrections, assuming the plasma wave remains sufficiently small. Although the Hamiltonian involves sums (or integrals) over all wavenumbers k, the relevant excitation will, by assumption, be confined to the neighborhoods of the probe carrier wavenumber $k_1 \sim O\left(\frac{\Omega_c}{c}\right)$, the pump carrier wavenumber $k_2 \sim O\left(\frac{\Omega_c - \omega_p}{c}\right) \sim O\left(\frac{\Omega_c}{c}\right)$, or the beat, or plasmon, carrier

wavenumber $k_p = k_1 - k_2 \sim O\big(\frac{\Omega_c \pm \omega_p}{c}\big)$ (depending on whether the pump co-propagates or counter-propagates relative to the probe). For plasma waves far below wave-breaking, the corresponding longitudinal particle displacements satisfy $|\xi_j| \ll \xi_0 \equiv \frac{c}{\omega_p}$, but because we shall also assume $\omega_p < \Omega_c \lesssim O(10)\,\omega_p$, it follows that $\frac{c}{\omega_p} \sim k_1^{-1}$, k_2^{-1}, k_p^{-1}, and therefore in this regime we may assume

$$|\xi_j| \ll k_1^{-1},\, k_2^{-1},\, k_p^{-1}, \tag{6.63}$$

or $|k\xi_j| \ll 1$ for all electrons and all relevant wavenumbers at which appreciable excitation might occur. Therefore, we can expand the complex exponentials appearing in $\mathcal{H}_{A^2}$ or $\mathcal{H}_{\rm int}$ as

$$e^{iz_j} - e^{i\bar{z}_j} = e^{i\bar{z}_j + i\xi_j} - e^{i\bar{z}_j} = e^{i\bar{z}_j}\left[e^{i\xi_j} - 1\right] \approx e^{ik\bar{z}_j}\left[ik\xi_j - \tfrac{1}{2}(k\xi_j)^2 - i\tfrac{1}{3}(k\xi_j)^3 + \ldots\right]. \tag{6.64}$$

Writing the Lagrangian displacement in terms of a linear combination of the plasmon creation and annihilation operators as

$$\xi_j = \tfrac{\sqrt{\hbar}}{\sqrt{2m_e\omega_p}}\big(f_j + f_j^\dagger\big) \to \tfrac{1}{\sqrt{\mathcal{A}\bar{n}_0}}\tfrac{\sqrt{\hbar}}{\sqrt{2m_e\omega_p}}\big(f(\bar{z}) + f(\bar{z})^\dagger\big), \tag{6.65}$$

we see that we can formally develop the remainder of the Hamiltonian as an infinite series of various interaction terms between the probe EM, pump EM, gyron, and Langmuir modes, which, quantum mechanically, may be interpreted as various kinds of "many-body" scattering/conversion/decay processes for the corresponding quanta, or quasi-particles. Consistent with our ongoing assumptions that the pump intensity greatly exceeds the probe intensity, and the plasma wave remains small, we retain only the leading-order (cubic) terms in this expansion, in which the pump photon field or pump-induced gyron field appears linearly and multiplies a bilinear product of the probe, probe-induced gyron, or plasmon modal excitation operators.

Arising from the $O(\|\boldsymbol{A}_\perp\|^2)$ terms, one category of such 3-wave processes involves the pump photon, probe photon, and plasma waves:

$$\begin{aligned}\mathcal{H}_{\mathcal{R}} = {}& \tfrac{\hbar\omega_p}{8\sqrt{\pi}}\tfrac{\sqrt{\hbar\omega_p}}{\sqrt{m_e c^2}}\sum_{\mu\,\mu'}\iint \tfrac{dk\,dk'}{\sqrt{\bar{n}_0\mathcal{A}}}\,\hat{\boldsymbol{\epsilon}}\cdot\hat{\boldsymbol{\epsilon}}'^{*}\,\tfrac{ic(k-k')}{\sqrt{\omega_0\omega_0'}}a_\mu(k)a_{\mu'}(k')^\dagger\left[f(k'-k) + f(k-k')^\dagger\right] + h.c. \\ & + \tfrac{\hbar\omega_p}{8\sqrt{\pi}}\tfrac{\sqrt{\hbar\omega_p}}{\sqrt{m_e c^2}}\sum_{\mu\,\mu'}\iint \tfrac{dk\,dk'}{\sqrt{\bar{n}_0\mathcal{A}}}\,\hat{\boldsymbol{\epsilon}}\cdot\hat{\boldsymbol{\epsilon}}'\,\tfrac{ic(k+k')}{\sqrt{\omega_0\omega_0'}}a_\mu(k)a_{\mu'}(k')\left[f(-k-k') + f(k+k')^\dagger\right] + h.c.\end{aligned} \tag{6.66}$$

where here we have used the abbreviated notation $\omega_0 = \omega_0(k)$, $\omega_0' = \omega_0(k')$, $\hat{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\epsilon}}_\mu(k)$, and $\hat{\boldsymbol{\epsilon}}' = \hat{\boldsymbol{\epsilon}}_{\mu'}(k')$, the hermitian conjugate terms have not been explicitly specified but must also be included, and the k-space integrations can be confined to a mutual range such that each

term includes the contributions from a product of one pump and one probe field operator only.

At the level of the RWA, this will correspond to what is specifically referred to as Raman scattering in plasma physics, namely the decay of a (transverse) EM probe photon into one (transverse) pump photon at lower frequency and one (longitudinal) plasmon, or the inverse of this process. In such processes frequency and wavenumber resonance conditions must be met, corresponding via the Planck-Einstein relations to conservation of total quasi-particle energy and momentum. Also certain action sums or differences are invariant, corresponding to the conservation of quanta. These action conservation laws are the famous Manley-Rowe relations, which arise quite naturally in the collective-mode Hamiltonian formalism when the RWA is invoked. For (RWA) Raman scattering in particular, we see that the total number of probe photons plus plasmons, and the excess of pump photons over plasmons, are both conserved.

Raman scattering in plasmas is analogous in many respects to what are termed Raman transitions in atomic physics: inelastic photon scattering normally involving a transition which is forbidden under single-photon processes, in which with the energy and/or momentum difference between absorbed and emitted photons is taken up by the medium – only in atomic physics it can occur in a single atom, while in the plasma it involves energy/momentum transfer to a collective mode involving the cooperative motion of many unbound electrons. In either the plasma or atomic cases, Raman scattering differs from fluorescence in that, although it is also a two-photon process, the intermediate state involved is effectively virtual and can be off-resonance, whereas fluorescence occurs as a distinct sequence of an on-resonance one-photon absorption event followed by a one-photon emission event. Note that in our derivation for the plasma-field couplings, the Raman scattering terms arise directly at first-order in the Hamiltonian (in the Dyson series for the propagator, for example), not as the second-order effect of an iterated one-photon term. (This is not always the case for single atoms.)

Expanding the remaining $O(\boldsymbol{P}_j \cdot \boldsymbol{A}_\perp)$ terms to first order in $k\xi_j$, we find another family of 3-wave terms involving the pump, plasma, and cyclotron waves:

$$\begin{aligned}\mathcal{H}_{\mathcal{L}} = &\tfrac{\hbar\omega_p}{2\sqrt{2\pi}}\tfrac{\sqrt{\hbar\Omega_c}}{\sqrt{m_e c^2}}\iint \tfrac{dk\,dk'}{\sqrt{\bar{n}_0\mathcal{A}}}\,\tfrac{-ick}{\sqrt{\omega_0(k)\omega_p}}a_+(k)^\dagger g(k')\left[f(k-k')+f(k'-k)^\dagger\right]+h.c.\\ &+\tfrac{\hbar\omega_p}{2\sqrt{2\pi}}\tfrac{\sqrt{\hbar\Omega_c}}{\sqrt{m_e c^2}}\iint \tfrac{dk\,dk'}{\sqrt{\bar{n}_0\mathcal{A}}}\,\tfrac{ick}{\sqrt{\omega_0(k)\omega_p}}a_-(k)g(k')\left[f(-k-k')+f(k+k')^\dagger\right]+h.c.\end{aligned} \tag{6.67}$$

In the veritable plethora of plasma dispersion relations, instabilities, and 3-wave interactions, the processes described by $\mathcal{H}_{\mathrm{L}}$ do not appear to have been named officially; we propose

the name Landau scattering, since Landau pioneered the correct classical analysis of one of the waves involved (plasmons) and the quantum-mechanics behind the other (gyrons). These terms describe two distinct types of scattering processes (and their inverse processes.) One is just like the earlier Raman scattering, but it is the probe-induced gyron, rather the probe photon itself, that decays into a pump photon and plasmon. The second type of process involves the decay of a probe photon into a pump-resonant gyron and plasmon. We will see that the first type of process is responsible for EIT, while the second is basically just a nuisance, a competing process that cannot contribute to the suppression of probe absorption.

The Discrete-Mode EIT Model Hamiltonian

In a slight abuse of our previous notation, for greater readability we now denote the wavenumber dependence as a parenthetical rather than subscripted argument even for the case of discrete modes. In the analysis to follow, the choice between discrete or continuum modes will either be explicitly specified or should be clear from context, so hopefully no confusion will arise.

Perhaps we should also obviate an opportunity for a certain confusion over the conventional names for these modes. We follow conventional EIT terminology, in that what we usually call the pump or control, the probe or signal, and the plasma wave or longitudinal excitation are known in the atomic case respectively as the pump or control field, the probe or signal field, and either the material excitation, atomic excitation, atomic polarization, or occasionally the pseudo-spin density (because the algebra of the atomic excitations is formally analogous to a spin system). However, because $\omega_p < \omega_1 < \omega_2$, what are here referred to as the pump/control, probe/signal, and plasma wave/longitudinal excitation correspond to what, in the contexts of either wave-wave interactions in plasma physics, or parametric amplification, frequency-doubling, or related nonlinear optical phenomenon, would be called the signal, pump, and idler waves, respectively. That is, the frequency ordering of pump and probe fields is here reversed relative to more conventional settings. Part of what makes true EIT interesting and distinctive is that its control field is downshifted in frequency relative to the probe field, but this breaks with tradition for what is called the pump and the signal.

For the case of discrete modes all assumed in exact one-quanta or two-quanta (Raman)

resonance as appropriate, the Hamiltonian becomes

$$\mathcal{H}_{\rm EIT} = \mathcal{H}_a + \mathcal{H}_f + \mathcal{H}_g + \mathcal{H}_{ag} + \mathcal{H}_{afa} + \mathcal{H}_{afg} \tag{6.68}$$

where the uncoupled Hamiltonians are

$$\begin{aligned}\mathscr{H}_a = {} & \hbar\omega_1 \Big[a_+(k_1)^\dagger a_+(k_1) + a_-(k_1)^\dagger a_-(k_1) + a_+(-k_1)^\dagger a_+(-k_1) + a_-(-k_1)^\dagger a_-(-k_1)\Big] \\ & + \hbar\omega_2 \Big[a_+(k_2)^\dagger a_+(k_2) + a_-(k_2)^\dagger a_-(k_2) + a_+(-k_2)^\dagger a_+(-k_2) + a_-(-k_2)^\dagger a_-(-k_2)\Big];\end{aligned} \tag{6.69}$$

$$\begin{aligned}\mathscr{H}_f = {} & \hbar\omega_p \Big[f(k_p)^\dagger f(k_p) + f(2\bar{k})^\dagger f(2\bar{k})\Big] \\ & + \hbar\omega_p \Big[f(-2\bar{k})^\dagger f(-2\bar{k}) + f(-k_p)^\dagger f(-k_p)\Big]; \qquad \text{and}\end{aligned} \tag{6.70}$$

$$\mathscr{H}_g = \hbar\Omega_c \Big[g(k_1)^\dagger g(k_1) + g(-k_1)^\dagger g(-k_1) + g(k_2)^\dagger g(k_2) + g(-k_2)^\dagger g(-k_2)\Big], \tag{6.71}$$

where here we have used the abbreviations $\omega_1 = \omega_0(k_1)$ and $\omega_2 = \omega_0(k_2)$ for the unperturbed EM frequencies, and $k_p = k_1 - k_2$ and $\bar{k} = \frac{1}{2}(k_1 + k_2)$ for the beat and average wavenumbers, respectively; the 2-wave couplings (including for now both RWA and non-RWA terms) are

$$\begin{aligned}\mathscr{H}_{ag} = {} & \tfrac{1}{\sqrt{2}}\hbar\omega_p \tfrac{\sqrt{\Omega_c}}{\sqrt{\omega_1}} \Big[a_+(k_1)^\dagger g(k_1) + a_+(k_1) g(k_1)^\dagger + a_-(k_1) g(-k_1) + a_-(k_1)^\dagger g(-k_1)^\dagger\Big] \\ & + \tfrac{1}{\sqrt{2}}\hbar\omega_p \tfrac{\sqrt{\Omega_c}}{\sqrt{\omega_1}} \Big[a_+(-k_1)^\dagger g(-k_1) + a_+(-k_1) g(-k_1)^\dagger + a_-(-k_1) g(k_1) + a_-(-k_1)^\dagger g(k_1)^\dagger\Big] \\ & + \tfrac{1}{\sqrt{2}}\hbar\omega_p \tfrac{\sqrt{\Omega_c}}{\sqrt{\omega_2}} \Big[a_+(k_2)^\dagger g(k_2) + a_+(k_2) g(k_2)^\dagger + a_-(k_2) g(-k_2) + a_-(k_2)^\dagger g(-k_2)^\dagger\Big] \\ & + \tfrac{1}{\sqrt{2}}\hbar\omega_p \tfrac{\sqrt{\Omega_c}}{\sqrt{\omega_2}} \Big[a_+(-k_2)^\dagger g(-k_2) + a_+(-k_2) g(-k_2)^\dagger + a_-(-k_2) g(k_2) + a_-(-k_2)^\dagger g(k_2)^\dagger\Big]\end{aligned} \tag{6.72}$$

and the relevant 3-wave couplings (including only resonant RWA terms) are

$$\mathscr{H}_{afa} = \tfrac{1}{2\sqrt{2}}\hbar\omega_p \tfrac{\sqrt{\hbar\omega_p}ck_p}{\sqrt{m_e c^2}\sqrt{\omega_1\omega_2}} \tfrac{1}{\sqrt{N}} \Big[i\, a_+(k_2)^\dagger a_+(k_1) f(k_p)^\dagger - i\, a_+(k_2) a_+(k_1)^\dagger f(k_p)\Big]; \qquad \text{and} \tag{6.73}$$

$$\begin{aligned}\mathscr{H}_{afg} = {} & \tfrac{1}{2}\hbar\omega_p \tfrac{\sqrt{\hbar\Omega_c}ck_2}{\sqrt{m_e c^2}\sqrt{\omega_p\omega_2}} \tfrac{1}{\sqrt{N}} \Big[i\, a_+(k_2) f(k_p) g(k_1)^\dagger - i\, a_+(k_2)^\dagger f(k_p)^\dagger g(k_1)\Big] \\ & + \tfrac{1}{2}\hbar\omega_p \tfrac{\sqrt{\hbar\Omega_c}ck_1}{\sqrt{m_e c^2}\sqrt{\omega_p\omega_1}} \tfrac{1}{\sqrt{N}} \Big[i\, a_+(k_1) f(k_p)^\dagger g(k_2)^\dagger - i\, a_+(k_1)^\dagger f(k_p) g(k_2)\Big].\end{aligned} \tag{6.74}$$

This Hamiltonian is equivalent to a discrete set of harmonic oscillators (one for each of the probe photon, pump photon, probe-induced gyron, pump-induced gyron, and plasmon modes) with bi-linear and tri-linear couplings between them. At first glance, the explicit dependence of the 3-wave coupling coefficients on $1/\sqrt{N}$ (in which N is itself proportional to

the non-physical quantization volume $\mathcal{V} = \mathcal{A}L$) may appear puzzling, but with the quantum conventions used here the pump photon operator is normalized so that $a(k_2)^\dagger a(k_2)$ corresponds to the total number of (right-handed) pump photons of momentum $\hbar k_2$ throughout this quantization volume $\mathcal{V}$ (and similarly for the pump gyron operators), so non-physical parameters will in fact cancel.

It might also appear puzzling at first that no ponderomotive interaction between the EM and Langmuir waves appears, until we realize that his interaction will not appear at first-order in the Hamiltonian, unless we consider the continuum case to allow for finite bandwidths, and perform some sort of averaging over the fast time-scale. In our theory, ponderomotive effects will arise in the propagator at second-order (in a Dyson series) in the continuum-mode Hamiltonian, involving the creation of plasmons simultaneously with the virtual destruction of photons and gyrons at one frequency and re-creation at an infinitesimally lower frequency.

6.3.2 Implications of the Plasma EIT Hamiltonian: Modes and Dynamics

With the EIT model Hamiltonian in hand, we turn to solutions of the corresponding equations of motion, either in terms of a harmonic (normal mode) analysis, to determine the effect of the pump on the effective dispersion relation, or later with adiabatic envelope effects included, to model the entry and exit of the probe wave packet or slow changes in the amplitude of the control field.

To summarize, our strategy will be as follows: in order to continue making analytic progress, and because the pump will typically greatly exceed the prove field in intensity, the pump will be treated as a classical (c-number), coherent, prescribed control field. In the absence of the pump. we may work directly in the Heisenberg picture with the full Hamiltonian and full time-dependence embedded in the mode operators. In the presence of the pump, we switch to an interaction-like (Dirac) picture to remove the explicit fast time dependence (associated with the pump carrier frequency) from the Hamiltonian. We use many-DOF generalizations of the well-known Bogoliubov/Tyablikov (BT) transformations[275, 276, 277, 278, 279], which are unitary transformations associated with symplectic linear combinations of the original (undressed) annihilation and creation operators which preserve the CCRs, in order to diagonalize the relevant (full or interaction) Hamiltonian. The resulting eigenfrequencies determine the dressed dispersion relation, and the transformed, or dressed operators correspond to what we will call pseudo-modes of the

full coupled system. The eigenmode with zero (or in practice, small) eigenvalue of the interaction Hamiltonian, and with no (or small) admixture of the gyron mode (associated with resonant absorption in the absence of the pump) corresponds to the so-called dark-state polariton mode[280, 246, 239, 240, 281, 241, 282, 248, 253, 265], which can propagate within a new transparency window opened up the pump-induced EIT interaction. Other modes in the EIT plasma consist of linear combinations of a complementary bright-state polariton mode and the gyron mode that are symplectically-orthogonal to the dark-state polariton. The dark-state polariton mode can be accessed by appropriately slowly-varying probe and pump envelopes and "counterintuitive" turn-on of the pump prior to the arrival of the probe. All this will emerge more fully as we proceed, and various additional simplifying assumptions will be introduced to aid qualitative understanding without getting bogged down in mathematical detail.

Diagonalization without Pump: Recovery of Standard Dispersion Relation in Magnetized Plasma

To help verify our Hamiltonian model and assess the accuracy of the RWA, it will be useful to first examine the simple case with no applied pump, and therefore no EIT, in which all 3-wave coupling terms drop out of the relevant Hamiltonian, along with any terms with direct $\pm k_2$ wavenumber dependence. The plasma wave can no longer be ponderomotively excited from an initially quiescent plasma, but is included anyway for completeness:

$$\begin{aligned}\mathscr{H}_{\rm mp} &= \hbar\omega_1 a_+(k_1)^\dagger a_+(k_1) + \hbar\omega_1 a_-(-k_1)^\dagger a_-(-k_1) \\ &+ \hbar\omega_p f(k_p)^\dagger f(k_p) + \hbar\Omega_c g(k_1)^\dagger g(k_1) \\ &+ \tfrac{1}{\sqrt{2}}\hbar\omega_p \tfrac{\sqrt{\Omega_c}}{\sqrt{\omega_1}}\left[a_+(k_1)^\dagger g(k_1) + a_+(k_1) g(k_1)^\dagger + a_-(-k_1) g(k_1) + a_-(-k_1)^\dagger g(k_1)^\dagger\right].\end{aligned} \tag{6.75}$$

(The case without the probe and with only the pump field present is of course exactly analogous, but with k_1 replaced by k_2.) Obviously, longitudinal plasmons decouple from the transverse photons and gyrons in this limit, and the plasmon mode operators are unchanged. All bare fields here vary spatially as pure plane waves, but we now seek those linear combinations of photon and gyron fields which remain associated with a single wavenumber *magnitude* $k = |k_1|$, but also vary harmonically in time at a fixed frequency ω. Because these eigenfrequencies ω in the axial magnetized plasma should not depend on the sign of the wavenumber k, but only on its absolute value $|k|$ and whether the electric field vector (in a circular polarization basis) rotates resonantly or anti-resonantly with respect to the

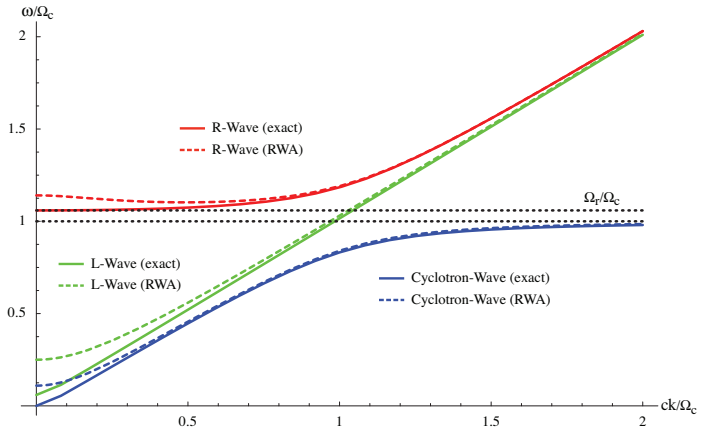

Figure 6.5. Dispersion relations for the transverse modes in an axially-magnetized, homogeneous, cold electron plasma with cyclotron frequency Ω_c and plasma frequency $\omega_p = 0.25\Omega_c$. The resonance at $\omega = \Omega_c$ is evident. Both the exact dispersion relations and Rotating Wave Approximations (RWA) are shown for comparison. The RWA L-polarized dispersion relations corresponds to the dispersion relation in the unmagnetized plasma. An additional longitudinal electrostatic mode at $\omega(k) = \omega_p$ is not shown.

gyrating electrons, the relevant characteristic polynomial will be a cubic in ω^2. In particular, keeping all terms (RWA and non-RWA) and performing the BT diagonalization procedure, the branches of the dispersion relation are determined as the roots of the characteristic equation

$$\Omega_c^2 \left(\omega^2 - c^2k^2\right)^2 - \omega^2 \left(\omega^2 - \omega_p^2 - c^2k^2\right)^2 = 0. \tag{6.76}$$

Choosing signs of the various square roots so as to ensure that frequencies on each branch are positive, continuous, and in this case even-parity functions of the wavenumber, this reproduces exactly the classical dispersion relation (6.4) for EM waves traveling parallel to a constant magnetic field in a cold electron plasma. Closed-form expressions can be written down for for the three distinct branches – an upper R-wave branch $\omega_{\rm R}(k)$, a counter-propagating, L-wave branch $\omega_{\rm L}(k)$, and the lower R-polarized cyclotron branch $\omega_{\rm C}(k)$ – but these are less less illuminating than a simple plot for a typical case, as in Fig. 6.5.

Classically, this dispersion relation is typically obtained using the cold Maxwell-fluid equations of motion, but it is reassuring to verify that it can also be reproduced using a Bogoliubov-type transformation of the collective-mode Hamiltonian – while perhaps not the simplest path to the EM dispersion relation, it enjoys a certain mathematical transparency in terms of diagonalizing an effective coupling matrix, and a direct physical interpretation in terms of dressed modes.

Somewhat less obviously, from the form of the Hamiltonian $\mathscr{H}_{\rm mp}$ containing just bare

photon, bare gyron, and (both RWA and non-RWA) photon-gyron couplings, for $k = k_1 > 0$, both the right-moving, R-wave mode and the right-moving, R-polarized cyclotron mode will include not only contributions from the positive-helicity, $k_1 > 0$ photon annihilation operator, but also coherent admixtures of the resonant gyron annihilation operator and the anti-resonant, left-moving ($-k_1 < 0$), photon creation operator. Specifically, the dressed-mode transverse annihilation operators may be written in terms of the eigenfrequencies as:

$$\chi_{\rm R}(k) = \frac{\frac{1}{\sqrt{2}}\omega_p\sqrt{\frac{\Omega_c}{\omega_0}}\left\{(\omega_{\rm R}+\omega_0)a_+(k) + (\omega_{\rm R}-\omega_0)a_-(-k)^\dagger\right\} + (\omega_{\rm R}^2-\omega_0^2)g(k)}{\sqrt{(\omega_{\rm R}^2-\omega_0^2)^2 + 2\omega_p^2\Omega_c\omega_{\rm R}}}; \quad (6.77a)$$

$$\chi_{\rm C}(k) = \frac{(\omega_0^2-\omega_{\rm C}^2)g(k) - \frac{1}{\sqrt{2}}\omega_p\sqrt{\frac{\Omega_c}{\omega_0}}\left\{(\omega_{\rm C}+\omega_0)a_+(k) + (\omega_{\rm C}-\omega_0)a_-(-k)^\dagger\right\}}{\sqrt{(\omega_{\rm C}^2-\omega_0^2)^2 + 2\omega_p^2\Omega_c\omega_{\rm C}}}; \quad (6.77b)$$

$$\chi_{\rm L}(-k) = \frac{\frac{1}{\sqrt{2}}\omega_p\sqrt{\frac{\Omega_c}{\omega_0}}\left\{(\omega_{\rm L}-\omega_0)a_+(k)^\dagger + (\omega_{\rm L}+\omega_0)a_-(-k)\right\} + (\omega_{\rm L}^2-\omega_0^2)g(k)^\dagger}{\sqrt{2\omega_p^2\Omega_c\omega_{\rm L} - (\omega_{\rm L}^2-\omega_0^2)^2}}. \quad (6.77c)$$

One can verify that these dressed operators satisfy equal-time CCRs just as the bare operators do. As $k \to \infty$, one can show that in this limit these dressed modes will coincide with the pure R-wave photon, gyron, and L-wave photon respectively, which explains the asymptotic behavior along their respective branches of the dispersion relation. As $\Omega_c \propto B_0 \to 0$, one can, with a little care, also show they reduce to the original uncoupled modes (in a different permutation, because of an avoided crossing of the eigenfrequencies). In terms of these eigenfrequencies and mode operators (and the still bare plasmon operators), the Hamiltonian (6.75) assumes the simple uncoupled form

$$\begin{aligned}\mathscr{H}_{\rm mp} = {} & \hbar\omega_{\rm R}(k_1)\chi_{\rm R}(k_1)^\dagger\chi_{\rm R}(k_1) + \hbar\omega_{\rm C}(k_1)\chi_{\rm C}(k_1)^\dagger\chi_{\rm C}(k_1) \\ & + \hbar\omega_{\rm L}(-k_1)\chi_{\rm L}(-k_1)^\dagger\chi_{\rm L}(-k_1) + \hbar\omega_p f(k_p)^\dagger f(k_p).\end{aligned} \quad (6.78)$$

The dressed modes clearly reveal what is sometimes obscured in textbook treatments – namely, that the true linear modes of the coupled field-and-magnetized-plasma system involve coherent superpositions of field and matter degrees-of-freedom, and moreover do so in ways that mix the left-moving and right-moving EM fields, and the bare creation and annihilation operators. The latter applies that the vacuum is dressed in addition to the modes: with photon-gyron coupling, the lowest energy state for collective excitations no longer corresponds to the absence of all bare quanta, and while the numbers of each kind of dressed quanta (i.e., the quantum actions) are separately conserved, bare transverse actions (i.e., undressed photon or gyron numbers) are not in general conserved.

This dressing of the vacuum may have important implications for quantum noise in the plasma, but because this mixing of creation and annihilation operators and vacuum dressing adds to the algebraic and physical complications, without really contributing to the dynamics essential to the EIT effect, it would be convenient to invoke the RWA and drop such terms, as long as the error is not too onerous. The resulting Hamiltonian,

$$\begin{aligned}\mathscr{H}_{\rm mp} \approx{}& \hbar\omega_1 a_+(k_1)^\dagger a_+(k_1) + \hbar\omega_1 a_+(-k_1)^\dagger a_+(-k_1) + \hbar\omega_p f(k_p)^\dagger f(k_p) \\ &+ \hbar\Omega_c g(k_1)^\dagger g(k_1) + \tfrac{1}{\sqrt{2}}\hbar\omega_p \tfrac{\sqrt{\Omega_c}}{\sqrt{\omega_1}}\Big[a_+(k_1)^\dagger g(k_1) + g(k_1)^\dagger a_+(k_1)\Big]\end{aligned} \tag{6.79}$$

consists only of products of creation with annihilation operators, and will separately conserve the total number (i.e., action) of plasmons and the numbers of bare L-polarized photons, as well as the sum of bare resonant gyrons and R-polarized photons, in addition to the number of each type of transverse dressed-mode quanta. Within the RWA, the L-wave remains unaltered (as does of course the uncoupled plasma wave), so that

$$\omega_{\rm L}(k) = \omega_0(k), \tag{6.80}$$

while the two R-polarized transverse waves involving in general both resonant photonic and gyronic excitation are associated with the coupled dispersion relation

$$(\omega_0(k) - \omega)(\Omega_c - \omega) = \tfrac{1}{2}\omega_p^2 \tfrac{\Omega_c}{\omega_0(k)} \equiv \Omega_{\rm G}(k)^2, \tag{6.81}$$

with solutions

$$\omega_{\rm R}(k) = \tfrac{1}{2}(\omega_0(k) + \Omega_c) + \tfrac{1}{2}\sqrt{(\omega_0(k) - \Omega_c)^2 + 4\Omega_{\rm G}(k)^2}; \tag{6.82a}$$

$$\omega_{\rm C}(k) = \tfrac{1}{2}(\omega_0(k) + \Omega_c) - \tfrac{1}{2}\sqrt{(\omega_0(k) - \Omega_c)^2 + 4\Omega_{\rm G}(k)^2}; \tag{6.82b}$$

where we have introduced the convenient definition

$$\Omega_{\rm G} = \Omega_{\rm G}(k) \equiv \tfrac{1}{\sqrt{2}}\omega_p \tfrac{\sqrt{\Omega_c}}{\sqrt{\omega_0(k)}} \geq 0 \tag{6.83}$$

for the gyro-photo-coupling frequency. The corresponding RWA dressed mode annihilation operators may be taken to be

$$\begin{aligned}\chi_{\rm R}(k) &= \frac{\Omega_{\rm G}(k)a_+(k) + (\omega_{\rm R}(k) - \omega_0(k))g(k)}{\sqrt{\Omega_{\rm G}(k)^2 + (\omega_{\rm R}(k) - \omega_0(k))^2}} \\ &= \cos\big(\theta_{ag}(k)\big)\, a_+(k) + \sin\big(\theta_{ag}(k)\big)\, g(k);\end{aligned} \tag{6.84a}$$

$$\begin{aligned}\chi_{\rm C}(k) &= \frac{\Omega_{\rm G}(k)g(k) - (\Omega_c - \omega_{\rm C}(k))a_+(k)}{\sqrt{\Omega_{\rm G}(k)^2 + (\Omega_c - \omega_{\rm C}(k))^2}} \\ &= \cos\big(\theta_{ag}(k)\big)\, g(k) - \sin\big(\theta_{ag}(k)\big)\, a_+(k);\end{aligned} \tag{6.84b}$$

$$\chi_{\rm L}(-k) = a_-(-k); \tag{6.84c}$$

where we have defined

$$\begin{aligned}\theta_{ag}(k) &\equiv \arctan\left[\frac{\omega_{\rm R}(k)-\omega_0(k)}{\Omega_{\rm G}(k)}\right] = \arctan\left[\frac{\Omega_c-\omega_{\rm C}(k)}{\Omega_{\rm G}(k)}\right] \\ &= \arctan\left[\frac{\Omega_{\rm G}(k)}{\omega_0(k)-\omega_{\rm C}(k)}\right].\end{aligned} \tag{6.85}$$

So in this approximation, the transverse dressed modes just correspond to orthogonal rotations of the bare modes in an abstract operator space by the mixing angle $\theta_{ag}(k)$, which increases with the coupling strength $\Omega_{\rm G}(k)$, or equivalently, with the magnitude $|\omega(k)-\omega_0(k)|$ of the deviation of the dressed from the bare eigenfrequencies. It is trivial to verify the CCRs for the RWA modes in this $\mathcal{SU}(2)$ form.

Here we have written the mixing angle formally in terms of the eigenfrequencies. It will also prove useful to invert these relations, expressing the mixing angle directly in terms of the known parameters, and then write the eigenfrequencies in terms of the mixing angle. After some algebra, we find

$$\theta_{ag}(k) = \arctan\left[\frac{\frac{1}{2}(\Omega_c-\omega_0(k)) + \frac{1}{2}\sqrt{(\Omega_c-\omega_0(k))^2+4\Omega_{\rm G}(k)^2}}{\Omega_{\rm G}(k)}\right], \tag{6.86}$$

and

$$\omega_{\rm R}(k) = \omega_0(k)\cos^2\theta_{ag}(k) + \Omega_c\sin^2\theta_{ag}(k) + 2\Omega_{\rm G}\sin\theta_{ag}(k)\cos\theta_{ag}(k); \tag{6.87a}$$

$$\omega_{\rm C}(k) = \omega_0(k)\sin^2\theta_{ag}(k) + \Omega_c\cos^2\theta_{ag}(k) - 2\Omega_{\rm G}\sin\theta_{ag}(k)\cos\theta_{ag}(k); \tag{6.87b}$$

so the eigenfrequencies are effectively weighted averages of the bare frequencies, weighted by the corresponding proportions of bare modes in the dressed modes, plus an additional coupling term.

For comparison, the RWA dispersion relation is also plotted along with the exact dispersion relation for the same parameters in Fig. 6.5. As might be expected, the high-frequency (short wavelength) asymptotic behavior is excellent, but the error increases with the wavelength, with behavior near the cutoffs rather poorly captured (and we already know the vacuum properties cannot be exactly right). As $k \to 0$, the upper R-wave frequency no longer approaches the cutoff precisely at ω_r, the cyclotron-wave frequency does not approach zero, and the L-wave dispersion relation just coincides with unmagnetized case, so the frequency approaches a cutoff at the unperturbed value of ω_p rather than at the value ω_ℓ appropriate to a magnetized plasma. However, in the context of EIT, the region of interest lies at intermediate values of k, where the transverse EM mode of the unmagnetized plasma passes through the stop band for the magnetized plasma. For typical parameters, the RWA

approximation remains reasonably accurate there, and so out of convenience will be used as we add the 3-wave contributions.

Because for typical parameters the pump-dependent 3-wave interaction energies will typically be comparable to or less than the pump-independent 3-wave coupling energies, we will also adopt the RWA for the Raman and Landau scattering terms, as is almost always done in classical plasma physics. For sufficiently large pump strengths $a_{\rm pump}$, the effects of the anti-resonant contributions will eventually become important, but by then our theory will already be invalidated by other neglected effects, such as relativistic detunings, plasma wave-breaking, and higher-order ("many-wave") but resonant scattering processes.

RWA EIT Pseudo-Modes

Even with just the resonant RWA 3-wave terms included, if the pump field were to be treated completely self-consistently as an independent dynamical variable, then the Hamiltonian would be cubic (and non-perturbatively so) in the annihilation and creation operators, and little further analytic progress could be made. Linear normal modes as such will not exist, and the meaning, if any, which might be attributed to dispersion relations becomes somewhat unclear. However, in regimes of interest exhibiting EIT, the pump field is generally much more intense than the probe wave, suggesting that a reasonable approximation might be to treat the pump EM wave and the pump-induced gyron wave as prescribed coherent classical fields (with photonic and gyronic components related by the RWA eigenmode forms), while continuing to describe the probe photon, probe-induced gyron, and ponderomotively-driven plasmon modes quantum-mechanically, with couplings parameterized by the applied pump strength. Essentially the same approximation is made almost universally in theoretical or numerical treatments of the atomic EIT case, and is commonly used as a first approximation in the classical analysis of Raman scattering when pump depletion effects can be neglected. Because the downshifted pump is R-polarized with respect to the magnetic field, we must account for both its photon-like and gyron-like components by ensuring that it lies on the appropriate lower R-polarized (cyclotron) transverse branch of the magnetized-plasma (RWA) dispersion relation. Within our quantum formalism and the RWA, treating the pump as a prescribed coherent classical field is equivalent to assuming the corresponding dressed cyclotron mode is excited in the corresponding Glauber coherent state, and additionally assuming that the dynamics of this coherent state are not determined self-consistently but rather are explicitly prescribed and satisfy the dispersion relation $\omega_2(k_2) = \omega_{\rm C}(k_2)$ formulated above. In essence, any feedback of the probe on the

pump is completely neglected. In the RWA, the relevant parametric EIT Hamiltonian then becomes bilinear in all dynamical quantum mode operators, with 3-wave coupling parameters associated with the Raman and Landau scattering terms tuned by the pump field, which is assumed given as a function of time and space (or frequency and wavenumber on the appropriate dispersion surface).

By definition, a normal mode normally refers to a fixed linear combination of the dynamical variables which can oscillate coherently at a fixed frequency (and therefore with fixed phase relationships between all constituent DOFs). While the dynamics of the parametric EIT Hamiltonian is now fully linear, because the pump has time-dependence at the frequency $\omega_2 \sim (\Omega_c - \omega_p)$, roughly comparable in magnitude to the bare frequencies of the dynamically-evolved bare modes, the system does not admit traditional normal modes *per se*. However, by shifting to an interaction-type (Dirac) picture, where the fast time-dependence is effectively stripped off, and diagonalizing the remaining interaction Hamiltonian using a BT transformation, we can identify effective interaction modes which do look like harmonic plane waves (still assuming a periodic or spatially-infinite medium). Back in the original Heisenberg picture, these correspond to fields patterns with reasonably simple time and space dependence, but where the transverse and longitudinal constituent DOFs oscillate harmonically at frequencies and vary spatially at wavenumbers which differ by exactly the frequency and wavenumber of the imposed pump, which beats with the bare probe or probe-induce gyron to make up this difference. Since the term "quasi-mode" usually refers to a "leaky" mode with some small imaginary frequency component (corresponding to slow damping or growth), we will instead call the dressed modes of the interaction Hamiltonian "pseudo-modes," particularly when transformed back into the Heisenberg representation so as to reacquire their full time-dependence. The effective dispersion relations for these pseudo-modes are then presented from the point-of-view of either the transverse or the longitudinal DOF, which differ by an offset just equal to the pump frequency.

In practice, the probe would typically be launched from some remote source outside the EIT medium with given frequency rather than wavenumber, but mathematically it is simpler to specify modes by wavenumber, and later invert the dispersion relation if necessary. That is, here we take the probe photon q-number field $a_1 \equiv a_+(k_1;t)$ to be a (spatial) plane wave, characterized in momentum space by the single given wavenumber k_1, while the corresponding probe-photon dressed frequency $\omega_1 = \omega_1(k_1)$ remains to be determined as an explicit function of k_1 and an implicit function of the pump strength, plasma density, and cyclotron frequency. Nearly in (at least spatial) resonance with the bare probe photon field

is a corresponding probe-induced bare gyron q-number field $g_1 \equiv g(k_1;t)$, with the usual bare cyclotron frequency Ω_c. In the context of EIT, we are of course most interested in cases where $\omega_1 \gtrsim \Omega_c$ lies in the stop-band, but the diagonalization procedure will reveal all three branches that can include admixtures of $a_1(k_1)$ or the corresponding probe-induced gyron field $g_1(k_1)$.

Interaction Hamiltonian

We next assume the prescribed pump is an harmonic cyclotron plane-wave given by the c-number photon and gyron fields

$$a_2 \equiv a_+(k_2;t) = \tilde{a}_2(k_2)\, e^{-i\omega_2 t}, \tag{6.88a}$$

$$g_2 \equiv g(k_1;t) \quad = \tilde{g}_2(k_2)\, e^{-i\omega_2 t}, \tag{6.88b}$$

for some particular values for the wavenumber k_2 and frequency $\omega_2 = \omega_{\rm C}(k_2)$ satisfying the RWA version of the dispersion relation for the unperturbed magnetized plasma along the lower R-polarized (cyclotron) branch. Within the RWA, the photon-like amplitude $\tilde{a}_2$ and gyron-like amplitude $\tilde{g}_2$ of the pump cyclotron-wave field are simply related by the mixing angle $\theta_{ag}(k_2)$. Inverting the RWA dressing relations (6.84) and then setting $\chi_{\rm R}(k_2) = 0$, we find

$$\tilde{g}(k_2) = -\tfrac{\cos\theta_{ag}(k_2)}{\sin\theta_{ag}(k_2)}\tilde{a}_2(k_2) = -\tfrac{1}{\sqrt{2}}\tfrac{\sqrt{\Omega_c}}{\sqrt{\omega_0(k_2)}}\tfrac{\omega_p}{\Omega_c - \omega_{\rm C}(k_2)}\tilde{a}_2(k_2). \tag{6.89}$$

The pump frequency ω_2 will be remain arbitrary for now, but typically we will be interested in the two-quanta-resonant cases where the pump frequency is either downshifted from from the bare cyclotron frequency by the plasma frequency, i.e., $\omega_2 = \Omega_c - \omega_p$, or else self-consistently downshifted from the dressed probe frequency by the linear plasma frequency, i.e., $\omega_2 = \omega_1 - \omega_p$. In the latter case, one will need to numerically solve an additional nonlinear equation after the dispersion relation for fixed ω_2 is obtained in order to determine this shift self-consistently. In the usual case where $\omega_1 \approx \Omega_c = \omega_2 + \omega_p$ and $\omega_p < \Omega_c$, note that $\frac{|\tilde{g}_2|^2}{|\tilde{a}_2|^2} \sim O(\frac{1}{2})$, so the gyration component of the (classical) pump action should not be neglected even though the frequency is downshifted by some not-insignificant amount ($O(25\%)$) below the cyclotron resonance.

For the ponderomotively-excited plasmon mode $f = f(k_p;t)$, the wavenumber $k_p = k_1 - k_2$ is then chosen to be in exact resonance with the beat wave between pump and probe, while the bare frequency is of course still chosen to be ω_p independent of k_p, consistent with cold plasma dynamics.

For this plane-wave case, the parametric RWA EIT Hamiltonian may be compactly written, apart from a trivial constant, as

$$\mathcal{H}_{\rm EIT} = \mathcal{H}_0 + \mathcal{H}_{\rm G} + \mathcal{H}_{\rm R} + \mathcal{H}_{\rm L'} + \mathcal{H}_{\rm L}, \tag{6.90}$$

where

$$\mathcal{H}_0 = \hbar\omega_0(k_1)\, a_1^\dagger a_1 + \hbar\Omega_c\, g_1^\dagger g_1 + \hbar\omega_p\, f^\dagger f; \tag{6.91a}$$

$$\mathcal{H}_{\rm G} = \hbar\Omega_{\rm G} a_1^\dagger g_1 + \hbar\Omega_{\rm G} a_1 g_1^\dagger; \tag{6.91b}$$

$$\mathcal{H}_{\rm R'} = \hbar\Omega_{\rm R'} e^{-i\omega_2 t} a_1^\dagger f + \hbar\Omega_{\rm R'}^* e^{+i\omega_2 t} a_1 f^\dagger; \tag{6.91c}$$

$$\mathcal{H}_{\rm L'} = \hbar\Omega_{\rm L'} e^{-i\omega_2 t} a_1^\dagger f + \hbar\Omega_{\rm L'}^* e^{+i\omega_2 t} a_1 f^\dagger; \tag{6.91d}$$

$$\mathcal{H}_{\rm L} = \hbar\Omega_{\rm L} e^{-i\omega_2 t} g_1^\dagger f + \hbar\Omega_{\rm L}^* e^{+i\omega_2 t} g_1 f^\dagger; \tag{6.91e}$$

where

$$\Omega_{\rm G} = \Omega_{\rm G}(k_1) = \tfrac{1}{\sqrt{2}}\omega_p\sqrt{\tfrac{\Omega_c}{\omega_0(k_1)}}; \tag{6.92}$$

is the real-valued 2-wave (probe photon-gyron) coupling constant defined previously; and where we have further introduced shorthand expressions for the various complex-valued, (Raman and Landau scattering) coupling coefficients, which are functions of the applied pump amplitude, and which we will therefore generically call Rabi frequencies in analogy with atomic physics. It is convenient to now return to a classical normalization for the pump field, whence:

$$\Omega_{\rm R'} = -i\tfrac{1}{2\sqrt{2}}\omega_p\sqrt{\tfrac{\hbar\omega_p}{m_e c^2}}\,\tfrac{ck_p}{\sqrt{\omega_0(k_1)\omega_0(k_2)}}\,\tfrac{1}{\sqrt{N}}\,\tilde{a}_2 = -i\tfrac{1}{2\sqrt{2}}ck_p\,\tfrac{\sqrt{\omega_p}}{\sqrt{\omega_0(k_1)}}\,\tilde{a}_{\rm pump}; \tag{6.93a}$$

$$\Omega_{\rm L'} = -i\tfrac{1}{2}\omega_p\sqrt{\tfrac{\hbar\Omega_c}{m_e c^2}}\,\tfrac{ck_1}{\sqrt{\omega_p\omega_0(k_1)}}\,\tfrac{1}{\sqrt{N}}\,\tilde{g}_2 \qquad = +i\tfrac{1}{2\sqrt{2}}ck_1\,\tfrac{\sqrt{\Omega_c}}{\sqrt{\omega_p}}\,\tfrac{\sqrt{\Omega_c}}{\sqrt{\omega_0(k_1)}}\,\tfrac{\omega_p}{\Omega_c-\omega_2}\,\tilde{a}_{\rm pump}; \tag{6.93b}$$

$$\Omega_{\rm L} = +i\tfrac{1}{2}\omega_p\sqrt{\tfrac{\hbar\Omega_c}{m_e c^2}}\,\tfrac{ck_2}{\sqrt{\omega_p\omega_0(k_2)}}\,\tfrac{1}{\sqrt{N}}\,\tilde{a}_2 \qquad = +i\tfrac{1}{2}ck_2\,\tfrac{\sqrt{\Omega_c}}{\sqrt{\omega_p}}\,\tilde{a}_{\rm pump}. \tag{6.93c}$$

Clearly, the 2-wave interactions in $\mathcal{H}_{\rm G}$ will be responsible for the cyclotron resonance, but only the 3-wave term $\mathcal{H}_{\rm L}$ directly couples the gyron, plasmon, and pump fields, and is therefore responsible for the "destructive interference" associated with the EIT. The interactions embodied in $\mathcal{H}_{\rm R'} + \mathcal{H}_{\rm L'}$ do not directly involve the gyron field, or in other words, they allow for transitions directly between Langmuir levels and do not pass – even virtually – through the Landau levels. Because they push the coupling in the direction of mixing the bare gyron and plasmon modes, rather than the probe photon and plasmon modes, these latter interactions can really only act to impede the establishment of a transparent dressed mode for which the effects of the magnetic field are in some sense canceled.

Roughly-speaking, we therefore expect the extent of pump-induced transparency – e.g., the width of the EIT pass-band, or the group velocity at the center of the band– to increase with increasing $|\Omega_{\rm L}|$, as compared to $\Omega_{\rm G}$, while deviations from exact "cancellation" of the resonance will become more apparent as $|\Omega_{\rm L'} + \Omega_{\rm R'}|$ increases relative to $\Omega_{\rm L}$.

For reasonable near-resonant parameters, where $\omega_1 \sim \Omega_c \sim \omega_2 + \omega_p$, $O(10^{-1}) \lesssim \frac{\omega_p}{\Omega_c} < O(1)$, and $|a_{\rm pump}| \lesssim O(\frac{1}{2})$, the ratio

$$\frac{|\Omega_{\rm L}|}{\Omega_{\rm G}} = \frac{1}{\sqrt{2}} \frac{ck_2}{\omega_p} \frac{\sqrt{\omega_0(k_1)}}{\sqrt{\omega_p}} |a_{\rm pump}| \sim \frac{1}{\sqrt{2}} \left(\frac{\Omega_c}{\omega_p}\right)^{3/2} |a_{\rm pump}| \tag{6.94}$$

is $O(1)$ or maybe a little larger, suggesting both that EIT can be achieved at these pump strengths and that it may be described reasonably well within the RWA. The relative strengths of the different 3-wave processes are roughly

$$\frac{\Omega_{\rm R'}}{\Omega_{\rm L}} = -\frac{1}{\sqrt{2}} \frac{k_p}{k_2} \frac{\omega_p}{\sqrt{\Omega_c \omega_0(k_1)}} \sim \begin{cases} -\frac{1}{\sqrt{2}} \frac{\omega_p}{\Omega_C - \omega_p} \frac{\omega_p}{\Omega_c} & \text{if } k_2 > 0 \text{ (co-propagating);} \\ -\frac{1}{\sqrt{2}} \frac{2\Omega_c - \omega_p}{\Omega_c - \omega_p} \frac{\omega_p}{\Omega_c} & \text{if } k_2 < 0 \text{ (counter-propagating);} \end{cases} \tag{6.95}$$

so in either geometry the effects of the Raman scattering should be somewhat smaller than those associated with the gyron-plasmon Landau scattering, but the Raman-scattering term is not really so small as to be totally negligible. If $\mathcal{H}_{\rm L'}$ were not also present, then it would be preferable to use co-propagating geometry for the pump and probe, since the undesirable Raman scattering effects are then smaller. However,

$$\frac{\Omega_{\rm L'}}{\Omega_{\rm L}} = \frac{1}{\sqrt{2}} \frac{k_1}{k_2} \frac{\sqrt{\Omega_c}}{\sqrt{\omega_0(k_1)}} \frac{\omega_p}{\Omega_c - \omega_p} \sim \frac{1}{\sqrt{2}} \frac{\Omega_c}{\Omega_c - \omega_p}, \tag{6.96}$$

so these terms are of comparable magnitude, which is mildly unfortunate, but the undesirable term is apparently unavoidable as it arises from the same type of interactions in the Hamiltonian as the transparency-inducing term. Finally,

$$\frac{\Omega_{\rm R'}}{\Omega_{\rm L'}} = -\frac{k_p}{k_1} \frac{\omega_p}{\Omega_c} \frac{\Omega_c - \omega_2}{\omega_p} \sim \begin{cases} -\frac{\omega_p^2}{\Omega_c^2} & \text{if } k_2 > 0 \text{ (co-propagating);} \\ -\frac{2\Omega_c - \omega_p}{\Omega_c} \frac{\omega_p}{\Omega_c} & \text{if } k_2 < 0 \text{ (counter-propagating);} \end{cases} \tag{6.97}$$

so the effects of the plasmon-probe photon Landau scattering term (i.e., that involving the pump gyron field) dominate the true Raman scattering in the co-propagating case, while these are more comparable but still typically larger in the counter-propagating case. The coupling constants are typically of opposite sign, so at least these deleterious scattering effects tend to partially cancel each other. In order to minimize the undesirable interactions, we will therefore assume counter-propagating geometry unless otherwise specified (This

geometry also makes the spectral separation of pump and probe fields simpler). The effects of both $\mathcal{H}_{\mathrm{R}'}$ and $\mathcal{H}_{\mathrm{L}'}$ can be combined into what looks like ordinary Raman scattering but with a renormalized photon-plasmon coupling constant:

$$\mathcal{H}_{\mathrm{R}} \equiv \mathcal{H}_{\mathrm{R}} + \mathcal{H}_{\mathrm{L}'} = \hbar\Omega_{\mathrm{R}} e^{-i\omega_2 t} a_1^\dagger f + \hbar\Omega_{\mathrm{R}}^* e^{+i\omega_2 t} a_1 f^\dagger, \tag{6.98}$$

where

$$\Omega_{\mathrm{R}} \equiv \Omega_{\mathrm{R}'} + \Omega_{\mathrm{L}'} = +i\tfrac{1}{2\sqrt{2}} c k_1 \tfrac{\Omega_c}{\sqrt{\omega_p \omega_0(k_1)}} \left(\tfrac{\omega_p}{\Omega_c - \omega_2} - \tfrac{\omega_p}{\Omega_c} \tfrac{k_p}{k_1} \right) \tilde{a}_2. \tag{6.99}$$

For the sake of simplicity, we henceforth refer to this combined term as the (renormalized) Raman scattering interaction, despite its origin in distinct Raman scattering and Landau scattering effects.

Next, we make a unitary transformation to an interaction-type (Dirac) picture in order to eliminate the fast explicit time-dependence in the 3-wave terms in the Hamiltonian. First we define the generator

$$\tilde{\mathcal{H}} = \hbar\Omega_c (a_1^\dagger a_1 + g_1^\dagger g_1) + \hbar\tilde{\omega}_p f^\dagger f, \tag{6.100}$$

and the corresponding unitary operator

$$U = e^{-\frac{i}{\hbar} t \tilde{\mathcal{H}}}, \tag{6.101}$$

in which $\tilde{\omega}_p \equiv \Omega_c - \omega_2$ is the pump downshift with respect to the Larmor frequency. The operator $\tilde{\mathcal{H}}$ is therefore close to the uncoupled Hamiltonian $\mathcal{H}_0$ in the resonant case where $\omega_1 \approx \Omega_c \approx \omega_2 + \omega_p$. Using this unitary transformation, we can define the slowly-varying interaction operators $\tilde{\chi} = U\chi U^\dagger$, where χ is any function of the bare operators a_1, g_1, and f, or their adjoints. In particular, we are interested in

$$\tilde{a}_1 = U a_1 U^\dagger = e_0^{+i\Omega_c t} a_1; \tag{6.102a}$$

$$\tilde{g}_1 = U g_1 U^\dagger = e^{+i\Omega_c t} g_1; \tag{6.102b}$$

$$\tilde{f} = U f U^\dagger = e^{+i\tilde{\omega}_p t} f, \tag{6.102c}$$

and their Hermitian adjoints, which in effect strip off most of the fast time dependence.

In the absence of any *explicit* time dependence in the Heisenberg operator χ (i.e., without additional dependence beyond that generated by $\mathcal{H}_{\mathrm{EIT}}$) a little algebra reveals that such interaction operators will evolve according to the equations of motion

$$i\hbar \tfrac{\partial}{\partial t} \tilde{\chi} = [\tilde{\chi}, \mathcal{H}_{\mathrm{int}}] \tag{6.103}$$

in terms of the interaction Hamiltonian

$$\begin{aligned}\mathcal{H}_{\rm int} = \mathcal{H}_{\rm EIT} - \tilde{\mathcal{H}} &= \hbar(\omega_0(k_1) - \Omega_c)\tilde{a}_1^\dagger \tilde{a}_1 + \hbar(\omega_p - \tilde{\omega}_p)\tilde{f}^\dagger \tilde{f} \\ &+ \hbar\Omega_{\rm G}\tilde{a}_1^\dagger \tilde{g}_1 + \hbar\Omega_{\rm G}\tilde{a}_1 \tilde{g}_1^\dagger \\ &+ \hbar\Omega_{\rm R}\tilde{a}_1^\dagger \tilde{f} + \hbar\Omega_{\rm R}^* \tilde{a}_1 \tilde{f}^\dagger + \hbar\Omega_{\rm L}\tilde{g}_1^\dagger \tilde{f} + \hbar\Omega_{\rm L}^* \tilde{g}_1 \tilde{f}^\dagger,\end{aligned} \tag{6.104}$$

where all couplings are now independent of time. This may be conveniently written as a quadratic form:

$$\begin{aligned}\mathcal{H}_{\rm int} &= \begin{bmatrix}\tilde{a}_1^\dagger & \tilde{f}^\dagger & \tilde{g}_1^\dagger\end{bmatrix} K \begin{bmatrix}\tilde{a}_1 \\ \tilde{f} \\ \tilde{g}_1\end{bmatrix} \\ &= \begin{bmatrix}\tilde{a}_1^\dagger & \tilde{f}^\dagger & \tilde{g}_1^\dagger\end{bmatrix} \begin{bmatrix}(\omega_0(k_1) - \Omega_c) & \Omega_{\rm R}^* & \Omega_{\rm G} \\ \Omega_{\rm R} & (\omega_p - \tilde{\omega}_p) & \Omega_{\rm L} \\ \Omega_{\rm G} & \Omega_{\rm L}^* & 0\end{bmatrix} \begin{bmatrix}\tilde{a}_1 \\ \tilde{f} \\ \tilde{g}_1\end{bmatrix}\end{aligned} \tag{6.105}$$

in terms of the indicated interaction-picture Hermitian c-number coupling matrix $K = K(k_1; k_2; \omega_2; \omega_p; \Omega_c; \tilde{a}_2)$.

Diagonalization

Now we effect an BT transformation[275, 276, 277, 278, 279] on this interaction Hamiltonian to determine the dressed interaction eigenfrequencies and eigenmodes. Within the RWA, we know that the dressed annihilation operators will be linear combinations of the form $\tilde{\chi} = \tilde{\alpha}\,\tilde{a}_1 + \tilde{\beta}\,\tilde{f} + \tilde{\gamma}\,\tilde{g}_1$, for c-number coefficients $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\gamma}$, satisfying

$$K \begin{bmatrix}\tilde{\alpha} \\ \tilde{\beta} \\ \tilde{\gamma}\end{bmatrix} = \tilde{\omega} \begin{bmatrix}\tilde{\alpha} \\ \tilde{\beta} \\ \tilde{\gamma}\end{bmatrix} \tag{6.106a}$$

$$|\tilde{\alpha}|^2 + \left|\tilde{\beta}\right|^2 + |\tilde{\gamma}|^2 = 1 \tag{6.106b}$$

for a particular real-valued interaction eigenvalue $\tilde{\omega} = \tilde{\omega}(k_1)$. Because of the coupling between transverse and longitudinal modes, we will now have three rather than two dressed branches involving the bare probe photon mode, probe-induced gyron mode, and ponderomotively-driven plasmon mode, the middle dispersion branch involving a new transparency window that emerges in a neighborhood of what, in the absence of the pump, was the bare cyclotron resonance.

We have in a sense made a somewhat long mathematical journey to recover a reassuringly familiar result; although we adopted an approach based on dressing of collective

excitation operators rather than quantum states of individual systems, determination of the dispersion relation is now seen to involve diagonalization of a 3×3 Hermitian matrix, just as in the case of non-interacting Λ-level atoms. In fact, if we were to neglect the residual detuning terms (remaining diagonal terms) and the renormalized Raman scattering terms compared to the Landau scattering terms, this matrix would be of *exactly* the same algebraic form as that arising in the usual EIT analysis of a single atom (when spontaneous emission effects are neglected), although the expansion coefficients here occur in the context of a Lie algebra of annihilation operators for collective field and matter excitations, rather than a Hilbert space of electronic states of an individual atom.

The dressed interaction-picture eigenfrequencies, as functions of the mode wavenumbers and the pump strength and frequency, are determined as the roots of the characteristic polynomial:

$$\begin{aligned}
&\tilde{\omega}^3 - \tilde{\omega}^2 \left[(\omega_0(k_1) - \Omega_c) + (\omega_p - \tilde{\omega}_p) \right] \\
&\quad - \tilde{\omega} \left[\Omega_{\mathrm{G}}^2 + |\Omega_{\mathrm{L}}|^2 + |\Omega_{\mathrm{R}}|^2 - (\omega_0(k_1) - \Omega_c)(\omega_p - \tilde{\omega}_p) \right] \\
&\quad + \left[(\omega_0(k_1) - \Omega_c) |\Omega_{\mathrm{L}}|^2 + (\omega_p - \tilde{\omega}_p)\Omega_{\mathrm{G}}^2 - \Omega_{\mathrm{G}}(\Omega_{\mathrm{L}}^* \Omega_{\mathrm{R}} + \Omega_{\mathrm{R}}^* \Omega_{\mathrm{L}}) \right] = 0;
\end{aligned} \tag{6.107}$$

and the corresponding interaction eigenmodes may be written in terms of these interaction eigenfrequencies:

$$\tilde{\chi}_\omega = \frac{\left[\tilde{\omega}^2 - \tilde{\omega} \left(\omega_p - \tilde{\omega}_p \right) - |\Omega_{\mathrm{L}}|^2 \right] a_1 + \left[\tilde{\omega} \Omega_{\mathrm{R}} + \Omega_{\mathrm{G}} \Omega_{\mathrm{L}} \right] f + \left[\Omega_{\mathrm{L}}^* \Omega_{\mathrm{R}} + \left(\tilde{\omega} - \omega_p + \tilde{\omega}_p \right) \Omega_{\mathrm{G}} \right] g_1}{\left[\left| \tilde{\omega}^2 - \tilde{\omega} \left(\omega_p - \tilde{\omega}_p \right) - |\Omega_{\mathrm{L}}|^2 \right|^2 + |\tilde{\omega} \Omega_{\mathrm{R}} + \Omega_{\mathrm{G}} \Omega_{\mathrm{L}}|^2 + |\Omega_{\mathrm{L}}^* \Omega_{\mathrm{R}} + \left(\tilde{\omega} - \omega_p + \tilde{\omega}_p \right) \Omega_{\mathrm{G}}|^2 \right]^{1/2}}. \tag{6.108}$$

Because the K matrix is Hermitian (damping having been neglected), the interaction eigenfrequencies $\tilde{\omega}$ are necessarily real. Also Ω_{G} and both $\Omega_{\mathrm{R}}/\Omega_{\mathrm{L}}$ and $\Omega_{\mathrm{L}}^* \Omega_{\mathrm{R}}$ are always real, so the expansion coefficients for both the probe photon and gyron contributions can be (and have been) taken to be real, while the phase of the complex plasmon coefficient is then just determined by $\arg[i\,\tilde{a}_2]$. Symplectic orthogonality of the dressed modes for distinct eigenvalues $\tilde{\omega}$ is not obvious, but is guaranteed by Hermiticity and the resulting form of the dispersion relation. Enforcing continuity of the interaction frequency as a function of wavenumber, the eigenvalues may be separated into three distinct branches:

$$\tilde{\omega}_c(k_1) < \tilde{\omega}_d(k_1) < \tilde{\omega}_r(k_1) \tag{6.109}$$

as explicit functions of the probe (or gyron) wavenumber k_1, and also implicit functions of the pump strength $\tilde{a}_2$ and pump frequency ω_2 and/or wavenumber k_2. In the interaction picture, K is not positive definite, and these frequencies are not in general non-negative.

As long as $\Omega_c \geq \omega_p > 0$, for fixed a_{pump} these interaction frequencies never intersect for any value of k_1 (as a consequence of the avoided crossing theorem), and in fact, generically we find $\tilde{\omega}_c(k_1) < 0$ and $\tilde{\omega}_r(k_1) > 0$, while $\tilde{\omega}_d(k_1) = 0$ for some particular value of k_1. As the dispersion relation is only cubic, explicit expressions can be written for these eigenvalues, but are somewhat long and not especially illuminating, so are omitted.

The dressed interaction eigenmodes oscillate harmonically at these eigenfrequencies. Restoring full time dependence using the inverse unitary transformation $U(t)^{-1} = U(t)^\dagger$ to map back to the complete Heisenberg representation, each of the three pseudo-modes $\chi_\nu = U(t)^\dagger \tilde{\chi}_\nu U(t)$, $\nu = c$, d, r, in general includes transverse (bare photon and/or gyron) contributions oscillating at wavenumber k_1 and frequency $\tilde{\omega}_\nu(k_1)+\Omega_c$, as well as longitudinal (bare plasmon) components at wavenumber $k_p = k_1 - k_2$ and frequency $\tilde{\omega}_\nu(k_1) + \tilde{\omega}_p = \omega_\nu(k_1) + \Omega_c - \omega_2$. That is, the transverse and longitudinal components of the pseudo-mode oscillate at frequencies differing precisely by the pump frequency ω_2, and vary spatially at wavenumbers differing precisely by the pump wavenumber k_2, as a result of beating with the pump field, which makes up the difference. So in order to recover an effective dressed dispersion relation from the point-of-view of the transverse photon mode or gyron mode, we just add back the frequency offset Ω_c to the interaction-picture frequencies, to obtain effective dressed pseudo-mode dispersion relations for the transverse components:

$$\omega_c(k_1) = \tilde{\omega}_c(k_1) + \Omega_c < \omega_d(k_1) = \tilde{\omega}_d(k_1) + \Omega_c < \omega_r(k_1) = \tilde{\omega}_r(k_1) + \Omega_c. \tag{6.110}$$

(Similarly, we could add back the frequency $\tilde{\omega}_p = \Omega_c - \omega_2$ to obtain effective pseudo-mode dispersion relations from the longitudinal plasmon perspective.)

If, as we will often assume, the pump is detuned such that $\omega_2 = \omega_c - \omega_p$ (or equivalently $\tilde{\omega}_p = \omega_p$) exactly, the characteristic polynomial simplifies considerably:

$$\begin{aligned} &\left(\tfrac{\tilde{\omega}}{\Omega_{\text{G}}}\right)^3 - \left(\tfrac{\tilde{\omega}}{\Omega_{\text{G}}}\right)^2 \tfrac{(\omega_0(k_1)-\Omega_c)}{\Omega_{\text{G}}} - \left(\tfrac{\tilde{\omega}}{\Omega_{\text{G}}}\right)\left[1 + \left(\tfrac{|\Omega_{\text{L}}|}{\Omega_{\text{G}}}\right)^2 \left\{1 + \left|\tfrac{\Omega_{\text{R}}}{\Omega_{\text{L}}}\right|^2\right\}\right] \\ &\qquad + \left(\tfrac{|\Omega_{\text{L}}|}{\Omega_{\text{G}}}\right)^2 \left[\tfrac{(\omega_0(k_1)-\Omega_c)}{\Omega_{\text{G}}} - 2\left(\tfrac{\Omega_{\text{R}}}{\Omega_{\text{L}}}\right)\right] = 0; \end{aligned} \tag{6.111}$$

depending on just the positive coupling frequency Ω_{G}, which is pump-independent, and three dimensionless real parameters: the relative detuning $\frac{(\omega_0(k_1)-\Omega_c)}{\Omega_{\text{G}}}$ from the unmagnetized dispersion relation, which is also independent of the pump amplitude; the relative strength of Landau scattering, $\frac{|\Omega_{\text{L}}|}{\Omega_{\text{G}}}$, which is proportional to the magnitude of the pump amplitude (but independent of its phase); and the (signed) relative strength of (renormalized) Raman versus Landau scattering, $\frac{\Omega_{\text{R}}}{\Omega_{\text{L}}}$, which is independent of pump strength as well. The dressed

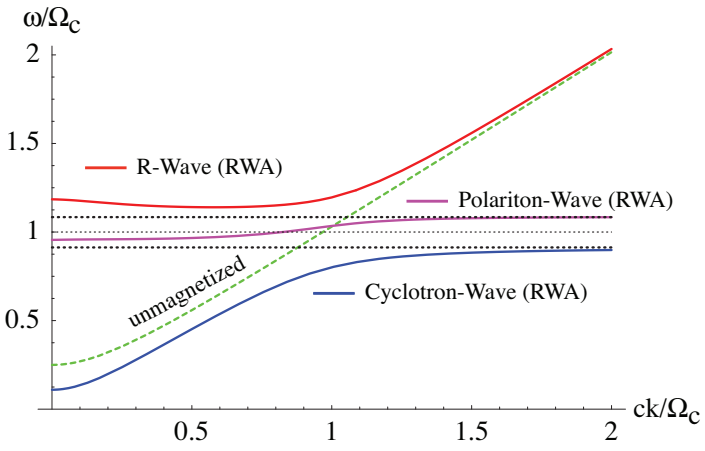

Figure 6.6. Effective dispersion relations for the transverse components of the dressed pseudo-modes in the presence of a plane-wave pump field of magnitude $|a_{\text{pump}}| = 0.1$ and frequency $\omega_2 = \Omega_c - \omega_p$, where $\omega_p = 0.25\Omega_c$. The split resonances are also indicated, and the EM dispersion relation for the unmagnetized plasma wave (or the L-polarized wave in the RWA) is shown for comparison.

eigenmodes simplify to

$$\tilde{\chi}_\omega = \frac{\left[\left(\frac{\tilde{\omega}}{\Omega_G}\right)^2 - \left(\frac{|\Omega_L|}{\Omega_G}\right)^2\right] a_1 + \frac{\Omega_L}{|\Omega_L|}\frac{|\Omega_L|}{\Omega_G}\left[\frac{\tilde{\omega}}{\Omega_G}\frac{\Omega_R}{\Omega_L} + 1\right] f + \left[\left(\frac{|\Omega_L|}{\Omega_G}\right)^2 \frac{\Omega_R}{\Omega_L} + \frac{\tilde{\omega}}{\Omega_G}\right] g_1}{\left[\left|\left(\frac{\tilde{\omega}}{\Omega_G}\right)^2 - \left(\frac{|\Omega_L|}{\Omega_G}\right)^2\right|^2 + \left|\frac{|\Omega_L|}{\Omega_G}\right|^2 \left|\frac{\tilde{\omega}}{\Omega_G}\frac{\Omega_R}{\Omega_L} + 1\right|^2 + \left|\left(\frac{|\Omega_L|}{\Omega_G}\right)^2 \frac{\Omega_R}{\Omega_L} + \frac{\tilde{\omega}}{\Omega_G}\right|^2\right]^{1/2}}, \tag{6.112}$$

and depend on the same parameters as the eigenvalues as well as one additional phase, $\frac{\Omega_L}{|\Omega_L|}$.

Eigenfrequencies

The effective dispersion relations inside the magnetized plasma in the presence of the pump are shown for typical parameters in Fig. 6.6. Behavior as $k \to 0$ is probably an artifact of the RWA, so should not be taken too seriously, but generally speaking all the features that we expect from EIT are reproduced. As suggested by the notation, the uppermost branch behaves similarly to the upper R-polarized (optical) branch in the absence of the pump, asymptotically approaching the vacuum dispersion relation $\omega = ck$ as $k \to \infty$. The lower branch behaves similarly to the lower R-polarized (cyclotron wave) branch in the absence of the pump, except that the resonance has been lowered below Ω_c, making room for a new branch with generally smaller slope (lower group velocity) within a narrow transparency band approximately centered on the cyclotron frequency Ω_c. In analogy with

condensed matter physics, this branch is referred to as a polariton, because it mixes both transverse EM field DOFs and certain material DOFs that couple dielectrically to the field. In the atomic EIT literature, the analogous excitation is referred to specifically as the Dark-State Polariton (DSP) mode, terminology we will borrow, because it corresponds to a small (or, ideally, vanishing) interaction eigenfrequency. That is, the excitation is *not* "dark" in the sense of having no probe photon content, but on the contrary because it largely de-couples from the gyron mode, and therefore looks dark from the perspective of the transverse electron DOFs, despite having an appreciable photon-like contribution at or near the cyclotron frequency.

In contrast to predictions for the atomic case, it is clear from the plot that the effects of the magnetic field on the refractive index are not exactly canceled, in that the DSP branch $\omega_d(k_1)$ crosses the bare resonance $\omega = \Omega_c$ at a wavenumber $\tilde{k}_1$ below that[6] where $\omega_0(k_1) = \Omega_c$. In [255], the authors conjectured that this shift is due to the participation of multiple Landau levels in the magnetized plasma case, compared to a single excited level in the atomic system, but this appears to be something of a red herring, as we will see more clearly later. Instead, we see that this shift arises out of the competing Raman scattering interactions between bare photons and plasmons, interactions which do not couple to the Landau levels at all, and therefore cannot contribute to the destructive interference that suppresses their population. As a result, the effects of the cyclotron resonance are not canceled precisely at the two-quanta resonance $\omega_1 = \Omega_c = \omega_2 + \omega_p$, as would happen in the absence of any Raman scattering. This can be seen graphically by plotting the effective dispersion relations for the same parameters as above, but with the deleterious Raman scattering effects artificially suppressed. Without such scattering, we see in Figure 6.7 that the passband lies centered on Ω_c, and the Raman-free DSP dispersion relation, denoted by $\bar{\omega}_d(k_1)$, satisfies $\bar{\omega}(k_0) = \omega_0(k_0) = \Omega_c$ at $k_0 \equiv c^{-1}\left[\Omega_c^2 - \omega_p^2\right]^{1/2}$. Using (6.111), but supposing contrary to fact that $\Omega_{\mathrm{R}} = 0$, we verify analytically that the Raman-free DSP dispersion relation would satisfy $\bar{\omega}_d(k_1) = \Omega_c$ exactly.

The crossover $\omega_d(k_1) = \Omega_c$ occurs at the zero of the DSP interaction eigenvalue: $\tilde{\omega}_d(\tilde{k}_0) = 0$. Numerical evidence suggests that the shift in crossing wavenumber is only very weakly dependent on a_{pump}, if at all, but that it is proportional to the relative strength of the Raman scattering compared to the Landau scattering, (which is independent of the pump strength). With a little Taylor expansion and implicit differentiation, this crossover

[6]The shift will be of larger magnitude and in the opposite sense for co-propagating pulses.

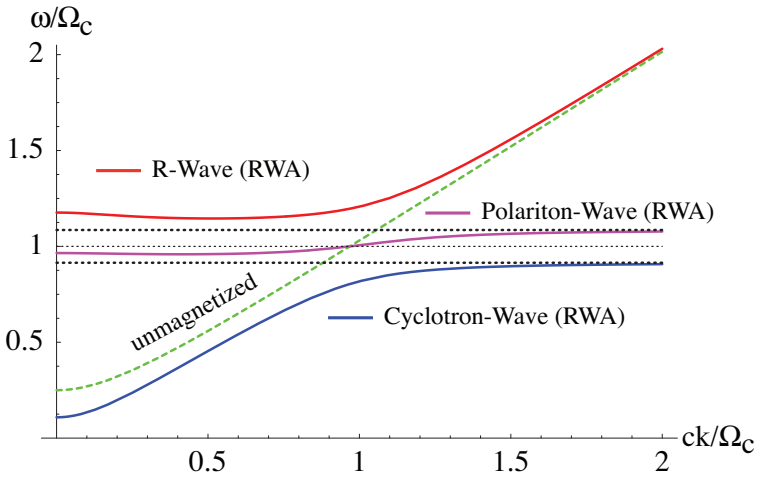

Figure 6.7. Effective dispersion relations for the same parameters as in Figure 6.6, but with the Raman scattering terms artificially suppressed. Here the index-of-refraction exactly matches that of the unmagnetized plasma exactly at the cyclotron resonance $\omega = \Omega_c$.

can be roughly estimated to occur at

$$\begin{aligned}\tilde{k}_0 &\approx k_0 - \left(\frac{\partial \omega_d'}{\partial k_1}\right)^{-1} \frac{\partial \omega_d}{\partial \Omega_{\mathrm{R}}}\Big|_{\Omega_{\mathrm{R}}=0} \Omega_{\mathrm{R}} + \dots \\ &= k_0 + 2\frac{\Omega_c}{c k_0}\left[\frac{\Omega_{\mathrm{G}}}{c}\frac{\Omega_{\mathrm{R}}}{\Omega_{\mathrm{L}}}\right]_{k_1=k_0} + O\left(\left|\frac{\Omega_{\mathrm{R}}}{\Omega_{\mathrm{L}}}\right|^2\right) + \dots .\end{aligned} \tag{6.113}$$

which seems to give reasonably accurate ($O(5\%)$ or better) estimates for typical parameters. While this approximation is rough, it can lead to such useful simplifications that we will repeatedly indulge in it.

Note that the EIT effect does not so much remove the cyclotron resonance as split it and move it away from Ω_c, opening the dressed polariton pass-band in between. In the presence of a fixed pump of amplitude a_{pump} and wavenumber k_2, these resonances are, to leading order in the pump strength, just offset from the bare cyclotron frequency by the Landau-scattering Rabi frequency:

$$\lim_{k_1\to\infty} \omega_c(k_1) = (\Omega_c - |\Omega_{\mathrm{L}}|) + O(\Omega_{\mathrm{L}}^*\Omega_{\mathrm{R}}) + O(|\Omega_{\mathrm{R}}|^2); \tag{6.114a}$$

$$\lim_{k_1\to\infty} \omega_d(k_1) = (\Omega_c + |\Omega_{\mathrm{L}}|) + O(\Omega_{\mathrm{L}}^*\Omega_{\mathrm{R}}) + O(|\Omega_{\mathrm{R}}|^2). \tag{6.114b}$$

The EIT pass-band is, to leading order in pump amplitude, centered on Ω_c, and its width grows (approximately linearly) with $|a_{\mathrm{pump}}|$ through its dependence on $|\Omega_{\mathrm{L}}|$, as expected. In fact, one can show that the errors in these estimated resonances are always proportional to (some power of) Ω_{R} so the limits (6.114) would hold exactly in the absence of the Raman scattering. The higher-order corrections due to Raman scattering introduce a slight asymmetry, but numerical evidence suggest this dependence is weak for typical parameters.

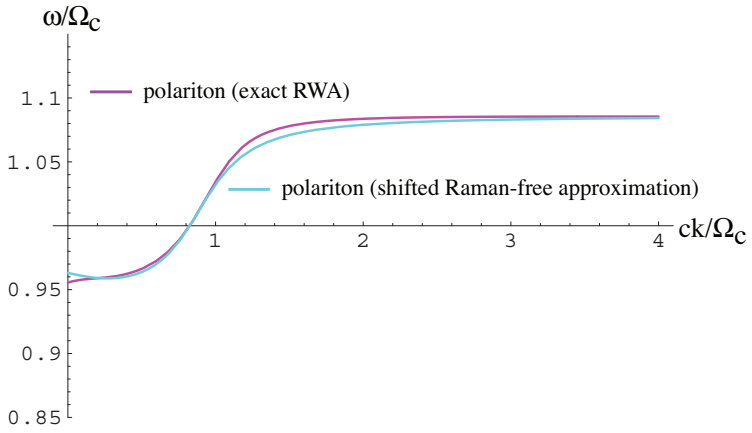

Figure 6.8. Comparison of the effective dispersion relation for the Polariton Mode to that estimated from the shifted-but-Raman-free (SRF) approximation, for pump magnitude $|a_{\rm pump}| = 0.1$ and frequency $\omega_2 = \Omega_c - \omega_p$, where $\omega_p = 0.25\Omega_c$.

Because the effect actually occurs at zeroth order in $a_{\rm pump}$, the most salient consequence of the Raman scattering term on the polariton dispersion relation for typical parameters (at least for $\omega_2 \approx \Omega_c - \omega_p$) is just the shift of the crossover point $\omega_d = \Omega_c$ described above. Since the locations of the resonances are only very weakly affected, the net effect is that $\omega_d(k_1)$ with Raman scattering effects included looks roughly like an overall horizontal translation of the dispersion curve $\omega'_d(k_1)$ in the absence of Raman scattering:

$$\omega_d(k_1) \approx \bar{\omega}_d(\bar{k}_1) + \ldots = \bar{\omega}_d\left(k_1 + (k_0 - \tilde{k}_0)\right) + \ldots \tag{6.115}$$

An example for typical parameters is shown in Fig. 6.8. This provides a rough but reasonably accurate approximation to the actual dispersion curve with Raman scattering effects, both near the crossover wavenumber $k_1 \approx \tilde{k}_0$, where the dispersion relation is approximately linear, and asymptotically as $k_1 \to \infty$ where the dispersion curve is nearly horizontal. It performs somewhat less well in the transition region in between. This will be referred to as the shifted-but-Raman-free (SRF) approximation.

Eigenmodes

Turning now from the eigenfrequencies to the eigenmodes, in Figs. 6.9-6.11 we plot the proportions (with respect to action) of the bare mode contributions to the various dressed pseudo-modes as a function of the (probe) wavenumber, for fixed pump strength and resonant pump frequency $\omega_2 = \Omega_c - \omega_p$. Again, behavior as $k \to 0$ should not be taken too seriously, but otherwise the dressing behaves as expected. The upper branch, or dressed R-wave-like mode, is mostly (but by no mean exclusively) gyron-like for low wavenum-

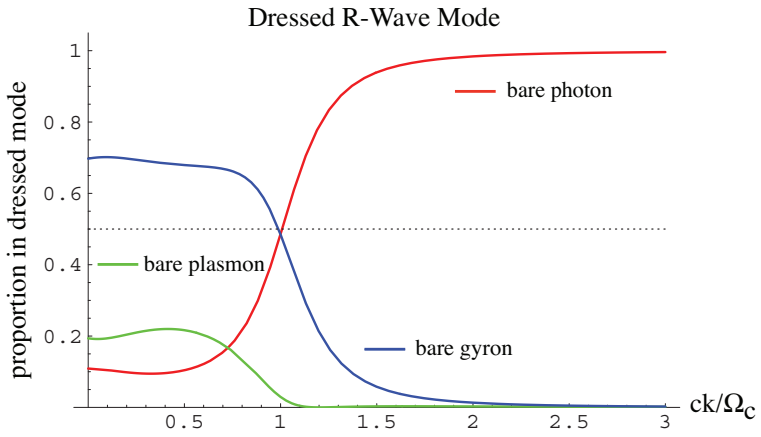

Figure 6.9. Relative proportions of bare mode actions, as a function of normalized probe wavenumber, present in the dressed R-wave-like mode, assuming a pump strength of $|a_{\text{pump}}| = 0.1$ and pump frequency $\omega_2 = \Omega_c - \omega_p$, where $\omega_p = 0.25\Omega_c$.

bers when the frequency lies nearest to he upper resonance, and then becomes increasingly photon-like as $k_1 \to \infty$ and the frequency $\omega_r(k)$ asymptotes to the dispersion relation for bare photons in an unmagnetized plasma. At low k, the cyclotron-like wave starts out as mostly (but not entirely) photon-like, but with some gyronic contribution as well at low k, where $\omega_c(k)$ has a similar shape to the bare photon dispersion curve, only shifted downward, but then asymptotically approaches a 50%/50% gyron-plasmon combination as k increases and the frequency $\omega_r(k)$ levels off asymptotically as it approaches the lower resonance.

The DSP mode shows more interesting behavior. At low wavenumber, it is mostly plasmonic, with some admixture of gyronic excitation (the bit of photonic excitation as $k \to 0$ is likely an artifact of the RWA.) At large k, the polariton becomes equal parts plasmon and gyron, just like the dressed cyclotron mode, although with different phases to ensure symplectic orthogonality. In between, however, the gyronic contribution dips to zero, and both the photon and plasmon contribution show local maxima. Note that the gyronic contribution vanishes at some particular wavenumber $\tilde{k}_d$ in between the crossover wavenumber $\tilde{k}_0$ and the bare resonant wavenumber k_0.

In the total absence of Raman scattering, these wavenumbers would all coincide. Over some range of wavenumbers in a neighborhood of $\tilde{k}_d$ corresponding to frequencies near the center of the EIT passband, the polariton has little or no gyron character. From a quantum-state perspective, this means the population of the Landau levels, which would imply absorption of probe photons, remains small.

At some fixed wavenumber near the center of the EIT pass-band, the DSP mode has

minimal gyronic contribution, but can be tuned between more photon-like or more plasmon-like excitation by adjusting the pump strength. Fig. 6.12 shows the bare mode contributions to the polariton as a function of applied pump strength for fixed wavenumber $k_1 = \tilde{k}_d$, still assuming a resonant pump at frequency $\Omega_2 - \omega_p$.

In the limit as $|a_{\text{pump}}| \to 0$, the Landau scattering necessary for EIT vanishes, and the polariton must coincide with the plasmon, without any admixture of transverse EM DOFs which would suffer resonant absorption. As the pump strength $|a_{\text{pump}}|$ increases, the polariton becomes less plasmon-like and increasingly photon-like in character, with a concomitant steepening of the dispersion curve near $\omega_d(k_1) = \Omega_c$. Note that at $\tilde{k}_d$, the gyron contribution to the polariton remains negligible for sufficiently small $|a_{\text{pump}}|$, and then begins to grow noticeably, eventually asymptoting to some moderately small but finite action fraction.

However, these larger pump values [$|a_{\text{pump}}| \gtrsim O(\frac{1}{4})$] are rather impractical experimentally at microwave frequencies, and will otherwise lead to neglected relativistic, wave-breaking, or higher-order scattering effects which will invalidate the current theory, so we expect to remain in the regime where the gyron contribution is small. Note that this gyronic admixture is purely a consequence of the Raman scattering, as can also be seen in Figure 6.12, where the polariton remains gyron-free at $k_1 = k_0$ for any pump level. Unphysically setting $\Omega_{\text{R}} = 0$ in (6.112), we verify that at k_0 the gyron content vanishes, $|\tilde{\gamma}|^2 = 0$, while the photon and plasmon components appear in the proportion $\left|\frac{\tilde{\alpha}}{\tilde{\beta}}\right| \sim \frac{|\Omega_{\text{L}}|}{\Omega_{\text{G}}}$ which grows monotonically with pump strength.

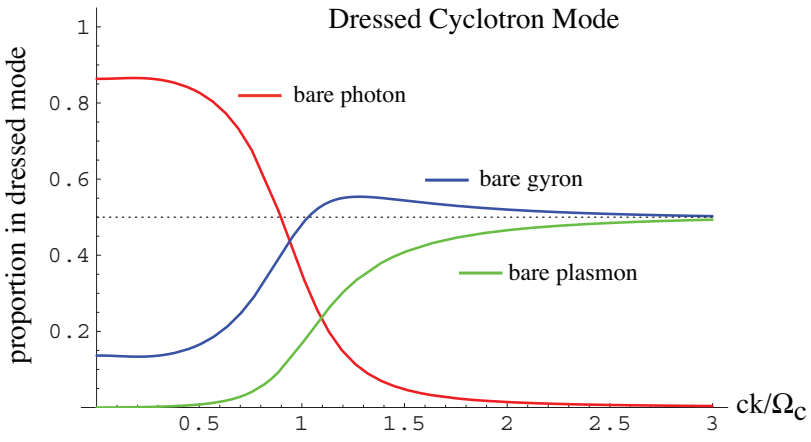

Figure 6.10. Relative proportions of bare mode actions, as a function of normalized probe wavenumber, present in the dressed cyclotron-wave-like mode, assuming a pump strength of $|a_{\text{pump}}| = 0.1$ and pump frequency $\omega_2 = \Omega_c - \omega_p$, where $\omega_p = 0.25\Omega_c$.

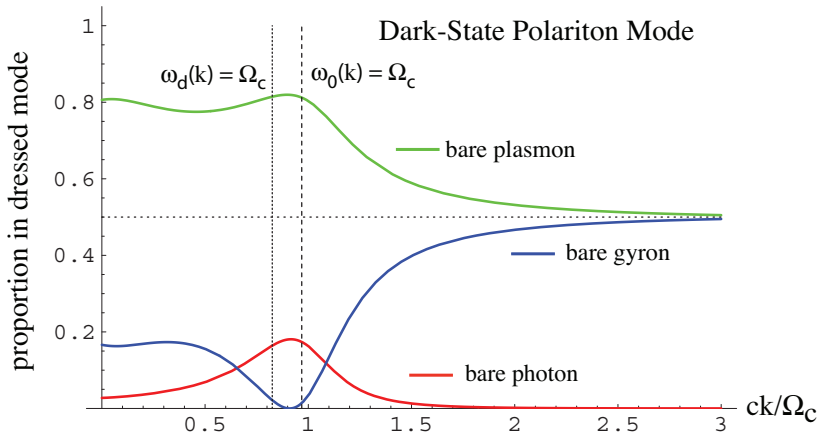

Figure 6.11. Relative proportions of bare mode actions, as a function of normalized probe wavenumber, present in the dark-state polariton mode, assuming a pump strength of $|a_{\rm pump}| = 0.1$ and pump frequency $\omega_2 = \Omega_c - \omega_p$, where $\omega_p = 0.25\Omega_c$.

Gyron-Free Approximation to the DSP

In counter-propagating geometry, for sufficiently small to moderate pump strengths and near-resonant pump frequency, and for dressed polariton frequencies $\omega_d \approx \Omega_c$ sufficiently near the center of the EIT pass-band, we have seen that the relative gyronic contribution to the DSP (proportional to the magnitude squared of the eigenmode expansion coefficient) remains small. Furthermore, at a nearby wavenumber, each of the three bare mode proportions exhibits a local extremum, so the dependence of the dressed mode composition on k_1 is relatively weak. In this regime one therefore expects to be able to approximate the DSP by a simplified two-mode description which disregards the gyron contribution, and

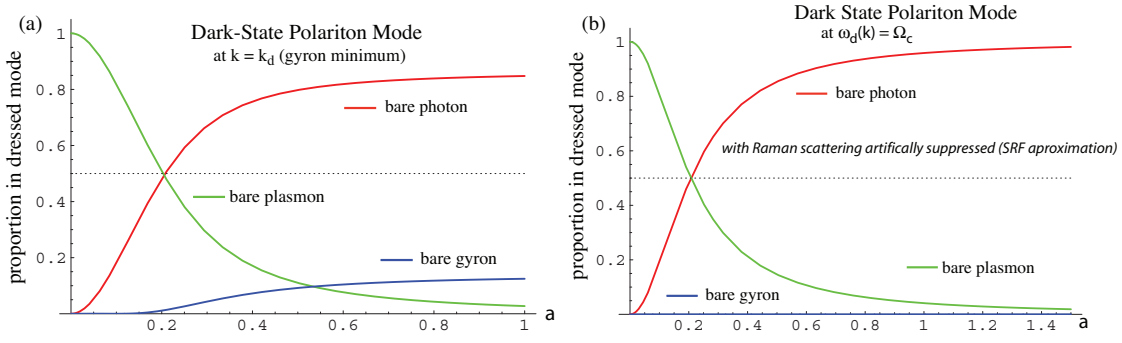

Figure 6.12. Relative proportions of bare mode actions in DSP mode, as a function of normalized pump strength at fixed wavenumber k_1, pump frequency $\omega_2 = \Omega_c - \omega_p$, and $\omega_p = 0.25\Omega_c$. In (a), full RWA dynamics are included, for wavenumber $k_1 = \tilde{k}_d$. For comparison, in (b) the Raman scattering effects are artificially suppressed, and mode contributions are shown for $k_1 = k_0$.

which can then allow us to use the elegant $\mathcal{SU}(2)$ formalism for dressing, simplifying the algebraic form for the DSP interaction eigenfrequency and eigenmode expansion coefficients, and offering some semi-quantitative insight into the DSP mode in the central part of the pass-band. Furthermore, in this regime the dependence of the DSP mixing angle θ_d on probe wavenumber k_1 is reasonably weak, and may be neglected in a first approximation.

Since K is Hermitian, a variational method offers the best means to obtain an approximation to the DSP constrained by the assumption that $\tilde{\gamma}(k_1) = 0$. (Note that this will amount to a Raleigh-Ritz-type technique applied at the level of first quantization to the c-number coupling matrix K, not at the level of second quantization, which would involve the q-number operators or quantum states themselves.) We seek an approximation to the DSP pseudo-mode corresponding to the interaction eigenfrequency $\tilde{\omega}_d(k_1)$ for the polariton mode, which is always the smallest in *magnitude* of the three branches, but K is not positive-definite, so we actually proceed by minimizing the quadratic form associated with $K^\dagger K = K^2$, which has the same eigenvectors as K itself but is non-negative:

$$\begin{bmatrix} \tilde{\alpha}(k_1) \\ \tilde{\beta}(k_1) \end{bmatrix} \approx \underset{\tilde{\alpha},\,\tilde{\beta}}{\arg\min} \begin{bmatrix} \tilde{\alpha}^* & \tilde{\beta}^* & 0 \end{bmatrix} K^2 \begin{bmatrix} \tilde{\alpha} \\ \tilde{\beta} \\ 0 \end{bmatrix} \tag{6.116a}$$

$$\text{such that: } |\tilde{\alpha}|^2 + \left|\tilde{\beta}\right|^2 = 1. \tag{6.116b}$$

After the approximate eigenvector of expansion coefficients is determined, the corresponding matrix element of K serves as an approximation to the eigenvalue:

$$\tilde{\omega}_d(k_1) \approx \begin{bmatrix} \tilde{\alpha}(k_1)^* & \tilde{\beta}(k_1)^* & 0 \end{bmatrix} K \begin{bmatrix} \alpha(\tilde{k}_1) \\ \beta(\tilde{k}_1) \\ 0 \end{bmatrix}. \tag{6.117}$$

For the dressed modes, we retain the convention implicit in (6.108) that the expansion coefficients for the bare photon and gyron contributions are chosen to be real, while the plasmon contribution (driven, after all, by a beat wave between probe and pump) has a relative phase shift determined by $\arg[i\tilde{a}_2(k_2)]$. That is, we parameterize $\tilde{\alpha}$ and $\tilde{\beta}$ in terms of a real mixing angle $\theta_d(k_1)$ and a plasmon phase shift $\psi_p(k_2)$. In particular, for the DSP mode $\tilde{\chi}_d$ and the complementary (i.e., symplectically-orthogonal) bright-state polariton (BSP) mode $\tilde{\chi}_b$ also involving the photon and plasmon excitations, the unitary transformation can be written as

$$\tilde{\chi}_d = \quad e^{i\psi_p} \cos\theta_d\, \tilde{f} + \sin\theta_d\, \tilde{a}_1; \tag{6.118a}$$

$$\tilde{\chi}_b = -e^{i\psi_p} \sin\theta_d\, \tilde{f} + \cos\theta_d\, \tilde{a}_1; \tag{6.118b}$$

which can be shown to satisfy equal-time CCRs for any choice of real probe wavenumber k_1, real mixing angle θ_d and real plasmon phase-shift ψ_d :

$$\left[\tilde{\chi}_d(k)\ , \tilde{\chi}_d(k')^\dagger\right] = \left[\tilde{\chi}_b(z,t)\ , \tilde{\chi}_b(z',t)^\dagger\right] = \delta_{k\,k'}; \tag{6.119a}$$

$$\left[\tilde{\chi}_d(k)\ , \tilde{\chi}_b(k')^\dagger\right] = \left[\tilde{\chi}_d(z,t)\ , \tilde{\chi}_b(z',t)\ \right] = 0. \tag{6.119b}$$

Note, however, that the BSP envelope $\tilde{\chi}_b(k_1)$, defined here so as to be orthogonal to the DSP, does not necessarily correspond to a dressed pseudo-mode of the field-plasma system, but under our current assumptions, the two remaining dressed eigenmode envelopes (i.e., the upper, or R-wave-like branch, and the lower, or cyclotron-wave-like branch) can be constituted as symplectic linear combinations of this BSP mode $\tilde{\chi}_b(k_1)$ and the gyron mode $\tilde{g}(k_1)$, with no admixture of the $\tilde{\chi}_d(k_1)$ mode.

After performing the variational minimization and some additional algebraic manipulations, we find for the variational approximation to the interaction eigenmodes in terms of

$$\theta_d(k_1) = \arctan\left[\left\{\frac{\sqrt{1+\mathcal{Y}(k_1)^2}-\mathcal{Y}(k_1)}{\sqrt{1+\mathcal{Y}(k_1)^2}+\mathcal{Y}(k_1)}\right\}^{1/2}\right]; \tag{6.120a}$$

$$\psi_p(k_2) = \sigma_d \frac{i a_{\text{pump}}}{|a_{\text{pump}}|}; \tag{6.120b}$$

$$\tag{6.120c}$$

and for the eigenvalue:

$$\tilde{\omega}_d(k_1) = (\omega_0(k_1)-\Omega_c)\sin^2\theta_d + (\omega_p-\tilde{\omega}_p)\cos^2\theta_d + 2\,\mathrm{Re}\left[e^{-i\psi_d}\Omega_{\mathrm{R}}\right]\sin\theta_d\cos\theta_d; \tag{6.121}$$

where for convenience we have defined the quantity

$$\mathcal{Y}(k_1) = \frac{(\omega_0(k_1)-\Omega_c)^2 - (\omega_p-\tilde{\omega}_p)^2 + \Omega_{\mathrm{G}}^2 - |\Omega_{\mathrm{L}}|^2}{2\sigma_d\,\mathrm{Re}\left[e^{-i\psi_p}\left(\Omega_{\mathrm{G}}\Omega_{\mathrm{L}} + \Omega_{\mathrm{R}}\left\{(\omega_0(k_1)-\Omega_c)+(\omega_p-\tilde{\omega}_p)\right\}\right)\right]}, \tag{6.122}$$

and where the sign $\sigma_d = \pm 1$ is chosen so that the *denominator* of $\mathcal{Y}(k_1)$ is positive. In terms of this coupling function, the trigonometric functions may be explicitly written as

$$\sin^2\theta_d = \tfrac{1}{2}\left(1 - \frac{\mathcal{Y}(k_1)}{\sqrt{1+\mathcal{Y}(k_1)^2}}\right); \tag{6.123a}$$

$$\cos^2\theta_d = \tfrac{1}{2}\left(1 + \frac{\mathcal{Y}(k_1)}{\sqrt{1+\mathcal{Y}(k_1)^2}}\right); \tag{6.123b}$$

$$\sin\theta_d\cos\theta_d = \tfrac{1}{2}\left(\frac{1}{\sqrt{1+\mathcal{Y}(k_1)^2}}\right). \tag{6.123c}$$

If $\tilde{\omega}_p = \omega_p$ and $\omega_0(k_1) = \Omega_c$ exactly, as would be appropriate if Raman scattering effects could be turned off, then this would reproduce the exact Raman-free result mentioned

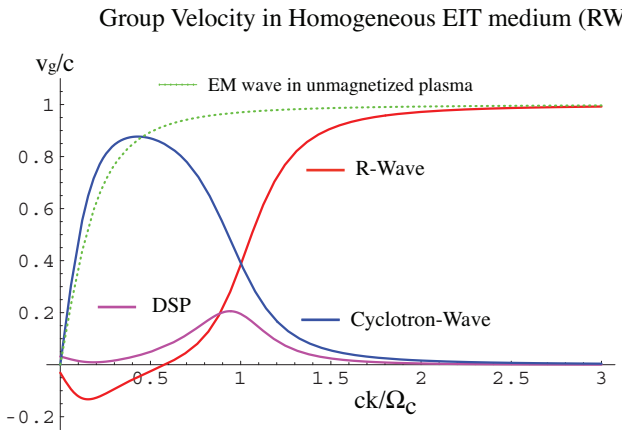

Figure 6.13. Group velocity versus wavenumber for the EIT pseudo-modes, assuming a plane-wave pump of strength $|a_{\rm pump}| = 0.1$, and pump frequency $\omega_2 = \Omega_c - \omega_p$, where $\omega_p = 0.25\Omega_c$. The odd behavior at low k is an artifact of the RWA.

above: $\bar{\theta}_d = \arctan\left[\frac{|\Omega_{\rm L}|}{\Omega_{\rm G}}\right]$, just depending on relative strength of the terms in the Hamiltonian responsible for resonant cyclotron absorption and for the Landau scattering that can suppress it. As a rough approximation to the (already approximate) gyron-free mixing angle (6.120a), we can use the translated version of Raman-free and gyron-free mixing angle:

$$\theta_d(k_1) \approx \bar{\theta}_d(\bar{k}_1), \tag{6.124}$$

which we expect to provide a reasonable estimate (within about 10% or less) for $k_1 \sim \tilde{k}_d$, leading to a shifted approximation for the interaction eigenvalue of the form

$$\tilde{\omega}_d(k_1) \approx \bar{\omega}_d\left(\bar{k}_1\right)) = \left(\omega_0(\bar{k}_1) - \Omega_c\right)\sin^2\bar{\theta}_d(\bar{k}_1) + (\omega_p - \tilde{\omega}_p)\cos^2\bar{\theta}_d(\bar{k}_1) \tag{6.125}$$

which is in effect just an action-weighted sum of the frequency differences for the constituent bare modes in the polariton.

Group Velocity in Homogeneous Medium

In the case considered so far, with an infinite (or periodic) homogeneous plasma medium and an applied plane-wave pump field, the effective group velocity for dressed mode wave packets of finite but narrow bandwidth can be taken to be the derivatives $\frac{\partial}{\partial k}\omega(k)$ of the corresponding effective dispersion relations for the pseudo-modes, just like in standard linear media. A plot for typical parameters is shown in Fig. 6.13, with the group velocity of the transverse EM waves in an unmagnetized plasma indicated for comparison.

Note that the behavior of the dressed R-wave mode or DSP mode as $k \to 0$ is an

unphysical artifact of the RWA, but the plot should be reasonably accurate for $k \gtrsim \omega_p/c$. We know the R-wave asymptotically approaches in content the unmagnetized EM mode, so we are not surprised to see the group velocities agree in that limit. Because of the competing scattering terms, the peak in the DSP group velocity does not occur precisely at the bare resonance.

We will defer discussion of the dependence of the group velocity on the pump strength a_2 until the next section, where we study group velocity effects more carefully, and also include possible effects of slow time-dependence in the pump envelope or gradual spatial dependence in the plasma density.

Slowly-Varying Envelope Approximation, Pulse Group Velocity, and Adiabatic Propagation Effects

Our previous treatment of group velocity effects has presumed near-CW probe and pump fields propagating in a completely homogeneous medium. In order to understand how a finite-duration wave-packet may enter, travel through, or leave the active EIT medium, and how the dark-state polariton excitation is established, or how it responds to slow variations in the plasma medium or in the amplitude of the control (pump) field, we need to look beyond discrete-mode dispersion relations in translation invariant media, and allow for plasma inhomogeneities and non-vanishing pulse bandwidths, or equivalently for envelopes with slow space and time dependence modulating pump and probe carrier oscillations that are assumed to satisfy local versions of the discrete-mode dispersion relations. This should be possible in principle starting with the exact RWA pseudo-modes, but would be quite involved, so in order to gain qualitative insight in a simplified analytic setting, we will pursue an even more approximate, or "cartoon" approach, where we use the slowly-varying envelope approximation (SVEA) and the adiabatic limit, but completely disregard the gyron contribution to the DSP (which is small but non-vanishing for small bandwidths around suitably matched pump and probe wavenumbers), neglect the explicit dependence of the mixing angle on wavenumber in the vicinity of the cyclotron frequency, allow for variation in the dressed group velocity with pump strength, but otherwise ignore higher-order group-velocity dispersion, and eventually neglect certain pump retardation or slippage effects and various other small terms that arise during the derivations. As we have seen, many of these approximations would be either exact or quite good if the (renormalized) Raman scattering could be neglected compared to the Landau scattering, but should still provide a reasonably faithful qualitative picture in the more realistic case as long as we account for the overall

Raman-induced shift in wavenumber at the cyclotron resonance, proportional to $\frac{\Omega_R}{\Omega_L}$, as in the SRF approximation described above.

In the SVEA, spatial and/or temporal bandwidths of the bare pump and probe and the resulting beat-wave driving the plasma oscillation are all assumed to be suitably small compared to the corresponding local carrier wavenumbers and frequencies, which themselves are at most slowly varying and taken to satisfy spatially-local versions of the approximate single-mode, gyron-free RWA polariton dispersion relations derived above, only for slowly-varying values of the plasma frequency $\omega_p(z,t)$ and/or the cyclotron frequency $\Omega_c(z,t)$. We continue to neglect diffraction or other transverse effects in the propagation.

Actually, to simplify the development, we will sacrifice some generality and will take the axial magnetic field B_0 to remain (temporally and spatially) constant, but will allow for the possibility of spatial inhomogeneities in the background density through a local plasma frequency $\omega_p(z)$, which is assumed to vary on scales long compared to all relevant carrier wavelengths. Although the SVEA formalism can in principle accommodate more general inhomogeneities or non-stationarity, within the context of our one-dimensional approximation, this likely covers realistic cases of interest: experimentally, we imagine that the plasma will be confined to the central region of some strong external solenoid where field variations are small, and inductive effects will limit temporal variations in overall axial magnetic field strength to time-scales which are likely very long compared to the duration of probe interaction in the plasma (even after allowing for slow-light effects). As for the plasma, in particular we suppose that the density approximately stabilizes at some more or less constant value over some reasonably long distance (compared to the probe pulse length) in the bulk plasma, but slowly changes in transition regions at the left and right boundaries of the plasma, eventually vanishing at sufficient distances in either direction from the center. For later convenience, we label a few typical spatial positions in the various regions: z_0 in the far upstream region, where the plasma density is negligible; $z = z_l$ in the leading transition region, where the density is changing most noticeably (yet still gradually compared to other scales); $z = z_b$ in the middle of the bulk region where the density flattens out; $z = z_t$ in the trailing transition region where the density drops; and $z = z_\infty$ in the far downstream region beyond the plasma where the pulse returns to near vacuum. Compared to any carrier wavelength, we may suppose that $z_0 \ll z_l \ll z_b \ll z_t \ll z_\infty$.

We further assume the carrier oscillations for both pump and probe remain near the one and two-photon resonances, so the dressed probe will propagate as a DSP wave-packet near the center of the EIT pass band. In particular, we suppose the probe and pump

are launched from remote sources with *fixed* carrier frequencies[7] $\omega_1 = \omega_1(z_0) \gtrsim \Omega_c$ and $\omega_2 = \omega_2(z_\infty) \approx \Omega_c - \omega_p(z_b)$ respectively, such that both $a_1(z_0,t) = e^{-\omega_1(z_0)t}\tilde{a}_1(z_0,t)$ and $a_2(z_\infty,t) = e^{i\omega_2(z_\infty)t}\tilde{a}_2(z_\infty,t)$ are prescribed slowly-varying functions of time. We assume that the probe is, if not strictly of compact support, at least weakly localized in time, in the sense that there exist times $t_0 < t_f$ such that $|\tilde{a}_1(z_0,t)|$ is negligible for $t < t_0$ or $t > t_f$.

Strictly speaking, these assumptions provide time-dependent boundary conditions, but since at remote times and positions the medium is assumed to be homogeneous, time-stationary, and dispersionless, they can be converted into equivalent initial conditions over all positions, which are more naturally handled in the Heisenberg picture of quantum dynamics. For the prescribed classical pump field, we ignore any nonlinear back-reaction as before, and define the c-number spatial envelope by removing the fast (spatio-temporal) carrier phase dependence:

$$\bar{a}_2(z,t) = a_2(z;t)e^{-i\psi_2(z,t)}; \tag{6.126a}$$

$$\psi_2(z,t) = \int_{z_0}^{z} dz'\, k_2(z') - \omega_2(t-t_0), \tag{6.126b}$$

where $k_2(z) = \frac{\partial}{\partial z}\psi_2(z,t)$ is the local wavenumber along the lower (RWA) R-polarized (cyclotron) branch, i.e., the positive solution of

$$\omega_2 = \tfrac{1}{2}\left[\omega_0(k_2(z),z) + \Omega_c\right] - \tfrac{1}{2}\sqrt{\left[\omega_0\left(k_2(z),z\right) - \Omega_c\right]^2 + 2\omega_p(z)^2\tfrac{\Omega_c}{\omega_0(k_2(z),z)}}, \tag{6.127}$$

in which

$$\omega_0(k,z) = +\sqrt{c^2k^2 + \omega_p(z)^2} \tag{6.128}$$

is the local dispersion relation for transverse EM waves in the unmagnetized plasma.

For the bare probe photon, probe gyron, and plasmon fields, we then define q-numbers (annihilation operators) corresponding to the slowly-varying envelopes modulating the fast carrier oscillations, in terms of the real-space Heisenberg-picture modes:

$$\bar{a}_1(z,t) \equiv a_1(z,t)e^{-i\psi_1(z,t)}; \tag{6.129a}$$

$$\bar{g}_1(z,t) \equiv g_1(z,t)e^{-i\psi_1(z,t)}; \tag{6.129b}$$

$$\bar{f}(z,t) \equiv f(z,t)e^{-i\psi_p(z,t)}; \tag{6.129c}$$

[7] Again, we could more generally allow for a time-dependent plasma medium or chirping of the pump or probe carriers, but such complications are avoided here. Sufficiently weak chirping could instead be incorporated into the envelopes, if desired.

where the c-number carrier phase $\psi_1(z,t)$ is related to the instantaneous local frequency and wavenumber in the conventional manner,

$$\frac{\partial}{\partial z}\psi_1(z,t) = k_1(z,t), \tag{6.130a}$$

$$-\frac{\partial}{\partial t}\psi_1(z,t) = \omega_1, \tag{6.130b}$$

$$\psi_1(z_0,t_0) = 0, \tag{6.130c}$$

which are determined by a local version of the dispersion relation in a specific manner spelled out shortly. For the longitudinal (plasmonic DOFs),

$$\psi_p(z,t) = \psi_1(z,t) - \psi_2(z) \tag{6.131}$$

is defined as the local beat phase between pump and dressed probe, such that $k_p(z,t) = \frac{\partial}{\partial z}\psi_p(z,t) = k_1(z,t) - k_2(z)$ is the local plasmon wavenumber, and $-\frac{\partial}{\partial t}\psi_p(z,t) = \omega_1 - \omega_2$ is the beat frequency. Note that here we have defined the carrier oscillation of the plasmon mode in terms of the beat between the gyron and pump, rather than in terms of the local spatially-varying plasma frequency. Under current assumptions, these frequencies are expected to be close in the bulk plasma (i.e., for $z \sim z_b$) , but in the transition region the beat frequency better reflects the frequency at which the plasma wave is actually being driven. Also note that, except for the prescribed pump, the carrier frequencies for the transverse and longitudinal envelopes here are all offset by the same frequency shift $(\omega_1 - \Omega_c)$ compared to the interaction-picture amplitudes which arose in our earlier discrete-mode scenario, but this just lead to a compensatory shift in the interaction eigenvalues, and will not affect the form of the eigenvectors of the coupling matrix.

To obtain truly slowly-varying envelopes, we have defined the over-barred operators with the fast spatial as well as temporal dependence in the phase extracted. For technical reasons emerging from our subsequent derivation of the SVEA dynamics, we also introduce the tilded envelopes

$$\tilde{a}_1(z,t) \equiv a_1(z,t) e^{-i\int_{t_0}^{t} dt' \frac{\partial}{\partial t'}\psi_1(z,t')}, \tag{6.132a}$$

$$\tilde{g}_1(z,t) \equiv g_1(z,t) e^{-i\int_{t_0}^{t} dt' \frac{\partial}{\partial t'}\psi_1(z,t')}, \tag{6.132b}$$

$$\tilde{f}(z,t) \equiv f(z,t) e^{-i\int_{t_0}^{t} dt' \frac{\partial}{\partial t'}\psi_1(z,t') - \omega_2(t-t_0)}, \tag{6.132c}$$

which remain slowly-varying in time but include the full carrier oscillations in space, and are closely related to the spatial Fourier transforms of the reciprocal-space interaction-picture

mode operators introduced above in the single-mode analysis. Obviously the magnitudes of the corresponding over-barred and tilded envelopes are equal, while their phases differ by rapidly-varying c-numbers.

Now, to determine equations of motion for the enveloped operators, we employ what is essentially a WKB-type analysis, but for collective fields whose transverse and longitudinal components can oscillate at different carrier frequencies. We can either start explicitly with a multi-component wave-packet and then effectively diagonalize to find the DSP spatial envelope, or else begin with the DSP envelope and later extract from it the photonic and plasmonic components if needed. The latter approach is probably easier, and will be adopted here.

By retaining the RWA but ignoring the gyron contribution to the DSP over the bandwidth of the probe, we can take advantage of the simple and elegant $\mathcal{SU}(2)$ framework for coupled-mode dressing (mentioned above in the RWA analysis of the magnetized-plasma dispersion relation without pump), effectively involving a rotation of the bare modes by a coherent mixing angle θ_d. This gyron-less approximation to the DSP improves to the extent that we can neglect the Raman scattering terms compared to the Landau scattering terms, so we will suppose a co-propagating geometry where their relative strength is more favorable. In its effects on the envelope dynamics, the RWA itself improves as the separation between bare frequencies increases or the coupling strengths decrease.

The simplest means to derive the SVEA Hamiltonian governing the DSP wave-packet dynamics is to make judicious use of Taylor expansions and Weyl[8] transforms. We start with the continuum k-space interaction Hamiltonian for the dark-state polariton in an infinite homogeneous medium:

$$\begin{aligned}\mathscr{H}_{\text{DSP}} &= \hbar \int dk\ [(\Omega_c - \omega_1) + \tilde{\omega}_d(k)]\, \tilde{\chi}_d(k)^\dagger \tilde{\chi}_d(k) \\ &= \hbar \iint dk\, dk'\, \delta(k-k')\, [(\Omega_c - \omega_1) + \tilde{\omega}_d(k)]\, \tilde{\chi}_d(k')^\dagger \tilde{\chi}_d(k),\end{aligned} \tag{6.133}$$

where we have included the overall shift in the definition of the carrier frequencies mentioned above. Using the well-known "covariance" of the Hilbert-Schmidt inner product under Weyl

[8] Also known in the literature as the Weyl-Wigner-Groenewald-Ville–Moyal transform, or some permutation thereof, eponymously named for those whose seminal publications were made, respectively, in 1931, 1932, 1946, 1948, and 1949. This transform and related algebraic structures have been discovered and re-discovered many times in various physical, mathematical, or engineering contexts, and extensively discussed in recent decades. Even in the context of quantum mechanics, the Wigner function was introduced before Wigner by Dirac in 1930 and Heisenberg in 1931, but they apparently believed it provided only an approximate rather than exact phase-space representation for quantum mechanics.

transforms, this can be written as

$$\mathscr{H}_{\rm DSP} = \tfrac{\hbar}{2\pi}\iint dz\, dk\, \left[(\Omega_c - \omega_1) + \tilde{\omega}_d(k)\right] W_d(z,k), \tag{6.134}$$

where $W_d(z,k)$ is the normally-ordered, operator-valued Weyl transform of the second-order DSP coherence operator $\tilde{\chi}_d(k')^\dagger \tilde{\chi}_d(k)$ with respect to the wave-kinetic phase space coordinates,[9] given explicitly by:

$$W_{dd}(z,k) = \int dq\, e^{+iqz}\, \tilde{\chi}_d\left(k + \tfrac{1}{2}q\right)^\dagger \tilde{\chi}_d\left(k - \tfrac{1}{2}q\right). \tag{6.135}$$

Next we make the replacements $\omega_1 \to \omega_1(z,t)$ and $\tilde{\omega}_d(k_1) \to \tilde{\omega}_d(k_1;z;t)$ to allow for slow space-time variation in addition to the fast carrier wavenumber dependence. For a sufficiently narrow-bandwidth probe pulse, we expect that we can then Taylor expand this interaction-frequency Weyl symbol in probe wavenumber k around the appropriate solution $k_1(z,t)$ to the local dispersion relation, resulting in:

$$\begin{aligned}(\Omega_c - \omega_1) + \tilde{\omega}_d(k_1;z;t) \approx [(\Omega_c - \omega_1(z,t)) + \tilde{\omega}_d(k_1(z,t);z;t) - v_d(x,t)k_1(z,t)] \\ + \tfrac{1}{2}\left(k\, v_d(z,t) + v_d(z,t)\, k\right) + \ldots,\end{aligned} \tag{6.136}$$

where $v_d(z,t)$ is the local polariton group velocity:

$$v_d(z,t) = \tfrac{\partial}{\partial k_1}\tilde{\omega}_d\Big|_{k_1 = k_1(z,t)}. \tag{6.137}$$

As we have seen, the dependence of the mixing angle on the wavenumber k_1 is weak near the resonance, so neglecting these derivatives and using the shifted-but-Raman-free (SRF) approximation, we find

$$v_d(z,t) \approx \sin^2\bar{\theta}_d(z,t)\, v_0\big(\bar{k}_1(z,t)\big) \equiv \sin^2\theta_d(z,t)\, v_0\,(z,t)\,, \tag{6.138}$$

which is just the local action-weighted average of the group velocity $v_0(k) = c^2\frac{k}{\omega_0(k)}$ of the bare photon mode and the group velocity of the plasmon mode, that vanishes in the cold limit.

With these same approximations, the local SVEA mixing angle itself is taken as

$$\theta_d(z,t) \approx \arctan\left[\frac{\left|\bar{\Omega}_{\rm L}(z,t)\right|}{\bar{\Omega}_{\rm G}(z,t)}\right], \tag{6.139}$$

[9] Any operator-valued or real-valued Weyl transforms on the so-called wave-kinetic phase space, with q-number coordinates (z,k), should not be confused with Weyl transforms that could be performed on the mode quadrature phase space (i.e., sine-like and cosine-like components of the oscillator operators), with q-number normal coordinates $\tilde{\chi}_d(k)$ and $\tilde{\chi}_d(k)^\dagger$ at each wavevector k. In the context of quantum optics, the former might be called a first-quantized Weyl transform, while the latter is a second-quantized Weyl transform. We will only make use of the first-quantized version here. We could also define Weyl transforms on an extended phase space which includes $(t,-\omega)$, as canonical coordinates rather than just t as a parameter. Such an approach can offer advantages in some situations, but since we only know the equal-time commutation relations of the field operators, here it seems preferable to stick with the one-time Wigner function.

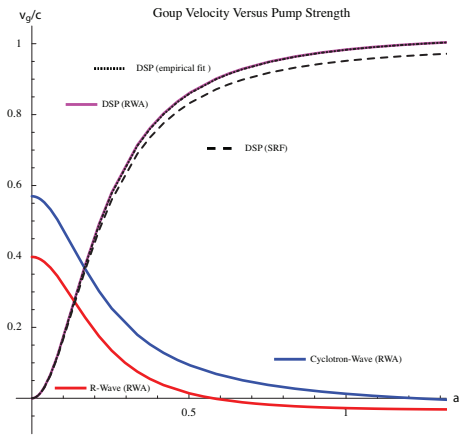

Figure 6.14. Group velocity versus pump strength for the EIT pseudo-modes, evaluated at a fixed carrier wavenumber k_1 corresponding to the point of resonance cancellation where the DSP and unmagnetized dispersion relations cross for the pump strength $|a_{\rm pump}| = 0.1$, and a fixed pump frequency $\omega_2 = \Omega_c - \omega_p$, where $\omega_p = 0.25\Omega_c$. The exact (with the RWA) group velocities for the dressed R-wave (red), dressed Cyclotron wave (blue), and the Dark-State Polariton (magenta) are shown, along with the estimate from the SRF approximation (dashed) and a simple empirical fit (dotted).

where

$$\bar{\Omega}_{\rm L}(z,t) = +i\tfrac{1}{2}\sqrt{\Omega_c\omega_2}\,\frac{ck_2(z)}{\sqrt{\omega_p(z)\omega_0(k_2(z),z)}}\,\bar{a}_{\rm pump}(z,t), \tag{6.140}$$

is the local version of the Rabi frequency associated with the Landau scattering interaction, depending on $k_2(z)$ but not $\bar{k}_1(z,t)$, and

$$\bar{\Omega}_{\rm G}(z,t) = \tfrac{1}{\sqrt{2}}\omega_p\sqrt{\frac{\Omega_c}{\omega_0(k_1(z,t))}} \tag{6.141}$$

is the local gyron-photon coupling frequency, which does depend on the local (and shifted) probe wavenumber $\bar{k}_1(z,t)$.

For a homogeneous medium and a steady-state pump, the SRF estimate for the DSP group velocity is shown in Fig. 6.14 as a function of the applied pump strength, along with the group velocities for each of the dressed modes as determined directly from the effective RWA dispersion relations. The SRF approximation (for fixed k_1) is reasonably accurate at low values of a_2 near the parameter values at which k_1 was determined via $\omega_d(k_1) = \Omega_c$, but at larger pump strengths it systematically underestimates the group velocity. Also shown is a simple empirical fit to the DSP group velocity using the functional form

$$v_d(\theta_d) = v_0\frac{(1+\delta_0)\frac{\bar{\Omega}_{\rm L}^2}{\bar{\Omega}_{\rm G}^2}}{1+(1+\delta_1)\frac{\bar{\Omega}_{\rm L}^2}{\bar{\Omega}_{\rm G}^2}} \tag{6.142}$$

for small adjustable parameters δ_0 and δ_1, such that (6.142) amounts to a small functional perturbation of the SRF form. The accuracy of this fit suggest that another approximation

nearly as simple but more accurate than the SRF is waiting to be discovered, but so far it has eluded us.

Actually, we should really continue to denote the mixing angle as $\bar{\theta}_d(z,t)$ to indicate that it arises within our shifted-but-Raman-free (SRF) approximation, but we will just use $\theta_d(z,t)$ for the sake of notational simplicity, as no other form of the mixing angle will be employed throughout our discussion of envelope propagation. For our purposes here, the exact form of the mixing angle $\theta_d(z,t)$ is not as important as its general behavior, increasing smoothly and monotonically with the local pump strength (magnitude $|\bar{a}_2(z,t)|$ of the pump envelope) or decreasing monotonically with the local background plasma density $\omega_p(x)$.

In all cases of interest to us, the local effective dispersion relation for transverse components can be taken as

$$(\Omega_c - \omega_1(z,t)) + \tilde{\omega}_d(k_1(z,t);z;t) = 0, \tag{6.143}$$

so performing inverse Weyl transforms, only this time back into the position representation, the Hamiltonian becomes

$$\begin{aligned}\mathscr{H}_{\rm DSP} \approx & -\hbar \int dz\ [v_d(x,t)k_1(z,t)]\ \tilde{\chi}_d(z;t)^\dagger \tilde{\chi}_d(z;t) \\ & - \tfrac{i\hbar}{2} \iint dz dz'\, \delta'(z-z')\left[v_d(z,t) + v_d(z',t)\right] \tilde{\chi}_d(z';t)^\dagger \tilde{\chi}_d(z;t),\end{aligned} \tag{6.144}$$

where the DSP envelope $\tilde{\chi}_d(z;t)$ is a spatially local version of the corresponding reciprocal-space pseudo-modes previously analyzed, and satisfies the equal-time CCRs:

$$\left[\tilde{\chi}_d(z;t)\ , \tilde{\chi}_d(z';t)^\dagger\right] = \delta(z-z'), \tag{6.145}$$

and hence evolves according to

$$\begin{aligned}\tfrac{\partial}{\partial t}\tilde{\chi}_d(z;t) &= \tfrac{1}{i\hbar}\left[\tilde{\chi}_d(z;t), \mathscr{H}_{\rm DSP}\right] \\ &= +iv_d(x,t)k_1(z,t)\tilde{\chi}_d(z;t) \\ &\quad - v_d(z,t)\tfrac{\partial}{\partial z}\tilde{\chi}_d(z;t) - \tfrac{1}{2}\tilde{\chi}_d(z;t)\tfrac{\partial}{\partial z}v_d(z,t).\end{aligned} \tag{6.146}$$

At microwave frequencies, linear measurements (i.e., of the field rather than of intensity) are still feasible in principle, but time-averaged intensity-based measurements, related to the action (line) density $\tilde{\chi}_d(z;t)^\dagger\tilde{\chi}_d(z;t)$ are often of greater theoretical interest and experimental practicality. For greater compactness of notation, we can write the DSP action density as $|\tilde{\chi}_d(z;t)|^2 = |\bar{\chi}_d(z;t)|^2$, and similarly for the bare-mode action densities, as long as it is understood that they always are taken to be normally-ordered. It follows from (6.146) and

the product rule that the DSP action density transport is governed by a continuity-type equation

$$\tfrac{\partial}{\partial t}\left|\bar{\chi}_d(z;t)\right|^2 + \tfrac{\partial}{\partial z}\left[v_d(z,t)\left|\bar{\chi}_d(z;t)\right|^2\right] = 0. \tag{6.147}$$

This might have been guessed from the start, but it is reassuring that it naturally emerges from our formalism.

We still need to relate the DSP envelope to the constituent bare photon and plasmon envelopes. Starting in k-space, the coherence operators for the bare photonic and plasmonic modes are related to the dark and bright state polariton modes via the dressing relations:

$$\begin{bmatrix} \tilde{f}(k_1')^\dagger \tilde{f}(k_1) & \tilde{f}(k_1')^\dagger \tilde{a}(k_1) \\ \tilde{a}(k_1')^\dagger \tilde{f}(k_1) & \tilde{a}(k_1')^\dagger \tilde{a}(k_1) \end{bmatrix} = V(k_1') \begin{bmatrix} \tilde{\chi}_d(k_1')^\dagger \tilde{\chi}_d(k_1) & \tilde{\chi}_d(k_1')^\dagger \tilde{\chi}_b(k_1) \\ \tilde{\chi}_b(k_1')^\dagger \tilde{\chi}_d(k_1) & \tilde{\chi}_b(k_1')^\dagger \tilde{\chi}_b(k_1) \end{bmatrix} V(k_1)^\dagger, \tag{6.148}$$

where

$$V(k_1) = \begin{bmatrix} e^{i\psi_p}\cos\theta(k_1) & -e^{i\psi_p}\sin\theta_d(k_1) \\ \sin\theta_d(k_1) & \cos\theta_d(k_1) \end{bmatrix} \tag{6.149}$$

is the unitary matrix effecting the canonical dressing transformation in reciprocal space. Note that the bright state polariton (BSP) mode is chosen to be orthogonal to the DSP and also gyron-free, but may or may not be a good approximation to one of the actual pseudo-modes of the system. To incorporate slow space-time variation, we now follow a strategy similar to that in deriving the Hamiltonian. Performing Weyl transforms, we obtain

$$\begin{bmatrix} W_{ff}(z,k) & W_{fa}(z,k) \\ W_{af}(z,k) & W_{aa}(z,k) \end{bmatrix} = V(k) \star \begin{bmatrix} W_{dd}(z,k) & W_{db}(z,k) \\ W_{bd}(z,k) & W_{bb}(z,k) \end{bmatrix} \star V(k)^\dagger, \tag{6.150}$$

where the diagonal elements of the Weyl tensors are defined as in (6.135) above, and the off-diagonal terms are defined as the obvious generalizations, e.g.,

$$W_{af}(z,k) = \int dq\, e^{+iqz}\, \tilde{a}\left(k+\tfrac{1}{2}q\right)^\dagger \tilde{f}\left(k-\tfrac{1}{2}q\right); \tag{6.151}$$

and $\star$ denotes the combined matrix/Moyal product for 2×2 matrices of Weyl symbols. We now make the substitution $V(k) \to V(k,z;t)$ so as to allow for slow variation in the density and pump amplitude, Taylor expand around $k \approx k_1(z,t)$, and perform inverse Weyl transforms back to a real-space representation. In general, the dressing relations that were local in k-space will not remain local when transformed to z-space, but contain convolutional corrections of higher-order (in certain eikonal slowness parameters measuring the ratio of envelope to carrier wavenumbers, which we need not define more precisely). However, as the leading-order corrections are all proportional to the derivative $\frac{\partial}{\partial k}\theta_d\big|_{k=k_1(z,t)}$ which has already been assumed small for $\omega_1 \approx \Omega_c$, we will neglect these terms, and the dressed

envelopes, in particular the dark state polariton envelope, can be taken to be just local versions of the corresponding pseudo-modes previously analyzed. In particular, in terms of the bare probe photon and plasmon envelope operators, the slowly-varying envelopes for the dark-state polariton (DSP) operator and the orthogonal bright-state polariton (BSP) operator become

$$\tilde{\chi}_d(z,t) = \quad e^{i\psi_p(z,t)}\cos\big(\theta_d(z,t)\big)\tilde{f}(z,t) + \sin\big(\theta_d(z,t)\big)\tilde{a}_1(z,t); \tag{6.152a}$$

$$\tilde{\chi}_b(z,t) = -e^{i\psi_p(z,t)}\sin\big(\theta_d(z,t)\big)\tilde{f}(z,t) + \cos\big(\theta_d(z,t)\big)\tilde{a}_1(z,t); \tag{6.152b}$$

which also satisfy equal-time CCRs

$$\Big[\tilde{\chi}_d(z,t)\ ,\tilde{\chi}_d(z',t)^\dagger\Big] = \Big[\tilde{\chi}_b(z,t)\ ,\tilde{\chi}_b(z',t)^\dagger\Big] = \delta(z-z'); \tag{6.153a}$$

$$\Big[\tilde{\chi}_d(z,t)\ ,\tilde{\chi}_b(z',t)^\dagger\Big] = \Big[\tilde{\chi}_d(z,t)\ ,\tilde{\chi}_b(z',t)\ \Big] = 0. \tag{6.153b}$$

Of course the spatially-local, 2-parameter form of (6.152) does not encompass the most general possible linear symplectic transformation between two bare and dressed mode envelopes, but is all that is needed for approximations consistent with a local version of (6.108) when the small gyron contribution is neglected along with the weak k-dependence of the mixing angle near resonance. As in the discrete mode case, the BSP envelope $\tilde{\chi}_b(z,t)$, defined here so as to be orthogonal to the DSP, does not necessarily correspond to a dressed pseudo-mode, but under our assumptions the two remaining dressed eigenmode envelopes (i.e., the upper, or R-wave-like branch, and the lower, or cyclotron-wave-like branch) can always be constituted as symplectic linear combinations of $\tilde{\chi}_b(z,t)$ and a slowly varying gyron envelope $\tilde{g}(z,t)$ with no admixture of $\tilde{\chi}_d(z,t)$.

Space-time dependence in the mixing angle $\theta_d(z,t)$ arises from density changes via $\omega_p(z)$ and/or modulations in the pump field via $\tilde{a}_2(z,t)$, while variation in the plasmon phase-shift $\psi_p(z,t) \sim \arg\left[i\tilde{a}_2(z,t)\right]$ arises only through implicit dependence of the Rabi frequencies on the pump envelope.

Obviously, the relations (6.152) can be inverted to express the bare photon and plasmon SVEA mode envelopes in terms of the dressed bright and dark polariton operators:

$$\tilde{f}(z,t) = \big[\cos\big(\theta_d(z,t)\big)\tilde{\chi}_d(z,t) - \sin\big(\theta_d(z,t)\big)\tilde{\chi}_b(z,t)\big]\, e^{-i\psi_p(z,t)} \tag{6.154a}$$

$$\tilde{a}_1(z,t) = \sin\big(\theta_d(z,t)\big)\tilde{\chi}_d(z,t) + \cos\big(\theta_d(z,t)\big)\tilde{\chi}_b(z,t)\ . \tag{6.154b}$$

For any quantum state such that the gyron and BSP modes remain in vacuum, or nearly so, and only the DSP is assumed to be appreciably excited, any *normally-ordered* expectation values, transition amplitudes, or matrix elements involving $\tilde{\chi}_b(z,t)$ can be taken to vanish,

and any bare photonic or plasmonic excitation may be assumed to be part of the DSP only, which means that we can effectively set

$$\tilde{\chi}_b(z,t) \to 0; \tag{6.155a}$$

$$\tilde{g}_1(z,t) \to 0; \tag{6.155b}$$

and write

$$\tilde{f}(z,t) \to e^{-i\psi(z,t)} \cos\big(\theta_d(z,t)\big)\tilde{\chi}_d(z,t); \tag{6.156a}$$

$$\tilde{a}_1(z,t) \to \sin\big(\theta_d(z,t)\big)\bar{\chi}_d(z,t); \tag{6.156b}$$

or

$$\big|\bar{f}(z,t)\big|^2 \approx \cos^2\big(\theta_d(z,t)\big)\,|\tilde{\chi}_d(z,t)|^2\,; \tag{6.157a}$$

$$|\bar{a}_1(z,t)|^2 \approx \sin^2\big(\theta_d(z,t)\big)\,|\tilde{\chi}_d(z,t)|^2\,. \tag{6.157b}$$

Thus as expected the local mixing angle $\theta_d(z,t)$ simply tunes the local relative proportions of bare photonic or bare plasmonic action in the dressed DSP mode. However, we must reiterate that the relations (6.155) and (6.156) cannot be valid as operator equations (the commutation relations are manifestly inconsistent), but instead are ultimately valid as substitutions in *normally-ordered* c-number matrix elements between states where the bright-state and gyronic excitations are negligible.

From the present perspective, the action transport is governed fundamentally by the continuity equation (6.147), representing a local conservation law for the number of dressed polaritons, where the effective group velocity for the dressed mode envelope is the locally action-weighted average of the group velocities for the constituent bare modes. The constituent bare mode contributions to the DSP envelope can then be determined from (6.157). By integrating (6.147) over all positions and assuming the wave-packet is at least weakly localized in space, we confirm that the total number of polaritons is conserved, i.e.,

$$\tfrac{\partial}{\partial t}\int dz\ |\bar{\chi}_d(z,t)|^2 = 0, \tag{6.158}$$

although the numbers of bare probe photons and bare plasmons are not conserved individually. This continuity equation is also consistent with expectations from local Manley-Rowe relations – namely that, in the absence of net coupling to the gyron mode and with the consequent suppression of resonant cyclotron absorption and Landau scattering, the combined number (line density) of probe photons and plasmons should be invariant, and in fact we find:

$$\big|\bar{f}(z,t)\big|^2 + |\bar{a}_1(z,t)|^2 = \cos^2\theta_d(z,t)\,|\bar{\chi}_d(z,t)|^2 + \sin^2\theta_d(z,t)\,|\bar{\chi}_d(z,t)|^2 = |\bar{\chi}_d(z,t)|^2\,. \tag{6.159}$$

Local conservation of the sum of bare probe photons and plasmons is just equivalent to conservation of the total number of dressed dark-state quanta, supposing the number of bright-state quanta and gyrons remains negligible.

Entry and Exit of the EM pulse: Establishing the Dark State and Spatial Pulse Compression

In the EIT literature, a great deal of confusion has surrounded issues of whether and how the probe pulse may be compressed by its slow effective group velocity inside the active medium, so let us briefly consider the transient dynamics involved in entry or exit of the probe, still assuming the near-resonant SVEA approximation developed above is essentially valid. In our opinion, much of this confusion has arisen because of insufficient attention paid to the distinctions between bare and dressed modes, between action density and energy density, between density and flux, between spatial and temporal compression, and between advective and conservative flow.

Again, for the moment we effectively imagine that the probe pulse is initially launched in the remote past ($t \sim t_0 \to -\infty$) into an unmagnetized, uniform plasma in some remote upstream region ($z \sim z_0 \to -\infty$), while we suppose a *time-independent* pump envelope $\bar{a}_2(z)$ is established prior to the arrival of the probe, and maintained in essentially a steady-state throughout the traversal of the probe through the plasma. This corresponds to the situation modeled classically by the PIC simulations discussed earlier, and any experiment with a nominally CW pump and large transverse spot sizes might be reasonably approximated by this particular $1D$ geometry. Under these conditions, as the cyclotron frequency Ω_c is also assumed constant and uniform, and the plasma frequency $\omega_p(z)$ is independent of time, the probe wavenumber $k_1(z)$, mixing angle $\theta_d(z)$, and corresponding effective group velocity $v_d(z)$ are all also slowly-varying functions of position z but are independent of time t, while the overall frequency ω_1 of the photon component of the polariton remains constant (i.e., independent of z and t), as does the beat frequency (driven frequency of the longitudinal component). The carrier phase for the transverse DOF can be taken to be

$$\psi_1(z,t) = \int_{z_0}^{z} dz'\, k_1(z') - \omega_1\,(t - t_0), \tag{6.160}$$

where the wavenumber $k_1(z) \geq 0$ satisfies the local EIT dispersion relation

$$\omega_1 = \Omega_c + \tilde{\omega}_d\,(k_1(z); z)\,. \tag{6.161}$$

Ideally, during the entry phase, the pump intensity can be made suitably large (in practice, probably as large as is experimentally feasible at microwave frequencies) in order to maximize the bandwidth of the DSP mode and its overlap with the bare probe photon mode.

Because changes in the medium are assumed to be very gradual, we have seen above that the number of dressed quanta (or the sum of probe photons and plasmons) in the right-traveling probe pulse remains invariant within the SVEA approximation. Actually, WKB theory can work in media which are only piece-wise slowly-varying, if allowance is made for reflection and appropriate matching conditions are enforced at points of discontinuity. But under the various simplifying assumptions made above, if the probe frequency lies in a sufficiently narrow bandwidth near the center of the EIT pass-band, the *phase velocity* for the polariton mode will be real and numerically close to that of the photon mode in the unmagnetized plasma even as the *group velocities* may greatly differ. At any sharp boundary, say at some point z_s, the $1D$ reflection coefficient just depends on the relative difference in the effective index of refraction $\eta(z)$:

$$\mathfrak{R}(z_s) = \frac{|\eta(z_s^+) - \eta(z_s^-)|}{\left|\eta(z_s^+) + \eta(z_s^-)\right|}, \tag{6.162}$$

so reflection losses are expected to be small even if the plasma density jumps discontinuously or varies more rapidly than would seem to be naively allowed under the SVEA, and (continuing to assume the RWA and near resonant conditions) the Manley-Rowe relations associated with our dressed-mode transport equation would continue to hold to good approximation for the *transmitted* quanta in even rather abruptly-changing media.

In an inhomogeneous but time-independent medium, the DSP action transport equation becomes

$$\tfrac{\partial}{\partial t}\left|\tilde{\chi}_d\right|^2 = -\tfrac{\partial}{\partial z}\left[v_d(z)\left|\tilde{\chi}_d\right|^2\right] = -v_d(z)\tfrac{\partial}{\partial z}\left|\tilde{\chi}_d\right|^2 - \left|\tilde{\chi}_d\right|^2 \tfrac{d}{dz}v_d(z), \tag{6.163}$$

where

$$\tfrac{d}{dz}v_d(z) = \sin\big(2\theta_d(z)\big)v_a(z)\tfrac{d}{dz}\theta_d(z) - \sin^2\theta_d(z)\tfrac{c^2}{v_a(z)}\tfrac{\omega_p(z)}{\omega_1^2}\tfrac{d}{dz}\omega_p(z). \tag{6.164}$$

While we generally commend the illuminating analysis of DSPs in atomic EIT presented in [281, 241], we must disagree with their claimed equation-of-motion for the slowly-varying polariton envelope in the case of a time-independent medium and pump because it is of advective rather than conservative form – in this case and in our notation, their equation would read

$$\tfrac{\partial}{\partial t}\left|\tilde{\chi}_d\right|^2 = -v_d(z)\tfrac{\partial}{\partial z}\left|\tilde{\chi}_d\right|^2, \tag{6.165}$$

for which the dressed action is not in general conserved except in the trivial case where θ_d also happens to be independent of the spatial position z. Their derivation follows a different mathematical route and is somewhat tersely summarized, so it is not entirely clear where the discrepancy arises, but it is well known that in an eikonal/WKB asymptotic expansions for linear Hermitian wave equations, the leading-order (in the slowness, or scale-separation parameter) dynamics amount to ray equations (or equivalently by way of the method of characteristics, a Hamilton-Jacobi-type equation) for the fast phase evolution, while at the next order a conservation law governing amplitude transport emerges, so we have some confidence that the envelope transport should be conservative rather than advective.

Given equation (6.163), the solution to the associated Cauchy problem can be written formally in terms of the initial ($t = t_0$) conditions and the local group velocity $v_d(z)$, assumed known. Because the probe is assumed to start out in near vacuum, we take the initial conditions to be

$$\tilde{\chi}(z, t_0) = \tilde{a}(z, t_0) = \tilde{a}\big(z_0, t_0 - \tfrac{z - z_0}{v_d(z_0)}\big) \tag{6.166}$$

which is assumed to vanish for $z > z_0$, and in which we can take $v_d(z_0) \approx c$. At later times, the solution is given by:

$$|\tilde{\chi}_d(z, t)|^2 = \left|\tfrac{v_d(\zeta_0(z,t))}{v_d(z)}\right| \left|\tilde{\chi}_d\big(\zeta_0(z, t), t_0\big)\right|^2, \tag{6.167}$$

where $\zeta_0(z, t)$ solves the implicit equation

$$t = t_0 + \int_{\zeta_0(z,t)}^{z} \frac{dz'}{v_d(z')}, \tag{6.168}$$

and may be interpreted as the initial position of a characteristic, or "ray," currently at position z that has been propagating at the local group velocity. (The coordinate $\zeta_0(z, t)$ for the rays is essentially the inverse of the Lagrangian coordinate for particles introduced above.) The mathematical solution expresses the physical fact that in order to locally conserve action along a flow which is not necessarily divergenceless, the *flux* of polaritons at position z and time t must equal the initial flux at position ζ_0 and time t_0. This form for the solution can be easily verified by noting that $\frac{\partial}{\partial t}\zeta_0(z, t) = -v_d(\zeta_0)$ and $\frac{\partial}{\partial z}\zeta_0(z, t) = -\big(\frac{\partial t}{\partial z}\big)_{\zeta_0} / \big(\frac{\partial t}{\partial \zeta_0}\big)_z = v_d(\zeta_0)/v_d(z)$. Equation (6.167) properly expresses how a *density* should transform under a time-dependent coordinate change between ζ_0 and z, with Jacobian determinant $\frac{|v_d(\zeta_0)|}{|v_d(z)|}$.

From the polariton wave-packet perspective, we see from (6.167) that as the probe pulse enters the boundary region around z_l with increasing $\theta_d(z)$, the total number of polaritons

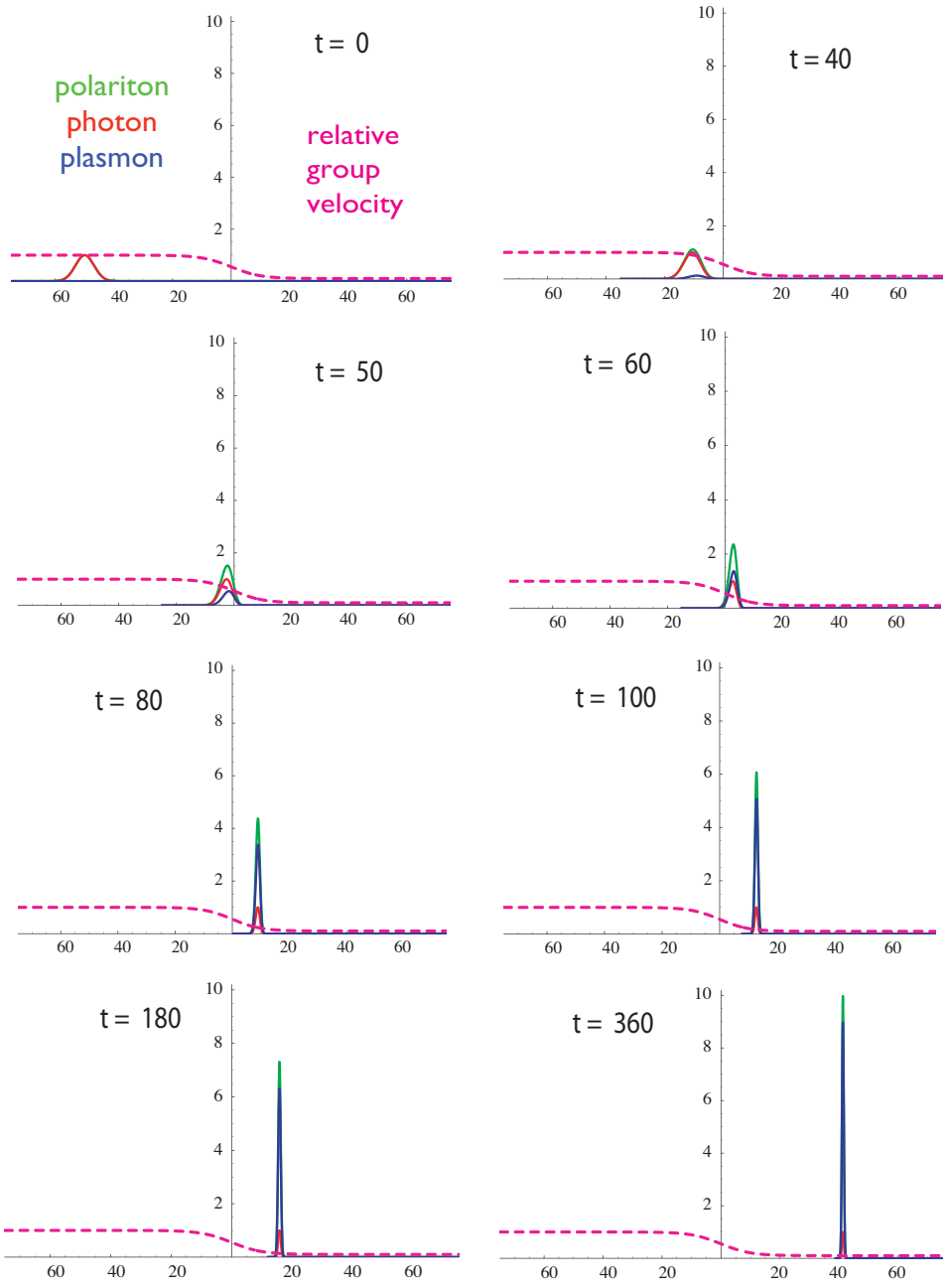

Figure 6.15. Numerical simulation of the pulse evolution through the transition region from vacuum into the plasma. Expected pulse action density (for a coherent state pulse with an initial Gaussian envelope) is plotted as a function of longitudinal velocity at a succession of times. The slow-down and resulting compression is evident, as is the fact that the peak bare photon intensity is essentially unchanged, although the gradient in group velocity has been exaggerated somewhat for effect compared to typical parameter values.

is conserved, but the linear group velocity decreases, so therefore the pulse bunches up in space – the pulse length $\sigma_z(z)$ (as measured, say, by either FWHM or action-weighted RMS spread) decreases, while the average action density $\left\langle |\bar{\chi}_d|^2 \right\rangle$ increases (as does the peak density), just as it would for any conventional wave packet in a linear but slowly-varying medium, so as to keep the average flux approximately constant: $\frac{\partial}{\partial t}\left[\left\langle |\tilde{\chi}_d|^2 \right\rangle \sigma_z\right] \approx 0$. Letting $t_0 \to -\infty$, and supposing that the pulse is sufficiently short compared to medium variations so that both $\sigma_z(z_0)\, \frac{\partial}{\partial z}\omega_p(z)\big|_{z=z_0} \ll \omega_p(z_0)$ initially and $\sigma_z(z_b)\, \frac{\partial}{\partial z}\omega_p(z)\big|_{z=z_b} \ll \omega_p(z_b)$ inside the bulk plasma, we see that at times $t \sim t_b$ when the centroid of the plasma is near z_b,

$$|\tilde{\chi}_d(z,t_b)|^2 \approx \tfrac{c}{v_0(z_b)} \left|\tilde{\chi}_d\left(\tfrac{c}{v_0(z_b)}\left[z - v_d(z_b)\, t - z_0\right], t_0\right)\right|^2, \tag{6.169}$$

so once it reaches the uniform bulk region, the pulse essentially reproduces in compressed (and of course translated) form the relative spatial profile of the initial pulse but with an overall action-preserving Jacobian change of scale, apart from an essentially trivial density-dependent factor involving the bare group velocity $v_0(z_b)$ that is only weakly-varying in the underdense regime.

The local dressed group velocity (6.138) varies weakly with density through $v_0(z_b)$ – which would occur even in the absence of any EIT effects – and also directly with the relative make-up of the dressed mode (also varying with density), as the proportion of photon-like (with high bare group velocity) versus plasmon-like (with zero bare group velocity) excitation in the polariton adjusts in response to the changing plasma frequency $\omega_p(z)$ so as to adiabatically maintain the dark-state. But from the dressed polariton point-of-view, this group velocity reflects the true rate of polariton action transport, and equivalently may be interpreted as the actual velocity of the conserved polariton quasi-particles associated with $\tilde{\chi}_d(z,t)^\dagger$.

In contrast, from the bare probe photon perspective, the total number of probe photons is not conserved, and the shortening of the pulse is interpreted (almost entirely) as a result of photon loss rather than energy or action compression. As the probe enters the EIT region, each photon lost from the leading edge of the probe is converted into one plasmon and one pump photon via 3-wave and/or 2-wave scattering processes. This conversion occurs so as to establish and maintain the local dark state, i.e., effectively minimizing the magnitude of the interaction energy and the involvement of the gyron mode in the vicinity of the cyclotron resonance, and allowing the pulse to "sneak" into the plasma in part as a longitudinal Langmuir oscillation. Contrary to some claims in the literature, individual probe photons neither really slow down nor bunch up due to the EIT effect (that is, via changes in the mixing angle $\theta_d(z,t)$), but only very slightly and rather mundanely

through the weak density dependence of $v_0(z)$ – an effect which is typically negligible in underdense plasmas. Likewise, the peak photon action density (or for that matter, energy density, or intensity) never actually increases through the EIT mechanism *per se* as the pulse propagates through the transition region into the bulk plasma (and changes at all only very slightly due to the density dependence of the bare linear group velocity). From (6.138), (6.157), and (6.167), we see that

$$\begin{aligned} |\bar{a}_1(z,t)|^2 &= \sin^2\theta_d(z)\,|\tilde{\chi}_d(z,t)|^2 = \left|\tfrac{v_d(z)}{v_0(z)}\right| |\tilde{\chi}_d(z)|^2 = \left|\tfrac{v_d(\zeta_0(z,t))}{v_0(z)}\right| \left|\tilde{\chi}_d\big(\zeta_0(z,t),t_0\big)\right|^2 \\ &= \left|\tfrac{v_0(\zeta_0(z,t))}{v_0(z)}\right| \left|\tfrac{v_0(\zeta_0(z,t))}{v_0(z)}\right| \left|\sin\theta_d\big(\zeta_0(z,t),t_0\big)\,\tilde{\chi}_d\big(\zeta_0(z,t),t_0\big)\right|^2 \\ &= \left|\tfrac{v_0(\zeta_0(z,t))}{v_0(z)}\right| \left|\bar{a}_1\big(\zeta_0(z,t),t_0\big)\right|^2, \end{aligned} \tag{6.170}$$

so, apart from the pre-factor $\left|\frac{v_0(\zeta_0)}{v_0(z)}\right| \approx 1$, the photon action density is in fact otherwise *constant* along each ray, while the total photon number decreases as the rays bunch and the volume occupied by photons diminishes due to the decreasing dressed group velocity.

From the undressed photon perspective, this effective slow group velocity instead reflects pulse-reshaping rather than true slowing, where action and energy are carved out of the leading edge and returned to the trailing edge. The effective group velocity $v_d(z)$ is not related to the velocity of individual photons, since their number is not conserved as the pulse propagates through the transition region ($z \sim z_l$) where $v_d(z)$ changes significantly, and even in the bulk plasma ($z \sim z_b$) where $v_d(z)$ is essentially constant, the slow pulse propagation involves the continual destruction and creation (at least virtually) of bare photons as mediated by the interaction Hamiltonian. Using the photon action density, obviously the temporal pulse duration (as calculated by FWHM or action-weighted moments) will agree with that calculated in terms of the full dressed polariton action density, and the bare and dressed spatial pulse lengths will also agree if the pulse is sufficiently short or the density sufficiently slowly-varying such that the effective group velocity is approximately constant over the entire spatial extent of the pulse. In [255], the authors refer to the "strictly geometrical" interpretation of the effective group velocity, in the sense that at early (pre-transition) and late (post-transition) times, when the wave packet may be taken to be in the neighborhoods, respectively, of z_0 in a homogeneous but rarefied region and z_b in the homogeneous bulk plasma region, the pulse lengths will satisfy $\frac{\sigma_z(z_b)}{\sigma_z(z_0)} \approx \frac{v_d(z_b)}{v_0}$.

From the bare plasmon perspective, quasi-particle number is obviously not conserved locally or globally. Rather, both total plasmon number and plasmon action density increase as the probe enters the EIT region. From (6.157), we see that

$$\left|\bar{f}(z,t)\right|^2 = \cos^2\theta_d(z)\,|\tilde{\chi}_d(z,t)|^2 = \cot^2\theta_d(z)\left|\bar{a}_1\big(z,t\big)\right|^2, \tag{6.171}$$

so the proportion of plasmons increases at the expense of photons as the pulse propagates through the transition region where $v_d(z)$ is decreasing. Bare plasmons generated by scattering processes at the leading edge of the probe establish the dark state and locally increase the plasmon action density, but these undressed plasmons have zero bare group velocity (in the assumed cold limit) and do not actually propagate anywhere, but persist (or rather are virtually created and destroyed continuously according to the interaction Hamiltonian) only so long as the remainder of the pulse passes over them, and then their action and energy is finally returned to the trailing edge of the pulse via inverse scattering processes in the presence of the pump. As with the bare photons, the effective group velocity $v_d(z)$ reflects the net effects of mode-conversion and pulse-reshaping, not the actual velocity of any individual bare plasmons. As the pulse travels, the plasmon components of the probe at widely separated spatial locations are effectively comprised of different bare plasmons, to the extent such a description is meaningful for indistinguishable bosons. At any instant in time, if the width of the dressed pulse is sufficiently small compared to the scale of slow variations in $\theta_d(z)$, then the overall width of the bare plasmon component will be the same as in the full polariton pulse.

Neglecting any reflection or other non-adiabatic losses, the Manley-Rowe relations guarantee that the total number of dressed polariton quanta will remain equal to the initial number of photon quanta in the probe, so in this sense the conversion efficiency between bare and dressed action will be $\eta_{\rm S} \sim 100\%$.

Energy transport inside the EIT medium is somewhat more subtle. Total energy will be invariant under the action of the full coupled Hamiltonian for fields and plasma, but we have sacrificed complete self-consistency and replaced the pump with a prescribed c-number field, so energy conservation is no longer manifest. From the perspective of just the dynamically-evolved DOFs in the probe pulse, the "system" is open, with the pump able to act as a source or sink for energy. In fact, as the probe propagates through the transition region, and its character transforms via scattering processes from more photon-like to more plasmon-like, the energy difference (determined by the Planck-Einstein relations to be $\sim \hbar(\omega_1 - \omega_p)$ per quanta) between these bare modes is of course taken up by the pump field. The energy loaned to one part of the pump as the probe enters the EIT region can be reclaimed from (in general another part of) the pump as the probe exits.

Because the plasma medium and the slowly-varying pump envelope are here both assumed to be stationary (time-independent), WKB theory predicts that all carrier frequencies will remain fixed, the local wavenumbers adjusting accordingly. From the dressed polari-

ton perspective, because dressed action is also conserved, this implies that a pseudo-energy $\int dz, \hbar\tilde{\omega}_d(z,t)\,|\tilde{\chi}_d(z,t)|^2$ is conserved as well, but this is not related in any obvious way to the total energy or the energy of any particular subset of the dynamical DOFs. Instead, the per-quanta energy contained within the probe is better thought of as the action-weighted average $\hbar\omega_1 \sin^2\theta_d + \hbar\omega_p \cos^2\theta_d$, so the energetic efficiency during the entry phase of the probe propagation is roughly $\eta_{\rm E} \sim \sin^2\theta_d(z_b) + \frac{\omega_p}{\omega_1}\cos^2\theta_d(z_b) = \frac{\omega_p}{\omega_1} + (1-\frac{\omega_p}{\omega_1})\sin^2\theta_d(z_b)$, and may be significantly less than unity. Again, however, in the ideal case this energy loss is completely reversible, and can be returned in total to the pulse at exit via coherent scattering interactions with the pump.

In addition to transport of action and energy, it is also illuminating to consider the transport of information (both classical and quantum), which is contained in both the amplitude and phase of the probe envelope. Following the pioneering insights of Shannon, information may be defined as that which resolves uncertainty, in this case uncertainty about outcomes of possible measurements on the probe DOFs, and is ultimately encoded or represented in the quantum state (density matrix) of the excitations. For Bosonic field theories, the same information may be represented by the Glauber-Sudarshan P function, which is a real-valued but possibly singular and/or non-positive quasi-distribution function, which for all but the most pathological quantum states is in turn uniquely determined by the set of all normally-ordered moments, or coherence functions, certain of which may be directly related to direct counting experiments.

Some snapshots of a numerical simulation of the probe's entry into the plasma is shown in Fig. 6.15. We consider the classical limit, where the pulse is described by a coherent state, and assume the envelope is originally described by a Gaussian. The expected action density is plotted as a function of space at a few different times, and clearly shows the slow-down and resulting compression as the pulse enters the plasma from vacuum. Note that the peak expected bare photon density hardly changes even as the pulse narrows, but the dressed polariton pulse becomes more plasmon-like as it travels down the group velocity gradient, which has been exaggerated beyond what is encountered for typical parameters for the purposes of illustration. The pulse is clearly narrowing in length at a given time, but not in the duration required to pass any fixed point in space.

As we have mentioned in passing "quantum" and "classical" information, perhaps we should briefly digress to address what might be meant by this distinction in the present context, as a certain measure of confusion persists. Because different possible observables may not commute, and because in general there is no unique way to decompose a given density

matrix into a classical ignorance mixture, information represented in a quantum state cannot in general be definitively and unambiguously decomposed into classical and quantum components. In certain states (namely, those with non-negative, normalizable P-functions, which therefore look like standard probability densities), normally-ordered measurement statistics can be completely and exactly reproduced by classical models, whereas all other states will exhibit at least some statistical features (that can be revealed by appropriate measurements), such as sub-Poissonian statistics or various types of squeezing, or Bell-type correlations arising from quantum entanglement, that apparently cannot be mimicked by any classical model. Because in classical physics measurements of the amplitude and phase of a wave are not limited by the Heisenberg uncertainty principle, it is often said that first-order moments (expectation values of the mode creation or annihilation operators) represent the classical aspect of Bosonic fields, while the higher-order moments represent the "quantum fluctuations" about these classical values. Strictly speaking, this is incorrect on two counts, as the non-zero higher-order moments represent statistical uncertainties rather than actual physical fluctuations, and these uncertainties may have their origin in classical ignorance as well as unavoidable quantum "fuzziness." Still, this distinction between average spatio-temporal envelopes and the remainder of information in the full quantum state can be a useful distinction in practice, just one that perhaps should be differently labeled. Other sources suggest a division between the first two orders of moments and higher-order ones, because it turns out that for Bosonic fields and normally-ordered observables, any consistent set of first and second moments (i.e., corresponding to a positive semi-definite covariance matrix) can be reproduced within a classical stochastic model, which is just to say that if one is performing quanta counting observations, at minimum two-quanta measurements (separated in time and/or space) are needed to reveal the kinds of correlations which are the signature of intrinsically non-classical states.

Rather than attempting to introduce some complete but necessarily arbitrary distinction in issues of the conversion or propagation of information, we will take refuge in a certain ambiguity, but describe as quantum mechanical those measurable aspects or features of the states or observables – predominately entanglement, or squeezing in quadrature components, number, phase, or observables – that definitely have no classical analogs or description.

Of course, the overall dynamics of the probe propagation are unitary, and therefore trivially conserve the von Neumann entropy, and indeed more generally all of the Hioe-Eberly trace invariants of the full density matrix for the combined field-matter system (i.e.,

traces of powers of the matrix), but besides the net amount of information, the specific form of the signal information is preserved in a more direct sense as the probe pulse propagates.

So far, we have mainly considered quadratic SVEA observables such as the DSP number (density) operator $|\bar{\chi}_d|^2 = \bar{\chi}_d^\dagger \bar{\chi}_d$, but now we need to examine directly the evolution of the actual envelope creation or annihilation operators. From (6.146) and (6.161), we have

$$\tfrac{\partial}{\partial t}\tilde{\chi}_d(z;t) = ik_1(z,t)v_d(x,t)\tilde{\chi}_d(z;t) - v_d(z,t)\tfrac{\partial}{\partial z}\tilde{\chi}_d(z;t) - \tfrac{1}{2}\tilde{\chi}_d(z;t)\tfrac{\partial}{\partial z}v_d(z,t). \tag{6.172}$$

Writing $\tilde{\chi}_d(z,t) = |\tilde{\chi}_d(z,t)|\, e^{i\psi_d(z,t)}$ in terms of a normally-ordered amplitude and real phase, and using (6.146), a straightforward calculation reveals that

$$\tfrac{\partial}{\partial t}\psi_d(z,t) = k_1(z,t)v_d(z) - v_d(z)\tfrac{\partial}{\partial z}\psi_d(z,t), \tag{6.173}$$

which can be formally integrated to obtain

$$\psi_d(z,t) = \int_{z_0}^{z} dz'\, k_1(z') + \tilde{\psi}_d\left(\zeta_0(z,t), t_0\right), \tag{6.174}$$

for some as yet undetermined function $\tilde{\psi}_d(z,t_0)$. The first term represents the fast carrier phase (at least for the photonic component of the DSP), while choosing the second term to match initial conditions leads to the simple result

$$\bar{\chi}_d(z,t) = \left|\tfrac{v_0(\zeta_0(z,t))}{v_0(z)}\right|^{1/2} \bar{\chi}_d\big(\zeta_0(z,t),t_0\big) = \left|\tfrac{v_0(\zeta_0(z,t))}{v_0(z)}\right|^{1/2} \bar{a}_1\big(\zeta_0(z,t),t_0\big), \tag{6.175}$$

such that $\bar{\chi}_d(z,t)$ can be written as a particular "square root" of the SVEA density (6.167), in which with our assumed geometry and initial conditions we may also equate $\bar{\chi}_d\left(z,t_0\right) = \bar{a}_1\left(z,t_0\right)$ if needed.

Although it is not immediately obvious, this time advance is in fact unitary, as it must be, preserving the equal-time CCRs:

$$\begin{aligned}
\Big[\bar{\chi}_d(z,t), \bar{\chi}_d(z',t)^\dagger\Big] &= \left|\tfrac{v_d(\zeta_0(z,t))}{v_d\theta(z)}\tfrac{v_d(\zeta_0(z',t))}{v_d(z')}\right|^{1/2} \Big[\bar{\chi}_d\big(\zeta_0(z,t),t_0\big), \bar{\chi}_d\big(\zeta_0(z',t),t_0\big)^\dagger\Big] \\
&= \left|\tfrac{v_d(\zeta_0(z,t))}{v_d(z)}\right| \delta\left(\zeta_0(z,t) - \zeta_0(z',t)\right) \\
&= \left|\tfrac{\partial\zeta_0(z,t)}{\partial z}\right| \delta\left(\zeta_0(z,t) - \zeta_0(z',t)\right) = \delta(z-z'),
\end{aligned} \tag{6.176}$$

where we made use of the fact that the Dirac delta function transforms as a density under a change of variables.

Because all frequencies are conserved in the time-independent medium, we expect that the temporal profile (or spectra, or action-weighted temporal moments, etc.) as measured

at any fixed position will actually be independent of z, apart from an overall scaling in amplitude depending on the local group velocity. Specifically, with a simple change of variables, the DSP envelope "power" spectrum may be written explicitly as

$$|v_d(z)|\left|\int dt\, e^{i\omega t}\bar{\chi}_d(z,t)\right|^2 = \iint d\zeta_0 d\zeta_0' \,\frac{\bar{\chi}_d(\zeta_0',t_0)^\dagger \bar{\chi}_d(\zeta_0,t_0)}{|v_d(\zeta_0')v_d(\zeta_0)|^{1/2}} e^{i\omega\int_{\zeta_0}^{\zeta_0'}\frac{dz'}{v_d(z')}}, \tag{6.177}$$

which is independent of the position of observation. In particular, the temporal duration $\sigma_t(z) \sim \sigma_\omega(z)^{-1}$ of the pulse, as measured either by FWHM or action-weighted moments, is completely independent of the position at which it is measured. As the probe propagates through a time-independent density gradient at a fixed pump strength, compression occurs spatially but not temporally.

More generally, we can see that all normalized and normally-ordered coherence functions essentially preserve their form under this evolution, apart from the overall spatial compression. Defining the (ℓ,ℓ')th-order normalized DSP spatio-temporal coherence function as

$$\bar{\gamma}_d(z_1,t_1;\ldots;z_\ell,t_\ell;z_1',t_1';\ldots;z_{\ell'}',t_{\ell'}') \equiv \frac{\left\langle \prod_{j'=1}^{\ell}\bar{\chi}_d(z_{j'}',t_{j'}')^\dagger \prod_{j=1}^{\ell}\bar{\chi}_d(z_j,t_j)\right\rangle}{\prod_{j'=1}^{\ell'}\left\langle\bar{\chi}_d(z_{j'}',t_{j'}')^\dagger\bar{\chi}_d(z_{j'}',t_{j'}')\right\rangle^{\frac{1}{2}}\prod_{j=1}^{\ell}\left\langle\bar{\chi}_d(z_j,t_j)^\dagger\bar{\chi}_d(z_j,t_j)\right\rangle^{\frac{1}{2}}} \tag{6.178}$$

with similar definitions made in an analogous manner for the normalized probe photon coherence function $\bar{\gamma}_a$ or plasmon coherence function $\bar{\gamma}_f$, and using (6.155) and (6.175), we see that

$$\begin{aligned}
\bar{\gamma}_a(z_1,t_1;\ldots;z_\ell,t_\ell;z_1',t_1';\ldots;z_{\ell'}',t_{\ell'}') &= \bar{\gamma}_f(z_1,t_1;\ldots;z_\ell,t_\ell;z_1',t_1';\ldots;z_{\ell'}',t_{\ell'}') \\
&= \bar{\gamma}_d(z_1,t_1;\ldots;z_\ell,t_\ell;z_1',t_1';\ldots;z_{\ell'}',t_{\ell'}') \\
&= \bar{\gamma}_d\left(\zeta_0(z_1,t_1),t_0;\ldots;\zeta_0(z_\ell,t_\ell),t_0;\zeta_0(z_1',t_1'),t_0;\ldots;\zeta_0(z_{\ell'}',t_{\ell'}'),t_0\right) \\
&= \bar{\gamma}_a\left(\zeta_0(z_1,t_1),t_0;\ldots;\zeta_0(z_\ell,t_\ell),t_0;\zeta_0(z_1',t_1'),t_0;\ldots;\zeta_0(z_{\ell'}',t_{\ell'}'),t_0\right) \\
&= \bar{\gamma}_f\left(\zeta_0(z_1,t_1),t_0;\ldots;\zeta_0(z_\ell,t_\ell),t_0;\zeta_0(z_1',t_1'),t_0;\ldots;\zeta_0(z_{\ell'}',t_{\ell'}'),t_0\right).
\end{aligned} \tag{6.179}$$

Normalized space-time coherences and correlations are actually invariant apart from propagation along rays. Analogous constructions may be established in the full frequency domain as well. All the information in the initial probe photon pulse just propagates undisturbed in dressed form along the (classical-looking) rays. At subsequent times, the information could be extracted by suitable direct measurement of the polariton field observables, if we knew how to do that, or more realistically by measurement of either the photonic or plasmonic

components, which subsequently both contain essentially the same information. Of course, this apparent doubling of the information does not violate unitarity, because the information in the longitudinal and transverse DOFs is perfectly correlated (neglecting damping, noise, or decoherence mechanisms). (If we were to evolve the pump self-consistently, the information would be embedded there as well, but as it stands the back-reaction on the pump is neglected).

Any squeezing or entanglement in the initial quantum state of the photon pulse is coherently transferred into the polaritons (and hence into the material, or plasmonic component thereof), just in a spatially compressed form. In a Schrödinger-like picture (with operators coinciding with their Heisenberg-picture counterparts at $t = t_0$), any initial pure state of the form

$$|\psi(t_0)\rangle \propto \prod_j \left\{ \bar{a}_1(\zeta_j, t_0)^\dagger \right\}^{\ell_j} |0\rangle \tag{6.180}$$

for some positive integers $\ell_j \in \mathbb{Z}^+$ and initial positions $\zeta_j \in \mathbb{R}$, evolves into

$$|\psi(t)\rangle \propto \prod_j \left\{ \bar{\chi}_d \left(Z(\zeta_j, t), t_0 \right)^\dagger \right\}^{\ell_j} |0\rangle \,, \tag{6.181}$$

where here $|0\rangle$ is the effective vacuum state (for excitations, not for the total number of plasma electrons), and $Z(\zeta, t)$ is a forward time-advance function satisfying $\zeta_0 \left(Z(\zeta, t), t \right) = \zeta$ for all $\zeta \in \mathbb{R}$ and $t \geq t_0$. Since all such states span the Fock spaces for photonic or polariton excitations, respectively, it follows that any particular pure or mixed initial quantum state of the probe photon field (including number, coherent, squeezed, or chaotic states, etc.) will be converted into a corresponding state of the polariton field with corresponding statistics and correlations modulo the classical compression of the rays.

We should confirm that such an injection process is actually consistent with our various physical and numerical assumptions. Some of the relevant parameters are plotted as a function of distance into the leading edge of the plasma in Fig. 6.16 for a typical case. With plane-wave pump of fixed intensity and frequency $\omega_2 = \Omega_c - \omega_p(z_b)$, the group velocity for a sufficiently narrow-band probe pulse of carrier frequency $\omega_1 \approx \Omega_C$ drops smoothly as the density rises to its bulk value. We might worry that before the transparency can be established, the pulse can be absorbed within the conventional cyclotron stop-band, but the width of this non-EIT frequency stop-band vanishes as $\omega_p(z) \to 0$, while the total EIT frequency pass-band only grows monotonically in this limit. The central carrier frequency ω_1 and the frequency bandwidth $\Delta\omega_1$ of the probe remain fixed during this stage, so in the frequency domain, propagation into the bulk requires that the initial probe frequencies can

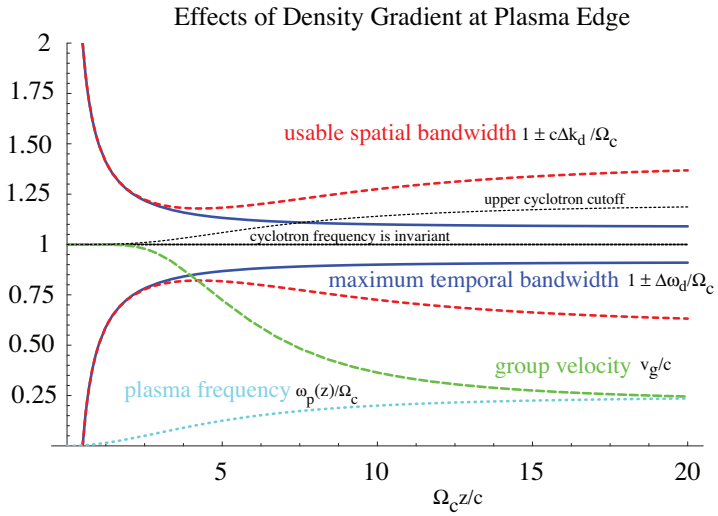

Figure 6.16. Plot of the available temporal and spatial bandwidths of the DSP mode, along with the local group velocity and plasma frequency, as a function of position of the probe pulse peak into the leading edge of the magnetized plasma, where we have assumed a gradual spatial density gradient, but no spatial variation, for a probe wave-packet with fixed carrier frequency $\omega_1 = \Omega_c$, and a plane-wave pump field of strength $|a_{\rm pump}| = 0.1$, and fixed frequency $\omega_2 = \Omega_c - \omega_p(z_b)$, where $\omega_p(z_b) = 0.25\Omega_c$. The full spatial bandwidth of the DSP pass-band is infinite, but the "usable" part refers to the region over which the dispersion relation can be approximately linearized, leading to pulse propagation without significant group velocity dispersion.

be accommodated within the polariton bandwidth $\Omega_c \pm \Delta\omega_d(z_b)$ in the bulk:

$$\Omega_c - \Delta\omega_d(z_b) < \omega_2 - \Delta\omega_2 \leq \omega_2 + \Delta\omega_2 < \Delta\omega_d(z_b). \tag{6.182}$$

The full spatial bandwidth Δk_d of the polariton pass-band is of course infinite, but the pulse should be limited to an effective, or useful bandwidth $\Delta\tilde{k}_d$, over which the dispersion relation can be linearized and the envelope can propagate without significant group velocity dispersion, which is to say about

$$\Delta k_1(z) \lesssim \Delta\tilde{k}_d \sim \epsilon_d \frac{\Delta\omega_d(z)}{v_d(z)}, \tag{6.183}$$

where $\epsilon_d \lesssim O(1)$ is some numerical factor (whose weak variation with $\omega_p(z)$ has been neglected). Now, as the pulse propagates up the density gradient, the pulse length $\sigma_{z\,1}(z)$ decreases, so the conjugate bandwidth $\Delta k_1(z) \sim \sigma_{z\,1}(z)^{-1}$ of the probe increases, which might appear troublesome, except that the required bandwidth $\Delta k_1(z)$ also increases with decreasing velocity by the same factor of $v_d(z)^{-1}$ as does the available EIT spatial bandwidth $\Delta\tilde{k}_d$, so the SVEA can remain valid as long as

$$v_d(z_0)\Delta k_1(z_0) \lesssim \epsilon_d \Delta\omega_d(z) \leq \epsilon_d \Delta\omega_d(z_b). \tag{6.184}$$

Since $v_d(z_0)\Delta k_1(z_0) \sim c\Delta k_1(z_0) \sim \Delta\omega_1$, this constraint is not much more severe than that already deduced in the frequency domain.

The photon intensity never increases past its initial peak value, so if $|a_1(z,t_0)| \ll |a_2|$ at all positions z, then $|a_1(z,t)| \ll |a_2|$ at all positions and for all subsequent times during pulse entry.

In order to ensure that the neglect of pump depletion is valid in all the various scattering processes, one might argue that the density of plasmons should also be much less than the density of pump photons, requiring the stronger condition

$$|\cot\theta_d(z_b)| \max_z[|a_1(z,t_0)|] \ll |a_2|\,. \tag{6.185}$$

As $|a_2| \to \infty$, $\cot\theta_d(z_b) \to 0$, so this is always possible for a sufficiently intense pump. However, as $|a_2| \to 0, \cot\theta_d(z_b) \sim a_2^{-1}$, so the initial peak probe intensity must decrease faster then the pump (as the square of the pump intensity) in order to maintain the validity of the prescribed-pump approximation.

One could make this argument, but we had better make a different one, if the validity of the prescribed pump approximation is to hold during the subsequent stage of envelope dynamics (to be discussed shortly), where the pump intensity can be tuned down or even

turned off completely. Basically, we suggest that the relative magnitudes of plasma wave and pump actions are largely irrelevant to the validity of the underlying assumption of a classical prescribed pump field, and smallness of the probe alone is necessary and sufficient. The apparent symmetry between the transverse (photon or gyron) and longitudinal (plasmon) envelopes in the 3-wave scattering terms is broken by their frequency difference, with the probe quanta privileged by virtue of being of higher energy than either pump photons or plasmons. If the density of probe photons is comparable to or exceeds the density of the lower-frequency pump photons, then forward scattering processes could spontaneously generate a significant change in the local value of the latter for essentially any non-zero value of the plasmon density, as governed by the Manley-Rowe bookkeeping. Conversely, if the probe intensity can be guaranteed to always remain negligible compared to the pump intensity, then forward scattering further depleting the already small number of probe photons cannot appreciably affect the pump, and inverse scattering events converting pump and plasma-wave quanta into probe quanta could not progress for long before invalidating the ordering assumption between pump and probe intensities.

We admit that this line of reasoning is not air-tight, because by prescribing the pump we may be somehow under-counting the quantum noise that is deposited back in the probe. Here, we simply leave this as an open question, and carry on with our envelope analysis treating only the probe and plasma wave as quantum mechanical. The evolution of bosonic modes where one or more is initially described in terms of Glauber coherent states (which are in a sense maximally classical) has been studied extensively for 2-wave couplings, but not to our knowledge for 3-wave processes.

Pump-Mediated Polariton Tuning: Adiabatic Transfer within the Magnetized Plasma

With a sufficiently large initial pump strength to accommodate the initial frequency bandwidth of the probe, the actual amount of mode conversion between photons and plasmons occurring throughout the entry of the probe pulse from vacuum generally remains relatively small. But once the probe pulse has propagated in the presence of a suitably-intense pump through the transition region into the bulk plasma region with more-or-less uniform density (i.e., in the region $z \gtrsim z_b$ for, say, times $t \gtrsim t_b$), the pump strength $|a_2(z,t)|$ can then be adjusted adiabatically over a broad range in order to tune the modal character and corresponding group velocity of the polariton wave-packet between a more photon-like or more plasmon-like excitation.

Now assuming propagation within a region with constant $\omega_p = \omega_p(z_b)$ and still constant Ω_c, the polariton mixing angle θ_d will vary only through changes in the envelope of the classical pump field, such that $\theta_d(z,t) = \theta_d\big(z - v_2[t-t_b], t_b\big)$, where the constant $v_2 = v_2(z_b)$ is the group velocity of the pump field (evaluated at its carrier wavenumber k_2 and frequency ω_2 for fixed ω_p and Ω_c), determining the speed at which (slow) variations imposed on the envelope of the pump field at its remote source (at $z \sim z_\infty$) now propagate through the bulk plasma.

By means of an appropriate coordinate transformation, a formal solution to the continuity equation (6.147) can also be written down in this case, but its somewhat complicated form tends to obscure the essential physics, so we will make additional assumptions so as to simplify matters further. Since the probe pulse has undergone spatial compression in propagating into the plasma, we suppose that the pulse length $\sigma_z(z_b)$ is now sufficiently small, and also suppose that temporal variations $\left|\bar{a}_2^{-1}\frac{\partial}{\partial t}\bar{a}_2\right|$ introduced in the pump envelope are sufficiently slow, yet assume that the pump group velocity v_2 is not too small, such that over the subsequent spatial extent $\sigma_z(t)$ of the dressed probe, the plasma density ω_p can be assumed constant, and also the pump strength $|\bar{a}_2(t)|$ may be taken to be independent of position. As a result, the associated polariton mixing angle $\theta_d(t)$ and group velocity $v_d(t)$ may also be regarded as z-independent, varying only in time through slow changes in the applied pump envelope.

In this case, the probe carrier wavenumber $k_1 = k_1(z_b)$ will be constant, equal to the previously spatially-varying wavenumber evaluated at $z = z_b$, while the carrier frequency $\omega_1(t)$ will be a slowly-varying function of time, satisfying the local dispersion relation

$$\omega_1(t) = \Omega_c + \tilde{\omega}_d(k_1; t), \tag{6.186}$$

and in particular $\omega_1(t = t_b) = \omega_1(z = z_0)$, so the carrier phase for the transverse (photonic) component of the DSP may be taken to be

$$\psi_1(z,t) = \psi_1(z_b, t_b) + k_1\,[z - z_b] - \int_{t_b}^{t} dt'\,\omega_1(t'), \tag{6.187}$$

for times $t \geq t_b$ (still assuming a sufficiently narrow pulse relative to the uniform central plasma region), in which

$$\psi_1(z_b, t_b) = \int_{z_0}^{z_b} dz'\,k_1(z') - \omega_1(z_b)\,[t_b - t_0] \tag{6.188}$$

is the full carrier phase as determined from the entry dynamics and evaluated at $z = z_b$ and $t = t_b$, ensuring continuity with our previous phase conventions.

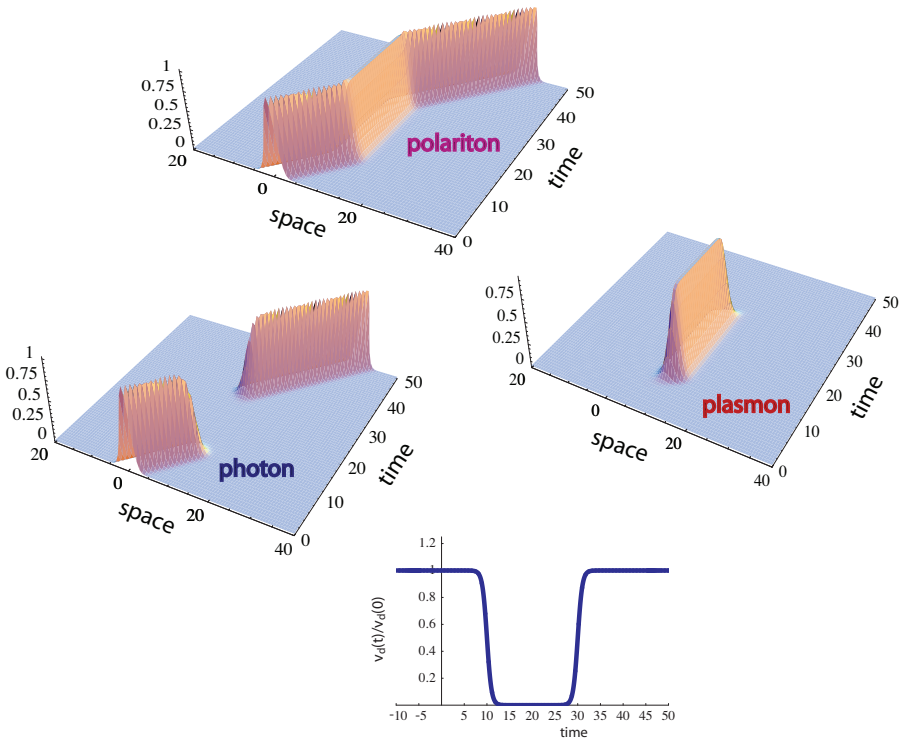

Figure 6.17. Numerical simulations of SVEA probe pulse propagation inside the magnetized plasma while the pump strength is varied adiabatically, showing expected action densities as a function of space and time. As the pump is decreased, the probe becomes mostly plasmonic in character and slows down accordingly. Slowly returning the pump to a higher value reverses the adiabatic conversion.

In this case, the envelope equation of motion (6.146) does reduce to a simple advection equation, namely

$$\frac{\partial}{\partial t}\tilde{\chi}_d(z;t) = i\, v_d(t) k_1(t) \tilde{\chi}_d(z;t) - v_d(t) \frac{\partial}{\partial z}\tilde{\chi}_d(z;t). \tag{6.189}$$

With "initial" ($t = t_b$) conditions chosen to match $\tilde{\chi}_d(z, t_b)$ from the previously-analyzed entry phase of propagation, the solution can be written

$$\tilde{\chi}_d(z,t) = e^{ik_1[z-\zeta_b(z,t)]} \tilde{\chi}_d\left(\zeta_b(z,t), t_b\right), \tag{6.190}$$

where $\zeta_b(z,t) = z - \int_{t_b}^{t} dt'\, v_d(t')$, implying that $\bar{\chi}_d\left(z,t\right) = \bar{\chi}_d\big(\zeta_b(z,t), t_b\big)$. The SVEA envelope $\bar{\chi}_d(z,t)$ propagates without change along classical characteristics, only now the rays converge or diverge in time rather than longitudinal position. Some of our previous SVEA analysis will hold only with the roles of time and space swapped.

Slow changes in the mixing angle $\theta_d(t)$ controlled by the magnitude $|\bar{a}_2(t)|$ of the pump envelope can be used to tune the modal character and corresponding group velocity of the probe pulse. As the pump intensity is slowly increased ($|a_2(t)| \to \infty$) the probe becomes predominately photonic, with increasing velocity approaching $v_0(z_b)$, while if the pump intensity is reduced ($|a_2(t)| \to 0^+$), its transverse contribution is correspondingly reduced, and it becomes increasingly plasmonic with ever slower group velocity $v_d(t) \to 0$.

The ratio of photonic to plasmonic action in the DSP is given by the same factor $\tan^2\theta_d(t) = \frac{v_d(t)}{v_0(z_b)-v_d(t)}$ everywhere in space. Neither the total number of bare photons, $\sin^2\theta_d(t)\int dz'\,|\bar{\chi}_d(z,t)|^2$, nor the number of plasmons in the probe, $\cos^2\theta_d(t)\int dz'\,|\bar{\chi}_d(z,t)|^2$, are separately conserved, but their sum is, because the polariton action density satisfies a continuity equation

$$\tfrac{\partial}{\partial t}\,|\bar{\chi}_d(z,t)|^2 + \tfrac{\partial}{\partial z}\Big[v_d(t)\,|\bar{\chi}_d(z,t)|^2\Big] = 0, \tag{6.191}$$

that indicates that the number of polaritons is conserved locally and globally throughout the adiabatic conversion. Neither the total number of photons nor the peak photon density can ever exceed the corresponding "initial" ($t = t_b$) values for the dressed polaritons.

Because of the time-dependence in the pump amplitude and the probe carrier frequency, neither probe energy nor any "pseudo-energy" is conserved, but rather energy is extracted from the pump as its intensity $|a_2(t)|^2$ is increased (by external means), or transferred to the pump as $|a_2(t)|^2$ is reduced. Any changes in the probe energy can in principle be reversed by slowly returning the pump intensity to its former value.

Apart from an overall translation given by the integrated group velocity, the overall spatial shape and amplitude of the dressed pulse are preserved. Because $\frac{\partial}{\partial z} z_b(z,t) = 1$, it follows for any integrable function $G(z)$ that

$$\int dz\, G(z)\,\bar{\chi}_d(z,t) = \int d\zeta_b\, G\Big(\zeta_b + \int_{t_b}^{t} dt'\, v_d(t')\Big)\bar{\chi}_d(\zeta_b,t_b) \tag{6.192a}$$

$$\int dz\, G(z)\,\bar{\chi}_d(z,t)^\dagger\bar{\chi}_d(z,t) = \int d\zeta_b\, G\Big(\zeta_b + \int_{t_b}^{t} dt'\, v_d(t')\Big)\bar{\chi}_d(\zeta_b,t_b)^\dagger\bar{\chi}_d(\zeta_b,t_b), \tag{6.192b}$$

so by choosing $G(z)$ to be either a complex exponential or a monomial, we see in particular that the *spatial* Fourier transforms are invariant in time apart from an overall translation, and spatial "power" spectra as well as RMS spatial spread or any higher-order normalized, centered, action-weighted moments are fully invariant. The bare photonic or plasmonic components share the same *relative* spatial profile as the polariton pulse, just with an overall change in amplitude depending on $\theta_d(t)$ which does not affect the spatial width $\sigma_z(t) \sim \sigma_k(t)^{-1}$ as calculated by FWHM, action-weighted moments, or other means.

Because the overall scale-factors involving the mixing angle cancel, it is easy to see that the normalized coherence functions satisfy

$$\begin{aligned}\bar{\gamma}_a(z_1,t_1;\ldots;z_\ell,t_\ell;z'_1,t'_1;\ldots;z'_{\ell'},t'_{\ell'}) &= \bar{\gamma}_f(z_1,t_1;\ldots;z_\ell,t_\ell;z'_1,t'_1;\ldots;z'_{\ell'},t'_{\ell'}) \\ &= \bar{\gamma}_d(z_1,t_1;\ldots;z_\ell,t_\ell;z'_1,t'_1;\ldots;z'_{\ell'},t'_{\ell'})\end{aligned} \tag{6.193}$$

so that the dressed polariton field and each of its bare components contain the same information in essentially the same form, while

$$\begin{aligned}\bar{\gamma}_a(z_1,t_1;\ldots;z_\ell,t_\ell;z'_1,t'_1;\ldots;z'_{\ell'},t'_{\ell'}) &= \bar{\gamma}_d(z_1,t_1;\ldots;z_\ell,t_\ell;z'_1,t'_1;\ldots;z'_{\ell'},t'_{\ell'}) \\ &= \bar{\gamma}_d\left(\zeta_b(z_1,t_1),t_b;\ldots;\zeta_b(z_\ell,t_\ell),t_b;\zeta_b(z'_1,t'_1),t_b;\ldots;\zeta_b(z'_{\ell'},t'_{\ell'}),t_b\right) \\ &= \bar{\gamma}_a\left(\zeta_b(z_1,t_1),t_b;\ldots;\zeta_b(z_\ell,t_\ell),t_b;\zeta_b(z'_1,t'_1),t_b;\ldots;\zeta_b(z'_{\ell'},t'_{\ell'}),t_b\right),\end{aligned} \tag{6.194}$$

so the form of this information in the pulse is preserved under the propagation in the presence of adiabatic pump tuning, apart from an overall spatial translation and possible temporal compression or divergence along the classical rays, depending on the time history of $\bar{a}_2(t)$. In principle, not just the average spatio-temporal profile but the full quantum state are adiabatically transferred between the transverse EM field and the longitudinal plasma excitation as $\bar{a}_2(t)$ is slowly changed. In particular, as the pump is slowly turned off ($\bar{a}_2(t) \to 0$), any squeezing of the photon field, or entanglement amongst the photons in the probe pulse or even between the probe photons and any other system is in principle coherently mapped into excitations of the plasma electrons. The conversion may be reversed and the information transported out of the plasma by subsequently increasing the pump intensity.

While the spatial width (or equivalently, spatial bandwidth) remains independent of the moment in time at which it is observed, the temporal duration (or its reciprocal, the temporal bandwidth) as measured at any fixed position varies with changes in the applied pump. For definiteness, suppose the pump intensity is slowly changed from $|a_2(t_b)|^2$ at $t = t_b$ to $|a_2(t_2)|^2$ at some time $t_2 \gg t_b$, and then held fixed, after which the probe is measured at some downstream position $z_2 \gg z_b$ where the amplitude is negligible for $t < t_2$. Then, the temporal duration will be approximately

$$\sigma_t(z_2) \approx \tfrac{\sigma_z(z_b)}{v_d(t_2)}, \tag{6.195}$$

according to either an RMS or FWHM definition for both spatial and temporal quantities.

An example of this adiabatic mapping between transverse and longitudinal excitation is shown in Fig. 6.17. Some average action density profiles of a numerical simulation of probe pulse propagation within the EIT medium while the pump field is slowly varied are shown

as a function of space and time. Again a classical coherent state with an initially Gaussian profile is assumed. With the probe residing fully within the homogeneous plasma region, the time-dependent pump envelope is slowly turned down in strength, maintained at a small value for some time, and then slowly returned to its initial value, adiabatically converting the probe pulse between photon-like and plasmon-like character, with commensurate changes in the effective group velocity, and temporarily storing[246, 239, 240, 241, 253, 283, 250, 284, 242] the pulse information in the medium.

From the dressed polariton perspective, it is clear that the pulse travels with variable velocity but otherwise no change in the overall spatial shape or overall amplitude in the action density profile. It just slows down as it becomes more plasmonic. In terms of the bare action densities, the *relative* shape remains unchanged, while the overall amplitude increases or decreases in response to the pump strength. Bare quanta are converted adiabatically between photons and plasmons, with corresponding effective group velocity approximately given by the action-weighted average.

Finally, we verify that such adiabatic conversion can actually occur self-consistently within the limitations of our dynamical model. Firstly, we note that the frequency content of the probe pulse can be accommodated within the transparency window even as the latter will shrink in width as the pump is turned off. As we have seen, as $a_2(t)$ is slowly varied, the spatial bandwidth Δk_1 of the probe remains fixed, but the frequency bandwidth $\Delta\omega_1(t)$ of the probe varies, approximately in proportion to the group velocity $v_d(t)$, while the available polariton bandwidth (i.e., transparency window) $\Delta\omega_d(t)$ also varies, approximately in proportion to $|a_2(t)|$, so their ratio is given approximately by

$$\frac{\Delta\omega_1(t)}{\Delta\omega_d(t)} \approx \frac{\Delta\omega_1(t)}{\Delta\omega_d(t)} \left|\frac{a_2(t)}{a_2(t_b)}\right| \frac{1+\delta_d(t_b)^2}{1+\delta_d(t_b)^2\left|\frac{a_2(t)}{a_2(t_b)}\right|^2}, \tag{6.196}$$

where the factor $\delta_d(t_b) \approx \left|\frac{\Omega_{\rm L}(t_b)}{\Omega_{\rm G}(t_b)}\right| \sim O(1)$ or so for typical parameters, or at least $O(10^{-1}) \lesssim \delta_d(t_b) \lesssim O(10)$. For sufficiently intense pulses, i.e., asymptotically as $|a_2(t)| \to \infty$, the transparency window grows but the group velocity and probe duration asymptote to some finite values, so the ratio (6.196) goes to zero and the frequency content of the probe can easily remain within the transparency window. As the probe is turned off, i.e., in the limit $|a_2(t)| \to 0$, the transparency window shrinks to zero, but more slowly than the probe bandwidth, so again (6.196) goes to zero, and the bandwidth can be accommodated within the EIT pass-band. At an intermediate value of the pump strength, $|a(t)| = |a(t_b)|\,\delta(t_b)$, the ratio (6.196) reaches a finite maximum value of $\frac{1}{2}\left[\delta_d(t_b) + \delta_d(t_b)^{-1}\right]\frac{\Delta\omega_1(t)}{\Delta\omega_d(t)} \lesssim O(10)\frac{\Delta\omega_1(t)}{\Delta\omega_d(t)}$,

and as long as this value is less than unity the pulse can be accommodated within the transparency window for any value of the pump intensity.

Secondly, we confirm that the probe photon density can remain small compared to the probe photon density, even as the latter is adiabatically reduced. Under present assumptions, a little algebra reveals that the ratio of probe to pump photon action satisfies

$$\max_z \left|\frac{a_1(z,t)}{a_2(t)}\right|^2 = \left|\frac{a_2(t_b)}{a_2(t)}\right|^2 \frac{v_d(t)}{v_d(t_b)} \max_z \left|\frac{a_1(z,t_b)}{a_2(t_b)}\right|^2 = \frac{1+\delta_d(t_b)^2}{1+\delta_d(t_b)^2\left|\frac{a_2(t)}{a_2(t_b)}\right|^2} \max_z \left|\frac{a_1(z,t_b)}{a_2(t_b)}\right|^2, \quad (6.197)$$

which vanishes in the limit $|a_2(t)| \to \infty$, and remains bounded (within about an order-of-magnitude of its value at $t = t_b$) for all values of the pump intensity, and in particular as $|a_2(t)| \to 0$.

6.4 Collective Quantum Formalism for Atomic EIT

EIT in a cold magnetized plasma has essentially the same observable signatures of EIT in cold atomic vapors, only in a different frequency regime, but the fundamental question remains as to whether they truly represent different manifestations of the same underlying interference phenomenon.

In order to effect precise (if still qualitative) comparisons between the plasma and atomic versions of EIT, we now develop a model for EIT in atomic vapors which will mirror that for the plasma, involving collective-looking coupled Bosonic excitations. We start with the complete Hamiltonian for the electronic and center-of mass (COM) degrees-of-freedom for a collection of atoms and the EM field DOFs with which they interact, and then under various simplifying assumptions aim to coax this Hamiltonian through a series of approximations and unitary transformations into a form manifestly analogous to that derived above for the magnetized plasma.

6.4.1 Preliminaries

We suppose our quantization volume $\mathcal{V}$ is now filled with a vapor consisting of $N_{\rm A}$ identical neutral atoms, of uniform average density $n_{\rm A} = \frac{N_{\rm A}}{\mathcal{V}}$, each of rest mass $M_{\rm A}$ and containing $\mathcal{Z}_e$ bound electrons. In order that the transparency effect be observable, we demand that the vapor is optically thick with respect to the probe radiation, in a sense made more precise below.

We specify the center-of-mass (COM) position of the jth atom by $\boldsymbol{R}_j = \bar{X}_j\hat{\boldsymbol{x}}+\bar{Y}_j\hat{\boldsymbol{y}}+\bar{Z}_j\hat{\boldsymbol{z}}$, and the displacement of the sth electron in the atom relative to this COM position by $\boldsymbol{r}_{sj} = x_{sj}\hat{\boldsymbol{x}} + y_{sj}\hat{\boldsymbol{y}} + \xi_{sj}\hat{\boldsymbol{z}}$. These are deliberately written in a form evocative of analogous quantities for the plasma, but we stress that here the longitudinal COM position $\bar{Z}_j$ is an an operator representing an independent DOF of the atom, while in the plasma case $\bar{z}_j$ just represented the c-number expected equilibrium position of the unbound electron. To make the distinction clear, we will now use the upper case $\bar{Z}_j$ to denote the operator, and $\bar{z}_j$ the corresponding eigenvalue (or expectation value, depending on context).

Conjugate to the COM longitudinal position is the total longitudinal (non-relativistic) momentum $\bar{P}_j$ with eigenvalues $\bar{p}_j = \hbar k_j$, while the relative electron momenta $\boldsymbol{P}_{sj}$ are conjugate to the corresponding relative positions for $s = 1, \ldots, \mathcal{Z}_e$ and $j = 1, \ldots, N_{\mathrm{A}}$. Transverse components of the total momenta will not be needed in our mostly 1D treatment.

For later use, we define in the usual way generalized eigenkets associated with atomic COM degrees-of-freedom inside the imagined cavity, i.e., the (continuous) COM position $\bar{X}_j$ and (discrete) total momentum $\bar{P}_j$, that satisfy:

$$\left\langle z \,\middle|\, z' \right\rangle = \sum_n \delta(z - z' + nL); \tag{6.198a}$$

$$\bar{Z}_j \left|\bar{z}_j\right\rangle = \bar{z}_j \left|\bar{z}_j\right\rangle; \tag{6.198b}$$

$$\left\langle k \,\middle|\, k' \right\rangle = \delta_{kk'}; \tag{6.198c}$$

$$\bar{P}_j \left|k\right\rangle = \hbar k \left|k\right\rangle; \tag{6.198d}$$

$$\left|z\right\rangle = \tfrac{1}{\sqrt{L}} \sum_k e^{-ikz} \left|k\right\rangle; \tag{6.198e}$$

$$\left\langle z \,\middle|\, k \right\rangle = \tfrac{1}{\sqrt{L}} e^{+ikz}; \tag{6.198f}$$

$$\left|k\right\rangle = \tfrac{1}{\sqrt{L}} \int dz\, e^{+ikz} \left|z\right\rangle; \tag{6.198g}$$

$$e^{ik'\bar{Z}_j} \left|k\right\rangle = \left|k + k'\right\rangle. \tag{6.198h}$$

Note that the sign conventions for the exponents in the Fourier transforms between kets must be opposite those for the expansion coefficients (i.e., wave functions) or for the corresponding second-quantized annihilation operators.

Initially, the COM DOFs are assumed to be in thermal equilibrium at a temperature T_{A} satisfying $T_{\mathrm{A}} \ll M_{\mathrm{A}}c^2$, in units where Boltzmann's constant $k_{\mathrm{B}} = 1$. To account for the possibility of different relaxation time-scales during preparation, the initial internal state of the electronic DOFs for each atom is assumed to be characterized by a Boltzmann-Gibbs

distribution at a possibly different temperature $T_e \ll m_e c^2$. Further constraints on these temperatures will be imposed as we proceed.

6.4.2 Full Hamiltonian

The full Hamiltonian is decomposed into contributions for the EM field DOFs, the atomic DOFs, and the interactions between them:

$$\mathscr{H} = \mathscr{H}_{\rm EM} + \mathscr{H}_{\rm A} + \mathscr{H}_{\rm int}. \tag{6.199}$$

We describe and simplify each in turn.

Electromagnetic Hamiltonian

In the usual way, we define bare photon annihilation and creation operators as complex normal coordinates for the $1D$ Coulomb-gauge vector potential and associated free EM fields in the absence of the atomic vapor. The derivation of the bare EM field Hamiltonian is very similar to that described in the case of the unmagnetized plasma (only simpler, without the complications of the background dielectric) so need not be repeated. The result is of the familiar uncoupled harmonic-oscillator form:

$$\mathscr{H}_{\rm EM} = \sum_{k\,\mu} \hbar\omega_0(k)\, a^\dagger_{k\mu} a_{k\mu} \leftrightarrow \sum_{\mu} \int dk\, \hbar\omega_0(k)\, a_\mu(k)^\dagger a_\mu(k), \tag{6.200}$$

where here we take the operator $a^\dagger_{k\mu}$ to create a bare photon of momentum $\hbar k$, helicity (or spin) $\mathrm{sgn}[k]\mu$, and energy $\hbar\omega_0(k)$, only now $\omega_0(k) = c\,|k|$ is the vacuum EM dispersion relation. More generally, and in even closer analogy with the unmagnetized plasma system, we could accommodate the dispersive effects of some additional non-resonant buffer gas through a real background refractive index η_0, and instead set $\omega_0(k) = \frac{c}{\eta_0}\,|k|$, provided absorption can be neglected in the vicinity of the pump and probe frequencies. This buffer gas would play a role analogous to the *unmagnetized* plasma, modifying the dispersion relation that would be seen if the resonant effects were absent or were perfectly canceled by EIT. In particular, we will consider an EM prove with carrier wavenumber k_1, polarization μ_1, and bare carrier frequency $\omega_1 = \omega_0(k_1)$, and a downshifted probe with carrier wavenumber k_2, polarization μ_2, and bare carrier frequency $\omega_2 = \omega_0(k_2) < \omega_1$.

Intra-Atomic Hamiltonian

We neglect any inter-atomic forces, so the atomic portion of the Hamiltonian (i.e., what would be the energy in the absence of all external EM fields) is additive in the contributions from each atom:

$$\mathscr{H}_{\rm A} = \sum_j \mathcal{H}_{{\rm A}j}. \tag{6.201}$$

After expanding the relativistic energy in powers of the normalized particle (ion and electron) velocities, we retain only the largest intra-atomic terms, obtaining for the contribution associated with the jth atom:

$$\mathcal{H}_{{\rm A}j} = \mathcal{H}_{{\rm COM},j} + \mathcal{H}_{K\,j} + \mathcal{H}_{V\,j} + \mathcal{H}_{{\rm RM}j} + \mathcal{H}_{{\rm SR}j} + \mathcal{H}_{{\rm SO}j} + \mathcal{H}_{{\rm D}j} + \mathcal{H}_{{\rm HF}j} + \ldots \tag{6.202}$$

where, in order: $\mathcal{H}_{{\rm COM},j} = \frac{1}{2M_{\rm A}}\bar{P}_j^2$ represents the (non-relativistic) kinetic energy associated with the total (COM) longitudinal momentum $\bar{P}_j$; $\mathcal{H}_{K\,j} = \frac{1}{2m_e}\sum_s |\boldsymbol{P}_{js}|^2$ represent in the absence of external fields the (non-relativistic) residual kinetic energy of all $\mathcal{Z}_e$ electrons in the atom in the center-of-momentum frame, in terms of the relative canonical momenta $\boldsymbol{P}_{js}$; $\mathcal{H}_{V\,j} = V_j(\boldsymbol{r}_{js},\ldots,\boldsymbol{r}_{js},\ldots,\boldsymbol{r}_{j\mathcal{Z}_e})$ represents the internal potential energy, involving the Coulomb interactions between all charges in the atom (or in actual practice, the effective screened Coulomb interactions), written in terms of the relative electron coordinates $\boldsymbol{r}_{js}$ with respect to the COM position $\bar{\boldsymbol{R}}_j$; $\mathcal{H}_{{\rm RM}j} = M_{\rm N}c^2 + \mathcal{Z}_e m_e c^2 \approx M_{\rm A}c^2$ is just a constant representing the total rest mass energy for the nucleus and all electrons in the atom, and can be dropped; $\mathcal{H}_{{\rm SR}j} = -\frac{1}{8m_e^3c^2}\sum_s |\boldsymbol{P}_{js}|^4$ includes the leading-order corrections to the electronic kinetic energy due to special relativity; $\mathcal{H}_{{\rm SO}j}$ represents the usual spin-orbit corrections, coupling the spin and orbital components of the angular momentum, but whose explicit form will not be needed; $\mathcal{H}_{{\rm D}j}$ represents the Darwin correction that arises out the low-velocity expansion of the Dirac equation and which acts to smear out the Coulomb interactions over a Compton wavelength, its exact form also immaterial for our purposes; and finally $\mathcal{H}_{{\rm HF}j}$ includes hyperfine interactions between nuclear and electronic spins, whose exact form too is omitted. (Henceforth, the atom index j will sometimes be dropped when the distinction between the Hamiltonian for a single atom and for all atoms is clear from the context).

Because the particle rest masses satisfy $m_e \leq \mathcal{Z}_e m_e \ll M_{\rm N} \approx M_{\rm A}$, we ignore the distinctions between nuclear mass and atomic mass, and between the reduced mass and actual rest mass of each electron. The relative orderings of the various terms in the atomic Hamiltonian are roughly as follows: $\frac{\mathcal{H}_V}{\mathcal{H}_K} \sim O(1)$; $\frac{\mathcal{H}_{\rm SR}}{\mathcal{H}_K} \sim \frac{\mathcal{H}_{\rm SO}}{\mathcal{H}_K} \sim \frac{\mathcal{H}_{\rm D}}{\mathcal{H}_K} \sim O(\alpha_e^2)$; and $\frac{\mathcal{H}_{\rm HF}}{\mathcal{H}_K} \sim O\big(\frac{m_e}{M_{\rm A}}\alpha_e^2\big)$, where here $\alpha_e = \frac{e^2}{\hbar c} \approx \frac{1}{137}$ is the fine structure constant.

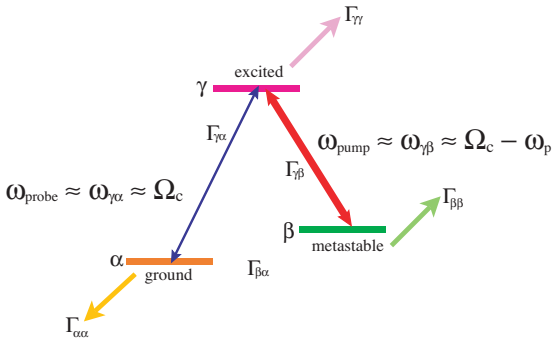

Figure 6.18. Schematic of the Λ-level manifold of the relevant internal electron states within each atom, along with the relevant couplings to the EM fields. Also shown are the decay and dephasing time-scales associated with various loss or decoherence mechanisms not included in our simple Hamiltonian model.

Now, we suppose that out of all of the complications of the electron dynamics relevant to our problem there emerges a Λ-level manifold of electronic energy eigenstates $|\alpha\rangle, |\beta\rangle, |\gamma\rangle$, with corresponding unperturbed energies $E_\alpha < E_\beta < E_\gamma$, respectively, and with Bohr frequencies close to resonance with the probe and pump fields: $\omega_{\gamma\alpha} \equiv \frac{(E_\gamma - E_\alpha)}{\hbar} \sim \omega_1$ and $\omega_{\gamma\beta} \equiv \frac{(E_\gamma - E_\beta)}{\hbar} \sim \omega_2$. By assumption, the symmetry of the states is such that selection rules allow the relative ground state $|\alpha\rangle$ and the excited state $|\gamma\rangle$ to be connected by electric dipole transitions compatible with the polarization of the probe wave, and also allow dipole transitions between the metastable state $|\beta\rangle$ and $|\gamma\rangle$ induced by the pump wave, but forbid direct transitions between $|\alpha\rangle$ and $|\beta\rangle$, at least at the single-photon level.

We essentially ignore all other electronic states, supposing them sufficiently uncoupled from any single-photon or multi-photon resonances with the pump and/or probe waves, and supposing the rates for any spontaneous decay into or out of the Λ-level manifold from or to other subspaces are slow compared to all relevant time-scales. The relevant level diagram is shown schematically in Fig. 6.18, along with the couplings to the EM pump and probe fields. Also shown are characteristic rates for decay and decoherence, which parameterize various physical loss or dephasing processes neglected in our reduced Hamiltonian model, such as spontaneous emission into extraneous states just mentioned, or collisional or other inter-atomic effects or uncontrolled environmental interactions. The rates $\Gamma_{\alpha\alpha}$, $\Gamma_{\beta\beta}$, and $\Gamma_{\gamma\gamma}$ reflect decay rates for the atomic level populations, or diagonal elements, in the reduced

atomic density-matrix, while $\Gamma_{\beta\alpha}$, $\Gamma_{\gamma\alpha}$ and $\Gamma_{\gamma\beta}$ reflect rates of decoherence of the relative phases reflected in the corresponding off-diagonal elements of a density matrix.

To develop a reduced but fully Hamiltonian model, here we will ignore these hopefully slow rates, or rather consider them only in passing when assessing the limitations of the present theory. In practice, for slow dephasing or decay, these effects can be incorporated phenomenologically into the dynamics at the end of the calculation by "analytic continuation" of various frequencies or coupling constants. In principle, any of these rates might be dependent on the atomic position or momentum or various external factors.

With the atoms assumed to be confined within the finite quantization volume, any energy eigenstate for the jth atom can then be specified by a discrete non-relativistic total momentum $\hbar q$ and the internal electronic state $\upsilon \in \{\alpha, \beta, \gamma\}$:

$$\left\langle \upsilon; q \left| \mathcal{H}_{\mathrm{A}j} \right| \upsilon'; q' \right\rangle = \delta_{\upsilon\upsilon'}\delta_{qq'} \left[E_\upsilon + \tfrac{\hbar^2 q^2}{2M_{\mathrm{A}}} \right], \tag{6.203}$$

where, to avoid notational clutter, explicit indexing of the eigenkets or eigenvalues by j is suppressed when no confusion should arise as a result.

We have already assumed $T_{\mathrm{A}} \ll M_{\mathrm{A}}c^2$, ensuring that initial atom momenta are predominately non-relativistic. We further suppose that additional energy scales satisfy

$$T_e \ll \min\left[\hbar\omega_2, \hbar(\omega_1 - \omega_2)\right] < \hbar\omega_1 \ll m_e c^2 \ll M_{\mathrm{A}}c^2, \tag{6.204}$$

implying that the atoms are sufficiently cold internally so that initially only the ground state is appreciably populated. This also implies that the recoil suffered by an atom when it absorbs or emits a probe or pump photon is small, in the sense that $\hbar\omega_2 < \hbar\omega_1 \ll M_{\mathrm{A}}c^2$ and $\hbar k_2 \leq \hbar k_1 \ll M_{\mathrm{A}}c$, so the kinetic energy and linear momentum of each atom will remain non-relativistic when exposed to the pump and/or probe.

We can also safely assume assume $T_{\mathrm{A}} \ll \min\left[\hbar\omega_2, \hbar(\omega_1 - \omega_2)\right] < \hbar\omega_1$, so any collision were it to occur would not excite internal states (although such collisions could still de-phase the metastable excited states, so are ignored more or less out of mathematical necessity here). It would be extremely useful if we could additionally assume the more stringent condition $\frac{T_{\mathrm{A}}}{\hbar\omega_2} \ll \frac{\hbar\omega_2}{M_{\mathrm{A}}c^2}$ or equivalently $\sqrt{M_{\mathrm{A}}T_{\mathrm{A}}} \ll \hbar k_2$, implying that the initial thermal momentum spread is small compared to the recoil momentum due to photon emission or absorption, because then we could treat the atoms as essentially condensed in their interaction with the fields. But for typical atomic parameters and optical frequency lasers, this would require ultra-cold atoms ($T_{\mathrm{A}} \sim O(\mu\mathrm{K})$) in a state-of-the-art trap, whereas EIT

is regularly observed in warmer gas cells. Therefore we regrettably conclude that we *cannot* assume the thermal momentum spread of the atoms is ultra-small in this sense.

Even with $1D$ geometry assumed for the fields, it may seem a little far-fetched to also restrict the motion of the COM of each atom to the longitudinal direction only. We have supposed the atoms are cold, but more importantly, regardless of their magnitudes, the transverse momenta (and positions) of the atoms are actually irrelevant if the fields are really one-dimensional to a good approximation and if collisions may be neglected – for then the atom experiences the same field independent of its transverse position, sees no doppler shift as a consequence of any transverse velocity, and recoils only in the longitudinal direction whenever it emits or absorbs a probe or pump photon. Transverse motion of each atoms is not constrained, just ignored as irrelevant for our purposes. Of course, transverse components of the momenta of the valence electrons have been included, to allow coupling to the laser fields.

Field-Atom Interactions

As in the plasma case, interaction terms arise from expanding all particle kinetic energies in the presence of a transverse vector potential $\boldsymbol{A}(z,t)$ representing the effects of the optical fields, leading to cross terms between the canonical momenta and the vector potential that are responsible for the usual multipole couplings (starting with the electric dipole transitions), and also to higher-order terms that are quadratic in the vector potential:

$$\mathscr{H}_{\rm int} = \sum_j \mathcal{H}_{mpj} + \mathcal{H}_{A^2j}, \tag{6.205}$$

where

$$\mathcal{H}_{mpj} = \tfrac{e}{2m_ec} \sum_s \left[\boldsymbol{P}_{js}\cdot\boldsymbol{A}(\bar{Z}_j+\xi_{sj},t) + \boldsymbol{A}(\bar{Z}_j+\xi_{sj},t)\cdot\boldsymbol{P}_{js} \right], \tag{6.206}$$

and

$$\mathcal{H}_{A^2j} = \tfrac{e}{2m_ec^2} \sum_s \|\boldsymbol{A}(\bar{Z}_j+\xi_{sj},t)\|^2, \tag{6.207}$$

We will retain the lowest-order (electric dipole) contributions from $\mathcal{H}_{mpj}$ as well as the lowest-order contributions from $\mathcal{H}_{A^2j}$. In atomic physics the latter are usually small and thrown away, but we keep them here for later comparison. Roughly speaking, we can estimate

$$\frac{|\mathcal{H}_{A^2j}|}{|\mathcal{H}_{mpj}|} \sim \tfrac{|a_{\rm pump}|}{2\alpha_e\sigma_{\rm A}\bar{Z}_e}, \tag{6.208}$$

where $\sigma_{\rm A} \sim O(1)$ is an effective screening term. Now, lasers based on chirped pulse amplification (CPA) can readily operate at optical or near optical frequencies with intensities

corresponding to $|a_{\text{pump}}| > 1$, but they would tend to overpower any signature of EIT, inducing tunneling ionization, harmonic generation, or other multi-photon processes. So we will assume a more typical value of $|a_{\text{pump}}| \lesssim O(10^{-1})$, such that for $\mathcal{Z}_e \sim O(20)$ to $O(100)$ we find $\frac{|\mathcal{H}_{A^2 j}|}{|\mathcal{H}_{mp\,j}|} \lesssim (10^{-1})$. In any case, while the terms quadratic in the laser vector potential may be small, we retain their lowest-order contributions for the purposes of comparison with the plasma case.

Dipole Interactions

To simplify what will be seen to be the dipole coupling terms, first notice that by using the atomic Hamiltonian (6.202), we find that for any (possibly complex) constant unit vector $\hat{\boldsymbol{\epsilon}}$,

$$\tfrac{1}{i\hbar}\left[\hat{\boldsymbol{\epsilon}}\cdot\boldsymbol{r}_{js}, \mathcal{H}_{\text{A}j}\right] = \tfrac{1}{m_e}\left(\hat{\boldsymbol{\epsilon}}\cdot\boldsymbol{P}_{js}\right)\left\{1 + O\left(\alpha_e^2\right) + O\left(\alpha_e^2 \tfrac{m_e}{M_{\text{A}}}\right) + \ldots,\right\}. \tag{6.209}$$

We can safely keep only the first term, and obtain the approximation

$$\begin{aligned}\sum_s \left\langle \upsilon \left| \hat{\boldsymbol{\epsilon}}\cdot\boldsymbol{P}_{js} \right| \upsilon' \right\rangle &\approx \tfrac{m_e}{i\hbar} \sum_s \left\langle \upsilon \left| \left[\hat{\boldsymbol{\epsilon}}\cdot\boldsymbol{r}_{js}, \mathcal{H}_{\text{A}j}\right] \right| \upsilon' \right\rangle \\ &= \tfrac{m_e}{i\hbar q_e}\left(E_{\upsilon'} - E_\upsilon\right) \sum_s \left\langle \upsilon \left| q_e \hat{\boldsymbol{\epsilon}}\cdot\boldsymbol{r}_{js} \right| \upsilon' \right\rangle = \tfrac{im_e}{e}\omega_{\upsilon'\upsilon}\hat{\boldsymbol{\epsilon}}\cdot\boldsymbol{\mu}_{\upsilon\upsilon'},\end{aligned} \tag{6.210}$$

in which $\omega_{\upsilon'\upsilon} \equiv \frac{E_{\upsilon'} - E_\upsilon}{\hbar}$ are the generalized Bohr frequencies (regardless of whether the corresponding transition is allowed), and

$$\boldsymbol{\mu}_{\upsilon\upsilon'} = \left\langle \upsilon \left| \sum_s q_e \boldsymbol{r}_{js} \right| \upsilon' \right\rangle \tag{6.211}$$

are the matrix elements for the total electron dipole moment operator of a given atom (i.e., summed over all electrons), with respect to the COM position (which is essentially the same as the position of the nuclear charge).

By considerations of parity, we deduce that $\boldsymbol{\mu}_{\alpha\alpha'} = \boldsymbol{\mu}_{\beta\beta'} = \boldsymbol{\mu}_{\gamma\gamma'} = 0$ always, while we *assume* that parity or other selection rules ensure that $\boldsymbol{\mu}_{\alpha\beta} = \boldsymbol{\mu}^*_{\beta\alpha} = 0$. The only non-vanishing dipole matrix elements are then $\boldsymbol{\mu}_{\alpha\gamma} = \boldsymbol{\mu}^*_{\gamma\alpha} \neq \mathbf{0}$ and $\boldsymbol{\mu}_{\beta\gamma} = \boldsymbol{\mu}^*_{\gamma\beta} \neq \mathbf{0}$.

Next, recall that, in the Coulomb gauge, $\boldsymbol{P}_{js}\cdot\boldsymbol{A}(\bar{Z}_j + \xi_{sj}, t) = \boldsymbol{A}(\bar{Z}_j + \xi_{sj}, t)\cdot\boldsymbol{P}_{js}$. Still assuming the atoms and fields are confined to the quantization volume, after substitution of the modal decomposition for the $1D$ vector potential we find

$$\left\langle \upsilon; q \left| \mathcal{H}_{dp\,j} \right| \upsilon'; q' \right\rangle = \tfrac{1}{2}\tfrac{e}{m_e}\tfrac{\sqrt{2\pi}}{\sqrt{\mathcal{V}}} \sum_{k\mu} \tfrac{\sqrt{\hbar}}{\sqrt{\omega_0(k)}} a_{k\mu} \left\langle q \left| e^{ik\bar{Z}_j} \right| q' \right\rangle \sum_s \left\langle \upsilon \left| e^{ik\xi_{sj}} \hat{\boldsymbol{\epsilon}}_\mu \cdot \boldsymbol{P}_{js} \right| \upsilon' \right\rangle + h.c. \tag{6.212}$$

Now for sufficiently low-$\mathcal{Z}_e$ atoms, which we shall assume, and for any wavenumber $k \sim \frac{\omega_1}{c}$ or $k \sim \frac{\omega_2}{c}$ comparable to resonant electronic transitions in such atoms, we will have $|k\xi_{js}| \lesssim kR_{\rm A} \sim \alpha\mathcal{Z}_e \ll 1$; in other words, the radiation wavelength (typically at optical or radio frequencies) greatly exceeds the size $R_{\rm A}$ of an atom. We can therefore set $e^{ik\xi_{sj}} \approx 1$ in the matrix elements, which is in fact the well-known electric dipole approximation. Using this and our previous results, the dipole terms simplify to:

$$\left\langle \upsilon; q \left| \mathcal{H}_{dp\,j} \right| \upsilon'; q' \right\rangle \approx \tfrac{1}{2}\hbar\omega_{\upsilon'\upsilon}\tfrac{\sqrt{2\pi}}{\sqrt{\mathcal{V}}}\sum_{k\mu}\tfrac{1}{\sqrt{\hbar\omega_0(k)}}\delta_{q,q'+k}\left[ia_{k\mu}\hat{\boldsymbol{\epsilon}}_\mu\cdot\boldsymbol{\mu}_{\upsilon\upsilon'} + h.c.\right] \tag{6.213}$$

We see that these terms represent emission or absorption of photons by an atom, with concomitant change in the internal energy of the bound electrons and recoil of the COM.

Two-Photon Terms

For atomic systems subject to radiation fields of moderate intensity at optical frequencies, it is generally the case that $\frac{\mathcal{H}_{A^2}}{\mathcal{H}_{dp}} \sim \alpha\mathcal{Z}_e \ll 1$ meaning these second-order terms are small, roughly of the same order as terms we have already neglected.

Nevertheless, in order to better illuminate the analogies between the atomic and plasma systems, we will retain leading-order contributions from these terms. After a bit of algebraic manipulation we obtain

$$\begin{aligned}\left\langle \upsilon; q \left| \mathcal{H}_{A^2\,j} \right| \upsilon'; q' \right\rangle \approx\ & \tfrac{e^2}{2m_e}\tfrac{2\pi}{\mathcal{V}}\sum_{k\mu}\sum_{k'\mu'}\tfrac{\hbar\hat{\boldsymbol{\epsilon}}_\mu\hat{\boldsymbol{\epsilon}}_{\mu'}}{\sqrt{\omega_0(k)\omega_0(k')}}a_{k\mu}a_{k'\mu'}\delta_{q,q'+k'+k}\left\langle \upsilon\left| e^{i(k'+k)\xi_{js}}\right|\upsilon'\right\rangle + h.c. \\ & + \tfrac{e^2}{2m_e}\tfrac{2\pi}{\mathcal{V}}\sum_{k\mu}\sum_{k'\mu'}\tfrac{\hbar\hat{\boldsymbol{\epsilon}}^*_\mu\hat{\boldsymbol{\epsilon}}_{\mu'}}{\sqrt{\omega_0(k)\omega_0(k')}}a^\dagger_{k\mu}a_{k'\mu'}\delta_{q,q'+k'-k}\left\langle \upsilon\left| e^{i(k'-k)\xi_{js}}\right|\upsilon'\right\rangle + h.c.\end{aligned} \tag{6.214}$$

Expanding $e^{i(k'\pm k)\xi_{sj}} \approx 1 + (k'\pm k)\xi_{sj} - \frac{1}{2}(k'\pm k)^2\xi_{js}^2 + \cdots$, the electronic matrix elements become

$$\left\langle \upsilon\left| e^{i(k'\pm k)\xi_{js}}\right|\upsilon'\right\rangle \approx \delta_{\upsilon\upsilon'} + i(k'\pm k)\left\langle \upsilon\left|\xi_{js}\right|\upsilon'\right\rangle - \tfrac{1}{2}(k'\pm k)^2\left\langle \upsilon\left|\xi_{js}^2\right|\upsilon'\right\rangle + \ldots . \tag{6.215}$$

From our previous definitions, $-e\left\langle \upsilon\left|\xi_{js}\right|\upsilon'\right\rangle = \hat{\boldsymbol{z}}\cdot\boldsymbol{\mu}_{\upsilon\upsilon'}$. By neglecting transitions from or to electronic states outside the Λ-level manifold, we implicitly assume $|\alpha\rangle\langle\alpha| + |\beta\rangle\langle\beta| +$

$|\gamma\rangle\langle\gamma| = I$ when acting on the electronic DOFs, and can deduce

$$e^2 \langle \alpha | \xi_{js}^2 | \alpha \rangle = |\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\alpha\gamma}|^2 ; \tag{6.216a}$$

$$e^2 \langle \beta | \xi_{js}^2 | \beta \rangle = |\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\beta\gamma}|^2 ; \tag{6.216b}$$

$$e^2 \langle \gamma | \xi_{js}^2 | \gamma \rangle = |\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\alpha\gamma}|^2 + |\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\beta\gamma}|^2 ; \tag{6.216c}$$

$$e^2 \langle \alpha | \xi_{js}^2 | \beta \rangle = (\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\alpha\gamma})(\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\gamma\beta}); \tag{6.216d}$$

$$e^2 \langle \alpha | \xi_{js}^2 | \gamma \rangle = 0; \tag{6.216e}$$

$$e^2 \langle \beta | \xi_{js}^2 | \gamma \rangle = 0. \tag{6.216f}$$

Next we invoke the RWA and drop the anti-resonant terms which do not conserve total photon number, to obtain

$$\langle \alpha; q | \mathcal{H}_{A^2} | \alpha; q' \rangle = \tfrac{e^2}{2m_e} \tfrac{2\pi}{\mathcal{V}} \sum_{k\mu} \sum_{k'\mu'} \tfrac{\hbar \delta_{\mu\mu'} \delta_{q,q'+k'-k}}{\sqrt{\omega_0(k)\omega_0(k')}} a_{k\mu}^\dagger a_{k'\mu'} \Big[\mathcal{Z}_e - \tfrac{(k-k')^2}{2e^2} |\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\alpha\gamma}|^2 \Big] + h.c.; \tag{6.217a}$$

$$\langle \beta; q | \mathcal{H}_{A^2} | \beta; q' \rangle = \tfrac{e^2}{2m_e} \tfrac{2\pi}{\mathcal{V}} \sum_{k\mu} \sum_{k'\mu'} \tfrac{\hbar \delta_{\mu\mu'} \delta_{q,q'+k'-k}}{\sqrt{\omega_0(k)\omega_0(k')}} a_{k\mu}^\dagger a_{k'\mu'} \Big[\mathcal{Z}_e - \tfrac{(k-k')^2}{2e^2} |\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\beta\gamma}|^2 \Big] + h.c.; \tag{6.217b}$$

$$\langle \gamma; q | \mathcal{H}_{A^2} | \gamma; q' \rangle = \tfrac{e^2}{2m_e} \tfrac{2\pi}{\mathcal{V}} \sum_{k\mu} \sum_{k'\mu'} \tfrac{\hbar \delta_{\mu\mu'} \delta_{q,q'+k'-k}}{\sqrt{\omega_0(k)\omega_0(k')}} a_{k\mu}^\dagger a_{k'\mu'} \left[\mathcal{Z}_e - \tfrac{(k-k')^2}{2e^2} \begin{pmatrix} |\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\alpha\gamma}|^2 \\ + |\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\beta\gamma}|^2 \end{pmatrix} \right] + h.c.; \tag{6.217c}$$

$$\langle \alpha; q | \mathcal{H}_{A^2} | \gamma; q' \rangle = \tfrac{e}{2m_e} \tfrac{2\pi}{\mathcal{V}} \sum_{k\mu} \sum_{k'\mu'} \tfrac{\hbar \delta_{\mu\mu'} \delta_{q,q'+k'-k}}{\sqrt{\omega_0(k)\omega_0(k')}} a_{k\mu}^\dagger a_{k'\mu'} \big[i(k-k') \hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\alpha\gamma} \big] + h.c.; \tag{6.217d}$$

$$\langle \beta; q | \mathcal{H}_{A^2} | \gamma; q' \rangle = \tfrac{e}{2m_e} \tfrac{2\pi}{\mathcal{V}} \sum_{k\mu} \sum_{k'\mu'} \tfrac{\hbar \delta_{\mu\mu'} \delta_{q,q'+k'-k}}{\sqrt{\omega_0(k)\omega_0(k')}} a_{k\mu}^\dagger a_{k'\mu'} \big[i(k-k') \hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\beta\gamma} \big] + h.c.; \tag{6.217e}$$

$$\langle \alpha; q | \mathcal{H}_{A^2} | \beta; q' \rangle = \tfrac{1}{2m_e} \tfrac{2\pi}{\mathcal{V}} \sum_{k\mu} \sum_{k'\mu'} \tfrac{\hbar \delta_{\mu\mu'} \delta_{q,q'+k'-k}}{\sqrt{\omega_0(k)\omega_0(k')}} a_{k\mu}^\dagger a_{k'\mu'} \Big[-\tfrac{(k-k')^2}{2} (\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\alpha\gamma})(\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\gamma\beta}) \Big] + h.c.; \tag{6.217f}$$

These terms all reflect direct two-photon terms in the Hamiltonian, in contrast to two-photon terms that arise at second order in a Dyson expansion for the propagator.

In the absence of the pump, we assume that the atomic density and the oscillator strength for the ground-excited state transition are sufficiently high so that the medium is optically dense to the probe radiation, i.e., the characteristic absorption length is much shorter than the interaction length in the plasma

$$L_{\text{abs}} \sim \frac{\mathcal{V}}{N_{\text{A}} \sigma_{\text{abs}}} \ll L \tag{6.218}$$

where $\sigma_{\text{abs}} \sim 2\,\frac{\omega_{\gamma\alpha}}{\Gamma_{\gamma\gamma}}\frac{\left|\hat{\boldsymbol{\epsilon}}_{\mu_1}\cdot\boldsymbol{\mu}_{\gamma\alpha}\right|^2}{\hbar c}$ is the total absorption cross section associated with this resonance.

Discrete Mode Hamiltonian in Second-Quantized Form

Following essentially the same strategy adopted in the plasma case, we simplify matters by assuming a discrete probe mode $a_1 \equiv a_{\mu_1}(k_1)$ of polarization μ_1 (assumed to satisfy relevant selection rules for the $|\alpha\rangle \to |\gamma\rangle$ transition), at wavenumber k_1, and corresponding bare frequency $\omega_0(k_1) \approx \omega_{\gamma\alpha}$, and by also assuming a downshifted pump mode $a_2 \equiv a_{\mu_2}(k_2)$ of polarization μ_2 (appropriate to the $|\beta\rangle \to |\gamma\rangle$ transition), at wavenumber k_2, and bare frequency $\omega_0(k_2) \approx \omega_{\gamma\beta} = \omega_{\gamma\alpha} - \omega_{\beta\alpha}$. Assuming lasers of sufficiently narrow bandwidth and of sufficiently moderate intensity, these frequencies will be well separated from each other, but much larger than the Rabi frequencies associated with the couplings, so we retain only resonant and RWA terms, with one exception: we also keep the lowest-order non-resonant (but not anti-resonant) effects of the pump, in particular the dipole-coupling between the pump and the $|\alpha\rangle \to |\gamma\rangle$ state transition, in part because the pump may be much more intense than the probe, but mostly just to better elucidate analogies with the plasma case. However, keep in mind that this off-resonant coupling may vanish exactly due to selection rules for some common Λ-level systems.

As before, for greater compactness of notation we now convey wavenumber dependence in parenthetical rather than indexed form, even though the modes are here still considered discrete.

For simplicity, we will assume the atoms are Bosonic, but in the end similar results can be obtained for Fermionic atoms in suitable temperature and density regimes, if the inter-atomic spacing is much smaller than the characteristic length-scale for slow modulations in the field envelopes. To describe the system consisting of the EM fields and all atoms, it is convenient to now adopt a fully second-quantized representation, and express the Hamiltonian in terms of the EM mode annihilation and creation operators as before as well as annihilation operators $\Psi_\upsilon(k)$ and their adjoint creation operators $\Psi_\upsilon(k)^\dagger$ for atoms of total momentum $\hbar k$ and internal quantum state $\upsilon = \alpha,\ \beta,$ or γ. For bosonic atoms, these operators satisfy CCRs:

$$\left[\Psi_\upsilon(k), \Psi_{\upsilon'}(k')^\dagger\right] = \delta_{\upsilon\upsilon'}\delta_{kk'}; \tag{6.219a}$$

$$\left[\Psi_\upsilon(k), \Psi_{\upsilon'}(k')\ \right] = 0. \tag{6.219b}$$

Their unitary Fourier transforms,

$$\Psi_\upsilon(z) \;= \tfrac{1}{\sqrt{L}} \sum_k e^{ikz} \Psi_\upsilon(k); \tag{6.220a}$$

$$\Psi_\upsilon(z)^\dagger = \tfrac{1}{\sqrt{L}} \sum_k e^{-ikz} \Psi_\upsilon(k)^\dagger; \tag{6.220b}$$

satisfy (continuum) CCRs, and annihilate or create, respectively, atoms in the electronic state υ at position z, assuming periodic boundary conditions.

In terms of these operators, the relevant part of the Hamiltonian can be written as

$$\mathscr{H} = \mathcal{H}_{\rm EM} + \mathcal{H}_{\rm A} + \mathscr{H}_{dp} + \mathscr{H}_{A^2}, \tag{6.221}$$

where

$$\mathscr{H}_{\rm EM} = \hbar\omega_0(k_1) a_1^\dagger a_1 + \hbar\omega_0(k_2) a_2^\dagger a_2 \tag{6.222}$$

describes the free electromagnetic fields;

$$\begin{aligned} \mathscr{H}_{\rm A} = {} & E_\alpha \sum_q \left[N_\alpha(q) + N_\beta(q) + N_\gamma(q) \right] \\ & + \sum_q \left[\tfrac{\hbar^2 q^2}{2M_{\rm A}} \right] N_\alpha(q) \\ & + \sum_q \left[\hbar\omega_{\beta\alpha} + \tfrac{\hbar^2 (q+k_1-k_2)^2}{2M_{\rm A}} \right] N_\beta(q+k_1-k_2) \\ & + \sum_q \left[\hbar\omega_{\gamma\alpha} + \tfrac{\hbar^2 (q+k_1)^2}{2M_{\rm A}} \right] N_\gamma(q+k_1) \end{aligned} \tag{6.223}$$

describes both the COM and internal DOFs of the unperturbed atoms;

$$\begin{aligned} \mathcal{H}_{dp} = {} & -\tfrac{1}{2}\hbar\omega_{\gamma\alpha} \tfrac{\sqrt{2\pi}}{\sqrt{\mathcal{V}}} \tfrac{\hat{\boldsymbol{\epsilon}}_{\mu_1} \cdot \boldsymbol{\mu}_{\gamma\alpha}}{\sqrt{\hbar\omega_0(k_1)}} i a_1 \sum_q \Psi_\gamma(q+k_1)^\dagger \Psi_\alpha(q) + h.c. \\ & -\tfrac{1}{2}\hbar\omega_{\gamma\beta} \tfrac{\sqrt{2\pi}}{\sqrt{\mathcal{V}}} \tfrac{\hat{\boldsymbol{\epsilon}}_{\mu_2} \cdot \boldsymbol{\mu}_{\gamma\beta}}{\sqrt{\hbar\omega_0(k_2)}} i a_2 \sum_q \Psi_\gamma(q+k_2)^\dagger \Psi_\beta(q+k_1-k_2) + h.c., \\ & -\tfrac{1}{2}\hbar\omega_{\gamma\alpha} \tfrac{\sqrt{2\pi}}{\sqrt{\mathcal{V}}} \tfrac{\hat{\boldsymbol{\epsilon}}_{\mu_2} \cdot \boldsymbol{\mu}_{\gamma\alpha}}{\sqrt{\hbar\omega_0(k_2)}} i a_2 \sum_q \Psi_\gamma(q+k_2)^\dagger \Psi_\alpha(q) + h.c., \end{aligned} \tag{6.224}$$

describes the resonant dipole couplings, and a possible off-resonant coupling involving the pump; and finally

$$\begin{aligned} \mathscr{H}_{A^2} = {} & \tfrac{\mathcal{Z}_e e^2}{m_e} \tfrac{2\pi}{\mathcal{V}} \tfrac{\hbar}{\omega_0(k_1)} a_1^\dagger a_1 \sum_q \left[N_\alpha(q) + N_\beta(q) + N_\gamma(q) \right] \\ & + \tfrac{\mathcal{Z}_e e^2}{m_e} \tfrac{2\pi}{\mathcal{V}} \tfrac{\hbar}{\omega_0(k_2)} a_2^\dagger a_2 \sum_q \left[N_\alpha(q) + N_\beta(q) + N_\gamma(q) \right] \\ & - \tfrac{1}{4m_e} \tfrac{2\pi}{\mathcal{V}} \tfrac{\hbar(k_1-k_2)^2}{\sqrt{\omega_0(k_1)\omega_0(k_2)}} (\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\gamma\alpha})^* (\hat{\boldsymbol{z}} \cdot \boldsymbol{\mu}_{\gamma\beta}) \, a_2^\dagger a_1 \sum_q \Psi_\beta(q+k_1-k_2)^\dagger \Psi_\alpha(q) + h.c., \end{aligned} \tag{6.225}$$

where

$$N_\alpha(q) = \Psi_\alpha(q)^\dagger \Psi_\alpha(q), \tag{6.226a}$$

$$N_\beta(q) = \Psi_\beta(q)^\dagger \Psi_\beta(q), \qquad \text{and} \tag{6.226b}$$

$$N_\gamma(q) = \Psi_\gamma(q)^\dagger \Psi_\gamma(q) \tag{6.226c}$$

are the usual number operators for atoms of a given momentum and internal state. In the expressions involving a sum over all possible atomic momentum, we used the resulting freedom to shift the dummy argument q in order to simplify our subsequent manipulations.

Bosonification of the Elementary Atomic Excitations and Other Approximations

Already the atomic EIT Hamiltonian is looking somewhat reminiscent of the Hamiltonian for magnetized plasma, but under further (hopefully reasonable) assumptions, some additional approximations and transformations can nudge them closer still.

First, let us re-write the unperturbed atomic Hamiltonian as

$$\begin{aligned}
\mathscr{H}_{\rm A} = {} & E_\alpha \sum_q \left[N_\alpha(q) + N_\beta(q) + N_\gamma(q) \right] \\
& + \sum_q \tfrac{\hbar^2 q^2}{2M_{\rm A}} \left[N_\alpha(q) + N_\beta(q+k_1-k_2) + N_\gamma(q+k_1) \right] \\
& + \sum_q \left[\hbar\omega_{\beta\alpha} + \tfrac{\hbar^2 (k_1-k_2)^2}{2M_{\rm A}} + \tfrac{\hbar^2 q (k_1-k_2)}{M_{\rm A}} \right] N_\beta(q+k_1-k_2) \\
& + \sum_q \left[\hbar\omega_{\gamma\alpha} + \tfrac{\hbar^2 k_1^2}{2M_{\rm A}} + \tfrac{\hbar^2 q k_1}{M_{\rm A}} \right] N_\gamma(q+k_1)
\end{aligned} \tag{6.227}$$

First, notice that the Hamiltonian obviously conserves the total atom number $\sum_q \left[N_\alpha(q) + N_\beta(q) + N_\gamma(q) \right]$, which we can therefore replace with the c-number $N_{\rm A}$ everywhere that operator appears. The first term in $\mathscr{H}_{\rm A}$ reduces to a constant term, $N_{\rm A} E_\alpha$, and can be dropped, while the combination of $\mathscr{H}_{\rm EM}$ and the diagonal terms from $\mathscr{H}_{A^2}$ just leads to renormalized EM frequencies for the bare photon modes:

$$\mathscr{H}_{\rm EM'} = \hbar\omega_0(k_1) \left[1 + \tfrac{1}{2} \tfrac{\omega_{p{\rm A}}^2}{\omega_0(k_1)^2} \right] a_1^\dagger a_1 + \hbar\omega_0(k_2) \left[1 + \tfrac{1}{2} \tfrac{\omega_{p{\rm A}}^2}{\omega_0(k_1)^2} \right] a_2^\dagger a_2, \tag{6.228}$$

where

$$\omega_{p{\rm A}} = \left[\tfrac{4\pi N_{\rm A} \mathcal{Z}_e e^2}{\mathcal{V} m_e} \right]^{1/2} \tag{6.229}$$

is the plasma frequency associated with all $N_e = N_{\rm A} \mathcal{Z}_e$ electrons if treated as unbound and uniformly distributed in the quantization volume $\mathcal{V}$. For high EM frequencies, i.e.,

$\omega_{p\mathrm{A}} \ll \omega_0(k_2) < \omega_0(k_1)$, we see that the RWA expression (6.228) is just the leading-order Taylor expansion of

$$\mathscr{H}_{\mathrm{EM}'} \approx \hbar\omega_0(k_1)\sqrt{1+\tfrac{\omega_{p\mathrm{A}}^2}{\omega_0(k_1)^2}}\, a_1^\dagger a_1 + \hbar\omega_0(k_2)\sqrt{1+\tfrac{\omega_{p\mathrm{A}}^2}{\omega_0(k_1)^2}}\, a_2^\dagger a_2, \tag{6.230}$$

which yields the classical asymptotic (plasma-like) results for EM dispersion relations to be expected in atomic vapors at frequencies far above any internal bound-state resonances. We could just incorporate these shifts into a modified definition for the bare frequency $\bar{\omega}_0(k)$, but at typical vapor densities the effective plasma frequency is negligibly small compared to optical or near-optical laser frequencies, and the bare dispersion relations are well approximated by the linear vacuum forms. For simplicity, we will make this assumption here, but as a matter of principle we do stress that, just like the plasma case, atomic EIT can at most cancel resonant effects of the medium, and will not, strictly speaking, cancel, erase, or otherwise suppress the background non-resonant effects of the medium. This was readily apparent in the magnetized plasma case and had been suggested as an important distinction between the two cases, but is actually also true in principle in the atomic case, but overlooked in practice because the non-resonant effects of the medium are typically quite small.

Next, we invoke what is really one of the central assumptions of our development – namely, that the atomic EIT dynamics remain very far from saturation, in the sense that the density of probe photons is assumed to remain much less than the density of atoms. Together with our previous assumption of a large number of initially cold atoms in the quantization volume, this implies that virtually all atoms initially occupy the internal ground state, and most atoms remain in the ground state throughout the interaction. It is this lack of saturation which will allow us to identify certain excitations of the atomic vapor as the analogs of the collective linear gyron and plasmon waves in the plasma system. From a Schrödinger state picture, recall that each plasma electron has available (in the cold non-relativistic limit) a very large number of evenly-spaced Landau levels, while each atom has available only a single excited state.[10] Likewise, each plasma electron can participate in longitudinal oscillations taking on amplitudes corresponding to long ladder of evenly-spaced Langmuir levels,[11] while each atom has only a single metastable state. To establish

[10]The gyron number operator $g_k^\dagger g_k$ also formally admits an infinite manifold of SHO-like states, and there is no analog of wave-breaking seen in longitudinal waves that limits the gyro-radius or velocity of transverse gyrations, but the non-relativistic kinematic approximation does effectively limit the accurately-modeled excitations to around $\left\langle g_k^\dagger g_k \right\rangle \ll \frac{1}{2}\frac{m_e c^2}{\hbar\Omega_c}$.

[11]In the linear plasma wave limit, each plasmon number operator $f_k^\dagger f_k$ formally possesses an infinite ladder of SHO-like Langmuir states, but recall that the description breaks down at the cold wave-breaking limit,

a meaningful correspondence, in effect we assume that over any small range of momentum (compared to the probe bandwidth) or position (compared to the scale of the probe envelope), the number of atoms available for excitation greatly exceeds the number of probe photons able to induce excitation. Under these conditions the level manifold for collective excitation can be effectively assumed infinite, and the excitation and de-excitation operators effecting transitions between these levels will assume the form of harmonic oscillator-like creation and annihilation operators.

Specifically, we assume that initially the populations are given by

$$\sum N_\alpha(q; t = t_0) \approx N_{\text{A}}, \tag{6.231a}$$

$$\sum N_\beta(q; t = t_0) \approx 0, \tag{6.231b}$$

$$\sum N_\gamma(q; t = t_0) \approx 0, \tag{6.231c}$$

while throughout the period of interaction between atoms and lasers,

$$N_\alpha(q) \gg N_\beta(q+k_2-k_1) \geq N_\gamma(q+k_1) > N_\gamma(q+k_2), \tag{6.232}$$

for all atom momenta q which are appreciably populated, and where these orderings can be understood in the sense of Heisenberg-picture expectation values over the initial thermal state.

This implies that the second term in $\mathscr{H}_{\text{A}}$ can also be approximated by

$$\begin{aligned}\sum_q \tfrac{\hbar^2 q^2}{2M_{\text{A}}} \left[N_\alpha(q) + N_\beta(q+k_1-k_2) + N_\gamma(q+k_1)\right] &\approx \sum_q \tfrac{\hbar^2 q^2}{2M_{\text{A}}} N_\alpha(q; t = t_0) \\ &= N_{\text{A}} \sum_q \tfrac{\hbar^2 q^2}{2M_{\text{A}}} n_\alpha(q; t = t_0) = N_{\text{A}} \left\langle \tfrac{\hbar^2 q^2}{2M_{\text{A}}} \right\rangle_0,\end{aligned} \tag{6.233}$$

which is just some constant that can be dropped.

Next, we employ a trick consisting of switching from the second-quantized operators tracking atoms and their energy levels (or populations, in density matrix language), to (approximately) Bosonic operators effecting transitions (or coherences) between the states, and of simultaneously replacing the ground state number operators with their expectation values after all operators have been commuted and normally-ordered. Such approximation methods involving the Bosonification.[12] of elementary collective excitations were pioneered Holstein and Primakof[285] and by Bogoliubov[275, 277]

corresponding to approximately $\left\langle f_k^\dagger f_k \right\rangle \sim \frac{1}{2}\frac{m_e c^2}{\hbar\omega_p} \gg 1$ plasmons. Neglect of relativistic effects additionally presumes $\left\langle f_k^\dagger f_k \right\rangle \ll \frac{1}{2}\frac{m_e c^2}{\hbar\omega_p}$, but for typical parameters this still leaves many plasmon states.

[12]The term "Bosonization" already has a precise meaning in Fermionic field theories that differs somewhat from our procedure, so we use the term "Bosonification" instead.

That is, for the ground state operators, we ultimately make the substitution

$$\Psi_\alpha(q) \to \sqrt{n_\alpha(q)} e^{i\phi_\alpha(q)}, \tag{6.234}$$

where $\phi_\alpha(q)$ is a random phase (uncorrelated for distinct momenta), and

$$n_\alpha(q) \equiv n_\alpha(q;t) \approx n_\alpha(q;t=t_0) = \frac{1}{e^{\frac{\hbar^2 q^2 - \mu_\mathrm{A}}{2M_\mathrm{A}T_\mathrm{A}}} - 1} \tag{6.235}$$

is the average number of (Bosonic) atoms in the internal ground state with momentum $\hbar q$, for some chemical potential μ_A chosen so that $\sum\limits_q n_\alpha(q) = N_\mathrm{A}$. This also implies $\Psi_\alpha(z)^\dagger \Psi_\alpha(z) \to \frac{N_\mathrm{A}}{L}$.

Next, we define the transition operators

$$g(k) = \tfrac{1}{\sqrt{N_\mathrm{A}}} \sum_q \Psi_\alpha(q)^\dagger \Psi_\gamma(q+k), \tag{6.236a}$$

$$f(k) = \tfrac{1}{\sqrt{N_\mathrm{A}}} \sum_q \Psi_\alpha(q)^\dagger \Psi_\beta(q+k), \tag{6.236b}$$

and their unitary Fourier transforms

$$g(z) = \tfrac{\sqrt{L}}{\sqrt{N_\mathrm{A}}} \Psi_\alpha(z)^\dagger \Psi_\gamma(z), \tag{6.237a}$$

$$f(z) = \tfrac{\sqrt{L}}{\sqrt{N_\mathrm{A}}} \Psi_\alpha(z)^\dagger \Psi_\beta(z), \tag{6.237b}$$

along with an additional transition operator

$$u(k) = \tfrac{1}{\sqrt{N_\mathrm{A}}} \sum_q \Psi_\beta(q)^\dagger \Psi_\gamma(q+k) \tag{6.238}$$

in momentum space, or the corresponding operator

$$u(z) = \tfrac{\sqrt{L}}{\sqrt{N_\mathrm{A}}} \Psi_\beta(z)^\dagger \Psi_\gamma(z) \tag{6.239}$$

in position space, but which are not independent of the $f(k)$ and $g(k)$ operators.

With a little algebra, it is straightforward to show

$$\left[g(z), g(z')^\dagger\right] = \tfrac{L}{N_\mathrm{A}} \left\{N_\alpha(z) - N_\gamma(z)\right\} \delta(z-z') = \delta(z-z') \left\{1 - O\left(\tfrac{N_\gamma}{N_\mathrm{A}}\right)\right\} \approx \delta(z-z'), \tag{6.240}$$

and similarly

$$\left[f(z), f(z')^\dagger\right] = \tfrac{L}{N_\mathrm{A}} \left\{N_\alpha(z) - N_\beta(z)\right\} \delta(z-z') = \delta(z-z') \left\{1 - O\left(\tfrac{N_\beta}{N_\mathrm{A}}\right)\right\} \approx \delta(z-z'), \tag{6.241}$$

while

$$\big[g(z), f(z')\big] = 0, \tag{6.242}$$

and

$$\Big[g(z), f(z')^\dagger\Big] = \tfrac{L}{N_\mathrm{A}}\Big\{-\Psi_\beta(z)^\dagger\Psi_\gamma(z)\Big\}\,\delta(z-z') = \delta(z-z')\,O\Big(\tfrac{\sqrt{N_\beta N_\gamma}}{N_\mathrm{A}}\Big) \approx 0. \tag{6.243}$$

Of course, an arbitrarily small but finite multiple of a Dirac delta function $\delta(z-z')$ still diverges at $z=z'$, and cannot be approximated pointwise by zero, but the above commutator will be small as $\frac{\left\langle a_1^\dagger a_1\right\rangle}{n_\mathrm{A}} \to 0$ when integrated over some interval in z in any physically meaningful observable.

From the unitarity of the Fourier transforms, it follows immediately that

$$\Big[g(k), g(k')^\dagger\Big] \approx \delta_{kk'}; \tag{6.244a}$$

$$\Big[g(k), f(k')\ \Big] = 0; \tag{6.244b}$$

$$\Big[g(k), f(k')^\dagger\Big] \approx 0; \tag{6.244c}$$

$$\Big[f(k), f(k')^\dagger\Big] \approx \delta_{kk'}. \tag{6.244d}$$

That is, in the weak (unsaturated limit), these transition operators approximately satisfy CCRs for Bosonic excitations.[13]

Under the same approximations, we also find in the spatial domain that

$$\Big[g(z), u(z')\ \Big] = 0, \tag{6.245a}$$

$$\Big[g(z), u(z')^\dagger\Big] = \tfrac{\sqrt{L}}{\sqrt{N_\mathrm{A}}} f(z)\delta(z-z') \approx \tfrac{\sqrt{L}}{\sqrt{N_\mathrm{A}}} f(z)\Big[g(z), g(z')^\dagger\Big], \tag{6.245b}$$

and

$$\Big[f(z), u(z')\ \Big] = \tfrac{\sqrt{L}}{\sqrt{N_\mathrm{A}}} g(z)\delta(z-z') \approx \tfrac{\sqrt{L}}{\sqrt{N_\mathrm{A}}} g(z)\Big[f(z), f(z')^\dagger\Big], \tag{6.246a}$$

$$\Big[f(z), u(z')^\dagger\Big] = 0, \tag{6.246b}$$

suggesting that $u(z)$ acts like $\frac{\sqrt{L}}{\sqrt{N_\mathrm{A}}} g(z) f(z)^\dagger$. In fact,

$$\begin{aligned} f(z)^\dagger g(z) \approx g(z) f(z)^\dagger &= \tfrac{L}{N_\mathrm{A}}\Psi_\alpha(z)^\dagger\Psi_\gamma(z)\Psi_\beta(z)^\dagger\Psi_\alpha(z) \\ &= \tfrac{L}{N_\mathrm{A}}\Big(\Psi_\alpha(z)^\dagger\Psi_\alpha(z)\Big)\Big(\Psi_\beta(z)^\dagger\Psi_\gamma(z)\Big) \approx \tfrac{L}{N_\mathrm{A}}\Big(\tfrac{N_\mathrm{A}}{L}\Big)\Big(\Psi_\beta(z)^\dagger\Psi_\gamma(z)\Big) \\ &= \Psi_\beta(z)^\dagger\Psi_\gamma(z) = \tfrac{\sqrt{N_\mathrm{A}}}{\sqrt{L}} u(z), \end{aligned} \tag{6.247}$$

[13] Because these transition operators are of second-order in the atom creation and annihilation operators, the same Bosonic commutation relations for the collective operators are still obtained in this limit even if the atoms are instead considered Fermionic and are associated with second-quantized operators that satisfy anti-commutation relations, as long as the density of atoms is sufficiently large.

just as anticipated. In particular, by using this relation and the properties of the unitary Fourier transforms, notice that

$$u(k_2) = \tfrac{1}{\sqrt{N_{\rm A}}} \sum_q \Psi_\beta(q)^\dagger \Psi_\gamma(q+k_2) = \tfrac{1}{\sqrt{N_{\rm A}}} \sum_k f(k)^\dagger g(k+k_2). \tag{6.248}$$

Since we are only retaining resonant 3-quanta terms, when $u(k)$ is inserted into the second piece of $\mathcal{H}_{dp}$, we may keep only the $k = k_1 - k_2$ contribution, i.e., effectively assume $u(k_2) \approx \frac{1}{\sqrt{N_{\rm A}}} \sum\limits_k f(k_1-k_2)^\dagger g(k_1)$ under the assumptions of single-mode, near-resonant pump and probe fields.

Consolidating all the approximations and transformation made so far, we have, apart from unimportant constants:

$$\begin{aligned}
\mathscr{H} \approx\ & \hbar\omega_0(k_1)\, a_1^\dagger a_1 + \hbar\omega_0(k_2)\, a_2^\dagger a_2 \\
& + \left[\hbar\omega_{\beta\alpha} + \tfrac{\hbar^2(k_1-k_2)^2}{2M_{\rm A}}\right] \sum_q N_\beta(q+k_1-k_2) + \left[\hbar\omega_{\gamma\alpha} + \tfrac{\hbar^2 k_1^2}{2M_{\rm A}}\right] \sum_q N_\gamma(q+k_1) \\
& + \sum_q \left[\tfrac{\hbar^2 q(k_1-k_2)}{M_{\rm A}}\right] N_\beta(q+k_1-k_2) + \sum_q \left[\tfrac{\hbar^2 q k_1}{M_{\rm A}}\right] N_\gamma(q+k_1) \\
& - \tfrac{1}{2}\hbar\omega_{\gamma\alpha}\sqrt{2\pi n_{\rm A}}\, \tfrac{\hat{\boldsymbol{\epsilon}}_{\mu_1}\cdot\boldsymbol{\mu}_{\gamma\alpha}}{\sqrt{\hbar\omega_0(k_1)}}\, i g(k_1)^\dagger a_1 + h.c. \\
& - \tfrac{1}{2}\hbar\omega_{\gamma\beta}\sqrt{2\pi n_{\rm A}}\, \tfrac{\hat{\boldsymbol{\epsilon}}_{\mu_2}\cdot\boldsymbol{\mu}_{\gamma\beta}}{\sqrt{\hbar\omega_0(k_2)}}\, \tfrac{1}{\sqrt{N_{\rm A}}}\, i a_2\, g(k_1)^\dagger f(k_1-k_2) + h.c. \\
& - \tfrac{1}{2}\hbar\omega_{\gamma\alpha}\sqrt{2\pi n_{\rm A}}\, \tfrac{\hat{\boldsymbol{\epsilon}}_{\mu_2}\cdot\boldsymbol{\mu}_{\gamma\alpha}}{\sqrt{\hbar\omega_0(k_2)}}\, i g(k_2)^\dagger a_2 + h.c. \\
& - \tfrac{1}{4m_e} 2\pi n_{\rm A}\, \tfrac{\hbar(k_1-k_2)^2}{\sqrt{\omega_0(k_1)\omega_0(k_2)}} (\hat{\boldsymbol{z}}\cdot\boldsymbol{\mu}_{\gamma\alpha})^* (\hat{\boldsymbol{z}}\cdot\boldsymbol{\mu}_{\gamma\beta}) \tfrac{1}{\sqrt{N_{\rm A}}}\, a_2^\dagger a_1 f(k_1-k_2)^\dagger + h.c..
\end{aligned} \tag{6.249}$$

To better elucidate analogies with the magnetized plasma case, we will need to make some additional assumptions.

Atomic EIT Hamiltonian in the Cold Limit

Now we will assume that the atoms are cold in the sense that Doppler effects associated with their COM motion may be neglected. This at least requires as a necessary condition that the Doppler frequency shifts for typical atomic momenta are small compared to the transition frequencies, which is equivalent to

$$\sqrt{\tfrac{T_{\rm A}}{M_{\rm A}c^2}} \ll 1, \tag{6.250}$$

and is easily achieved for typical parameters. However, this weak necessary condition is not sufficient. Although we are using an approximation involving discrete optical and

atomic polarization modes, we recognize that the upper state has an natural lifetime limited by spontaneous emission, or a corresponding effective line-width $\delta\omega_\gamma \sim \Gamma_\gamma^{-1} \ll \omega_{\gamma\alpha}$ for transitions into or out of this state. To be truly negligible, the relevant Doppler shifts should also be small compared to this line-width, which is a far more stringent constraint. Actually, if we do everything right, the upper state will remain negligibly populated, and one-photon transitions into the upper state are suppressed, so actually we need only worry about the Doppler shift for the two-photon Raman transitions[286, 283] between the effective ground state and the metastable state. We require $\frac{\hbar^2 q k_1}{M_\mathrm{A}} \ll \delta\omega_\gamma$ for typical momenta, or equivalently

$$\sqrt{\frac{T_\mathrm{A}}{M_\mathrm{A}c^2}} \ll \frac{\delta\omega_\gamma}{|k_1 - k_2|}. \tag{6.251}$$

Admittedly, for typical alkali or other atomic systems, this restriction can fail for warm atomic vapors in cells at room temperature, but can be achieved for co-propagating pulses and sufficiently cold samples, say in magneto-optic traps.

The deleterious effects of warmer temperatures can be visualized by interpreting the vapor as a multi-species fluid, one for each momentum state occupied before the arrival of the pulses. If the lasers are tuned to a particular two-photon resonance to establish transparency, certain of the other "species" can still absorb the probe.

For the sake of in-principle comparison, we pursue the case of a cold vapor in some detail, because the similarities to the case of magnetized plasma readily emerge. The first-order Doppler-shift terms (those terms linearly proportional to the atomic momentum q in the Hamiltonian) are all dropped. Then, the only atomic operators which appear are the collective coherence operators g_1, g_2, and f, and their adjoints, as well as the Hermitian population operators $N_\beta = \sum_q N_\beta(q)$ and $N_\gamma = \sum_q N_\gamma(q)$.

With atomic DOFs restricted to the algebra of just these collective observables, we can actually express the excited and metastable population operators in terms of the coherence operators. If we try to do this directly, we encounter some divergence difficulties involving normal ordering, but we can show this indirectly. First, notice that:

$$\Big[g(k)\ , N_\beta\Big] = \quad 0; \tag{6.252a}$$

$$\Big[g(k)^\dagger, N_\beta\Big] = \quad 0; \tag{6.252b}$$

$$\Big[g(k)\ , N_\gamma\Big] = \quad g(k); \tag{6.252c}$$

$$\Big[g(k)^\dagger, N_\gamma\Big] = -g(k)^\dagger; \tag{6.252d}$$

and

$$\Big[f(k)\ , N_\beta\Big] = \quad f(k); \tag{6.253a}$$

$$\Big[f(k)^\dagger, N_\beta\Big] = -f(k)^\dagger; \tag{6.253b}$$

$$\Big[f(k)\ , N_\gamma\Big] = \quad 0; \tag{6.253c}$$

$$\Big[f(k)^\dagger, N_\gamma\Big] = \quad 0. \tag{6.253d}$$

These commutation relations are exact, and imply that as far as the dynamics of the coherence and population operators (or any functions of them) are concerned, we can make the replacements

$$N_\beta(k) = f(k)^\dagger f(k), \tag{6.254a}$$

$$N_\gamma(k) = g(k)^\dagger g(k), \tag{6.254b}$$

and, still under our assumptions of near-resonant narrow-band pump and probe fields, the Hamiltonian becomes:

$$\begin{aligned}\mathscr{H} &= \hbar\omega_0(k_1)\, a_1^\dagger a_1 + \hbar\omega_0(k_2)\, a_2^\dagger a_2 \\ &+ \hbar\omega_f\, f(k_1-k_2)^\dagger f(k_1-k_2) + \hbar\omega_g\, g(k_1)^\dagger g(k_1) \\ &+ \hbar\Omega_{ag1}\, g(k_1)^\dagger a_1 + \hbar\Omega_{ag2}\, g(k_2)^\dagger a_2 + h.c. \\ &+ \Omega_{gf}\, a_2\, g(k_1)^\dagger f(k_1-k_2) + \Omega_{af}\, a_2\, a_1^\dagger f(k_1-k_2) + h.c.,\end{aligned} \tag{6.255}$$

where we have defined certain transition frequencies

$$\omega_f = \omega_{\beta\alpha} + \tfrac{\hbar(k_1-k_2)^2}{2M_{\rm A}} \approx \omega_{\beta\alpha}, \tag{6.256a}$$

$$\omega_g = \omega_{\gamma\alpha} + \tfrac{\hbar k_1^2}{2M_{\rm A}} \approx \omega_{\gamma\alpha}, \tag{6.256b}$$

effective dipole coupling frequencies

$$\Omega_{ag1} = -i\tfrac{1}{2}\hbar\omega_{\gamma\alpha}\sqrt{2\pi n_{\rm A}}\,\tfrac{\hat{\boldsymbol{\epsilon}}_{\mu_1}\boldsymbol{\mu}_{\gamma\alpha}}{\sqrt{\hbar\omega_0(k_1)}}, \tag{6.257a}$$

$$\Omega_{ag2} = -i\tfrac{1}{2}\hbar\omega_{\gamma\alpha}\sqrt{2\pi n_{\rm A}}\,\tfrac{\hat{\boldsymbol{\epsilon}}_{\mu_2}\boldsymbol{\mu}_{\gamma\alpha}}{\sqrt{\hbar\omega_0(k_2)}}, \tag{6.257b}$$

and certain Rabi frequencies normalized to a pump-field intensity corresponding to one pump photon per quantization volume:

$$\Omega_{gf} = -i\tfrac{1}{2}\hbar\omega_{\gamma\beta}\sqrt{2\pi n_{\rm A}}\,\tfrac{\hat{\boldsymbol{\epsilon}}_{\mu_2}\boldsymbol{\mu}_{\gamma\beta}}{\sqrt{\hbar\omega_0(k_2)}}\,\tfrac{1}{\sqrt{N_{\rm A}}}, \tag{6.258a}$$

$$\Omega_{af} = \tfrac{\pi}{2m_e} n_{\rm A}\,\tfrac{(k_1-k_2)^2}{\sqrt{\omega_0(k_1)\omega_0(k_2)}}(\hat{\boldsymbol{z}}\cdot\boldsymbol{\mu}_{\gamma\beta})^*(\hat{\boldsymbol{z}}\cdot\boldsymbol{\mu}_{\gamma\alpha})\,\tfrac{1}{\sqrt{N_{\rm A}}}. \tag{6.258b}$$

This is the central result towards which we have been slowly plodding, and is quite satisfying. Although the various coefficients differ in their numerical values and their dependence on the pump and probe wavenumbers and polarizations, this Hamiltonian describing EIT in an atomic vapor has *exactly* the same form as that describing EIT in the magnetized plasma.

To accommodate finite pulse and/or probe bandwidths, multimode generalizations are possible, and should agree in form with the corresponding plasma analogs, although we have not made any explicit calculations. To obtain SVEA transport equations, we suppose that the usual eikonal conditions on the probe envelope hold, i.e., $\left|a_1^{-1}\frac{\partial}{\partial t}a_1\right| \ll \omega_1$ and $\left|a_1^{-1}\frac{\partial}{\partial z}a_1\right| \ll k_1$, and similarly for the probe, i.e., $\left|a_2^{-1}\frac{\partial}{\partial t}a_2\right| \ll \omega_2$ and $\left|a_2^{-1}\frac{\partial}{\partial z}a_2\right| \ll k_2$, but must also assume certain conditions on the medium. In order that the Bosonification of the atomic excitations can proceed locally, we require that $\left\langle a_1(z)^\dagger a_1(z)\right\rangle \ll n_\alpha(z)$ throughout the vapor, so that the number of atoms always greatly exceeds the number of probe photons in any region, and that $n_\alpha(z)\left\langle\left|\mathscr{H}^{-1}\frac{\partial}{\partial z}\mathscr{H}\right|\right\rangle^{-3} \ll 1$, so that the number of atoms within a region over which the pump and probe locally resemble plane waves is also large.

6.5 Discussion

For simplicity, let us here assume the case of discrete modes unless otherwise specified, so the creation and annihilation operators can be interpreted as changing the number of quanta rather than some number density. Any arguments or conclusions can be generalized to the continuum case, only with somewhat more cumbersome notation and terminology. We assess the extent of the parallels between EIT in the atomic vapor and the magnetized plasma, as well as point out effects so far mostly neglected which will influence details of the dynamics differently in the two cases, without altering our conclusion that at its most basic, the EIT phenomenon appears fundamentally similar in the two media, when expressed in a common language of Bosonic collective modes, or, more prosaically, coupled oscillators.

6.5.1 Comparisons of Operators and States

The pump and probe photon operators $a_1(k)$ and $a_2(k)$ have essentially the same interpretation in the atomic and plasma systems in terms of creating or annihilating quanta of excitation of the transverse EM fields at momentum $\hbar k$, even including in principle similar dressing effects from the background response of the constituent charges, although the latter is essential to the behavior of EM fields in the plasma except in the extremely underdense

regime, but quite negligible for the atomic vapor in typical parameter regimes. In both cases, the probe is normally assumed to be excited to a high-amplitude coherent state, so can be replaced with its c-number expectation value, greatly simplifying the dynamics.

In the plasma, the $f(k)^\dagger$ operators create plasmons, quanta of bare energy $\hbar\omega_p$ and momentum $\hbar k$ associated with the longitudinal density and electrostatic potential oscillations constituting Langmuir waves, i.e., to electrons oscillating in the electrostatic field.

These are truly collective phenomena, where electrons feel each other's fields, but in an ideal plasma can be treated via a mean field analysis. Plasmons can be created or destroyed through Raman scattering or other three-wave processes and instabilities, or else ponderomotively, involving photon red-shift, or the virtual annihilation of a photon at one frequency and virtual re-creation at a slightly lower one.

The $g(k)^\dagger$ operators create gyrons of bare energy $\hbar\Omega_c$ and momentum $\hbar k$, quanta of excitations associated with the transverse plasma response to the longitudinal magnetic field, i.e., to electrons gyrating about their own guiding centers. These modes are also collective, involving the distributed excitation of many plasma electrons, but involve only single-particle dynamics and no many-body interactions, i.e., they are collective but not cooperative.

Both the plasmons and gyrons are Bosons, even though the underlying particle DOFs are Fermionic. The DSP creation operator $\chi_d(k)$ involves particular linear combinations of the probe photon, plasmon, and gyron operators, so is also Bosonic.

When the density of atoms is assumed to remain much higher than the density of probe photons that can drive the resonant transitions, the reservoir of atoms in the atomic ground state is never exhausted, the system never runs out of room at the top of the ladder of excitation energies, and the excitations $f(k)^\dagger$ and $g(k)^\dagger$ behave nearly like Bosonic creation operators in the unsaturated limit. The corresponding quanta of excitation, or quasi-particles, involve changes to the internal electronic DOFs of the atoms, or in other words the polarization of the atomic medium, and with compensatory changes to the atomic momenta (involving the COM DOFs) as a consequence of the recoil that must accompany the change in internal state when these transitions are actually driven by probe and/or pump photons. In particular, $g(k)^\dagger$ effects transitions between the ground state and excited states of atoms, corresponding to a change in electronic energy of $\hbar\omega_{\gamma\alpha}$ and with a commensurate change of $\hbar k$ in the atomic momentum. The $f(k)^\dagger$ operators effect Raman transitions between ground states and metastable states, separated by internal energy $\hbar\omega_{\beta\alpha}$, with an

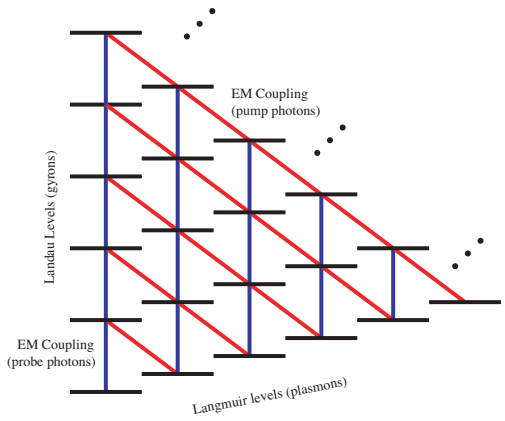

Figure 6.19. Schematic of the EIT collective state manifold, along with electromagnetic couplings

accompanying change in total momentum of $\hbar k$, transitions which are assumed forbidden at the one-photon level. The total change in energy, including the kinetic energy of the COM DOFs, depends on the initial state, but starting from rest, the kinetic energy associated with each excitation is $\frac{(\hbar k)^2}{2M_{\rm A}}$ in the non-relativistic regime assumed here. The excitations involve no inter-atomic forces or interactions, but still look collective.

Turning to the Schrödinger picture, if the initial state of the plasma is assumed completely quiescent, i.e., with negligible thermal momentum spread, and no longitudinal or transverse excitations, then, the accessible level diagram is shown schematically in Fig. 6.5.1. Although the photonic couplings effecting the various transitions are indicated, the energy spacings reflect the bare (uncoupled) energies of the plasmons and gyrons. Since the quasiparticle are identical bosons, each basis state can be uniquely specified (up to an unimportant phase) by the number of gyrons and plasmons, or equivalently, by what we have called Landau levels and Langmuir levels, respectively. All accessible states of the plasma medium can be written in terms of linear combinations of these, but we stress that arbitrary many-body momentum states of the underlying plasma electrons cannot be so generated, but only collective states of Langmuir oscillation and Landau gyration superposed on the initial (quiescent) state. If the initial radiation field (before the arrival of the probe) is also assumed to be in vacuum, then only certain collective states consistent with energy, momentum, and action conservation can be reached under the Hamiltonian evolution.

If the initial atomic state is also assumed cold, i.e.,

$$|\psi(0)\rangle = |\cdots;0;N_\alpha(0) = N_{stextA};0;\cdots\rangle\,, \tag{6.259}$$

then for sufficiently low-lying excitations (well below saturation), the manifold of accessible

atomic atomic states is exactly isomorphic to that in the plasma case. A basis can be chosen by specifying the number of one-photon, g-type (ground-excited) transitions, analogous to the number of gyrons, and the number of two-photon f-type (excited-metastable) transitions, analogous to the number of plasmons. That is, we can maintain the bookkeeping by keeping track of transitions, or coherences, rather than atoms or populations. The DSP polaritons $\chi_d(k)$ contain certain linear combinations of the probe photons and these atomic transitions, in effect shared in a symmetric and delocalized manner across all the atoms because of the Bosonic symmetry.[14]

From the perspective of collective excitations, what at first appears to be an important distinction between the atomic and plasma cases, namely that a single atom has only one available excited state or metastable state, while a single electron will have access an infinite ladder of evenly-spaced Landau levels (in the non-relativistic limit), a very large number of evenly-spaced Langmuir levels up to the cold wave-breaking limit,[15] is in fact irrelevant as long as the number density of probe photons is far less than the density of atoms. The atomic medium collectively appears to have long ladders of evenly-spaced excited-state and metastable-state transitions.

Note that the algebra of operators generated by $g(k)^\dagger$, $g(k)$, $f(k)^\dagger$ and $f(k)$ do not generate all pure states starting from the initial state $|\psi(0)\rangle$ or any other state, unlike the fundamental atomic operators generated by $\Psi_\upsilon(q)$ and $\Psi_{\upsilon'}(q)$ for υ, $\upsilon' = \alpha$, β, γ and arbitrary q, which do generate all possible pure states in the atomic Fock space considered here (or at least a dense covering of these states), starting from the many-body vacuum state or indeed any pure state. (This is really how we define the Fock space for identical atoms). In particular, the transition operators only generate those states that can be reached by absorption of suitable combinations of probe and pump photons, so must remain consistent with conservation of energy, momentum, total atom number, and actions.

If the initial state of the plasma or the atomic vapor has some non-negligible thermal spread in momentum, and so consists of some thermal or other mixed state, then the accessible states no longer form some manifold of pure states as indicated in Fig. 6.5.1, but the diagram can still be interpreted for both systems to represent the generators of the algebra of excitation operators which produce the possible transition out of the initial pure

[14]This symmetry still emerges even if the atoms are assumed from the beginning to be distinguishable atoms, leading to Dicke-like collective atomic states.

[15]In the cold, non-relativistic limit, recall that this number is approximately $N_e \frac{m_e c^2}{\hbar \omega_p}$, where N_e is the total number of electrons in the interaction volume.

state. The creation operators $f(k)^\dagger$ and $g(k)^\dagger$, then effect collective excitations out of this noisy background of atoms or electrons.

6.5.2 Nonlinear Effects

The exact analogy between plasma and atomic EIT occurs in a cold, linear dynamical limit, in which (with the pump regarded as a classical control field), the Hamiltonian is completely positive definite and bilinear in the annihilation and creation operators corresponding to various collective excitations with evenly-spaced, harmonic-oscillator-like energy levels. In general, any nonlinearities that have so far been neglected will enter in different ways for the two systems.

For plasma-based EIT, probably the most important nonlinearities so far blithely ignored are relativistic detunings or anharmonicities. Classically, as the velocities of the electrons increase, the effective plasma frequency is shifted to $\omega_{p\,\mathrm{eff}} \approx \frac{\omega_p}{\sqrt{\gamma}}$ while the effective cyclotron frequency becomes $\Omega_{c\,\mathrm{eff}} \approx \frac{\Omega_c}{\gamma}$, exhibiting an even stronger detuning. These effects will tend to shift the beat-waves out of resonance and inhibit EIT. Quantum-mechanically, this corresponds to dynamic frequency-shift terms in the Hamiltonian: terms of the form $(f^\dagger f)^2$, $(g^\dagger g)^2$, and higher-order powers in a series expansion. For cold systems, such nonlinear shifts will invalidate the exact one-to-one correspondence between manifolds of collective states for plasma excitations and their atomic analogs. Neglecting any inter-atomic interactions, $\mathcal{N}$ atoms excited to a given level have $\mathcal{N}$ times more energy than one atom occupying that level, even in the presence of a pump field sufficiently intense to non-perturbatively distort the energy levels themselves. But with relativistic detuning, excitation of $\mathcal{N}$ electrons into the first Landau level will no longer be energetically equivalent to the excitation of one electron into the $\mathcal{N}$th level.

In the plasma, other problems emerge with increasing pump or probe power. Raman terms that directly couple the ground and metastable states become increasingly important, and as we have seen, shift the index-of-refraction at the cyclotron resonance further from its value in the unmagnetized case. Higher-order (multi-photon, or many-wave) scattering/decay process can occur, and any number of nonlinear terms may become important in the Hamiltonian – for example, the two-plasmon decay instability, involving terms of the form $(a^\dagger)^2(f)^2$. Also, the RWA will perform increasingly poorly, as coupling between left-handed and right-handed EM waves and the cyclotron motion becomes important. In the atomic system, sufficiently high probe strengths can saturate the vapor, while high pump

strengths can lead to other optical nonlinearities. In the plasma, the Langmuir wave can break if driven excited too strongly.

While the nonlinearities enter somewhat differently in the different systems, and break the essentially exact analogy between EIT in the atomic and plasma media, this will not invalidate our conclusion that the EIT mechanism is fundamentally similar in the two cases, despite differences in microscopic details.

6.5.3 Thermal and Decoherence Effects

EIT relies on carefully established coherences between harmonic-oscillator-like DOFs, so hydrodynamic/thermal/damping effects that can degrade these coherences can also impair the resulting EIT. Even if we can neglect all ongoing damping and decoherence processes over the duration of pulse propagation, thermal spread in the initial momentum distribution can be harmful. As we have previously mentioned, in the atomic case, Doppler broadening that is larger than the intrinsic width of the two-photon resonance will lead to imperfect cancellation of the transition amplitudes into the upper state and degrade transparency. Mathematically, the complications arising from initial momentum spread are immediately evident. If all the atoms are initially condensed into the $q = 0$ state, then the convolution over momentum appearing in the definitions of $g(k)$ and $f(k)$ can be dropped, and these operators can be written as local bilinear products of the underlying atomic operators in both position and momentum space.

In the case of the magnetized plasma, EIT relies crucially on the fact that the longitudinal motion influences the electron response to the transverse fields via doppler shifts, so any background thermal motion in addition to the coherent Langmuir oscillation can be deleterious, although the specific effects of doppler shifts enter a bit differently in the two cases because of the different microscopic physics.

In the magnetized plasma, a preliminary classical analysis of thermal effects has been performed in [287]. Hamiltonian treatment of these thermal effects would require a more complicated formalism than that developed here, since our representation for the plasma wave assumed that the charge "sheets" associated with electrons in the 1D limit never cross, which can in fact happen if they are subject to thermal motion in addition to the coherent plasma oscillation, although the effects do tend to average away. It appears that fairly simple 1D Hamiltonian fluid models of an initially Maxwellian plasma are still possible, only with somewhat more complicated relationships between the canonical coordinates and

the physical fluid observables, i.e., velocity and either particle density or the longitudinal electric field.

So far we have neglected the effects of non-Hamiltonian damping and decoherence terms. In a Schrödinger-picture, the former are associated with decays of the diagonal elements of the density matrix while the latter are associated with dephasing of the off-diagonal elements. In the atomic system, the damping is due primarily to spontaneous emission out of the Λ-level EIT manifold, but may also include the effects of collisions or other processes that transfer energy to DOFs not included as part of the system. Decoherences can arise from interatomic effects or environmental jitter. In the plasma, damping and dephasing both could arise from collisional effects, but in typical plasma regimes, are dominated by collisionless wave-particle effects such as Landau damping of the longitudinal plasma wave or cyclotron damping of the transverse waves.

However, because obviously one of the more interesting things about EIT is the ability to effectively slow or trap the light pulse, total interaction times for the probe in the medium may not be very small compared to the characteristic time-scales for the damping or detuning. Note that atomic EIT is largely immune to the finite lifetime of the upper level because by design it is never populated, and can remain immune to the dephasing times between the upper and lower levels as long as these time-scales are sufficiently long compared to the inverse Rabi frequencies. A rough estimate suggests the proper condition is

$$\frac{\Omega_{\mathrm{Rabi}_1}\Omega_{\mathrm{Rabi}_2}}{\Gamma_{\alpha\gamma}\Gamma_{\beta\gamma}} \geq \frac{N_{\mathrm{A}}\sigma_{\mathrm{abs}}}{\mathcal{A}} \gg 1. \tag{6.260}$$

Transmission and storage are much more sensitive to decay of the metastable level, or worse, dephasing of the polariton coherence between the metastable and ground levels. Clearly the maximum storage time τ_s for trapping the pulse *qua* polariton is limited by the lifetime of the metastable state, i.e.,

$$\tau_s < \Gamma_{\beta\beta}^{-1}, \tag{6.261}$$

but because of the collective nature of the DSP state, a pulse involving about N_β material excitations will de-phase about N_β times faster than a single-particle state, so the storage time must be limited to

$$\tau_s \ll \left(\Gamma_{\alpha\beta}\left\langle N_\beta \right\rangle\right)^{-1} \tag{6.262}$$

to avoid decoherence. Analogous remarks hold for the plasma case. It should be relatively insensitive to any cyclotron damping, but much more sensitive to linear or nonlinear Landau damping of the plasmons.

In both cases, the higher-order moments are expected to degrade faster, so "classical" storage of just the average envelope should be more robust than "quantum" storage of the actual coherences or full quantum state.

Despite all the ominous sounding news, we expect that the degradation due to moderate thermal effects will be gradual, not catastrophic. Almost all of plasma physics as a distinct state of matter relies on screening to provide a separation between collective and single-particle DOFs with some residual interaction between them, as pointed out by Landau[288], Bohm and Pines[289], and other pioneers of the field. In the presence of moderate thermal effects, collective modes tend to persist, only with some real frequency shift, corresponding to thermal dispersion, and an imaginary frequency shift, corresponding to Landau or other collisionless damping. By "moderate" temperature we mean that the division between collective and single-particle DOFs remains meaningful, requiring perturbations of wavelength $\lambda > \lambda_e$, where

$$\lambda_e = \left(\frac{T_e}{4\pi n_0 e^2} \right)^{1/2} \tag{6.263}$$

is the electron Debye screening length in Gaussian units except for temperature, which is measured in energy units. A full analysis of thermal effects is probably needed, but is left for future research.

6.5.4 Some Final Comparisons

Translation of EIT dynamics into a mode-based, collective description reveals fundamental similarities and analogies, regardless of whether the important interactions are cooperative or single-body, or classical versus quantum. At its most basic, EIT requires a polarizable medium supporting multiple coherent pathways for spatially-localized excitation. From a quantum mechanical perspective, it requires appropriate bound states coupled by electric-dipole or other transitions.

In individual atoms, the restoring force on a bound electron is provided by Coulomb interactions with its own nucleus. In the plasma, the transverse restoring force is provided by the external magnetic field, and is not derivable from a potential, but rather is velocity-dependent. However, the plasma electron orbiting its guiding center is quite similar to the valence electron orbiting its nucleus. Longitudinal restoring forces in the plasma are provided by Coulomb interactions, but in the self-consistent field of all the electrons and background ions.

The longitudinal dynamics are truly collective in the plasma, but in the cold, 1D limit

are also expressible in terms of single electron Lagrangian coordinates. Both the transverse gyration in the plasma and the atomic polarization are non-cooperative effects, but are describable by collective modes.

Atomic EIT first emerges as a single-atom effect, but propagation in bulk vapor is more naturally described collectively, revealing Bosonic polariton modes which excite symmetric Dicke-like states in the cold, optically dense, unsaturated limit.

At the level of the collective description, the differences between EIT in atomic systems and magnetized plasmas are largely of degree rather than kind. In typically atomic systems, involving an effective ground state and metastable level split by a hyperfine transition, but the ground and excited states coupled by an optical frequency, the various frequencies satisfy

$$\Omega_{\text{Rabi1}} \ll \Omega_{\text{Rabi2}} \ll \omega_{\beta\alpha} \ll \omega_{\gamma\beta} \sim \omega_2 \ll \omega_{\gamma\alpha} \sim \omega_1, \tag{6.264}$$

while the temperature will typically satisfy $T_A \ll \hbar\omega_{\gamma\alpha}$. (Often $T_{\text{A}} \ll \hbar\omega_{\beta\alpha}$ as well, but if not the ground state can still be populated by appropriate optical pumping.)

In the plasma, however, the frequencies will tend to follow the ordering

$$\Omega_{\text{Rabi1}} \ll \Omega_{\text{Rabi2}} \lesssim \omega_p \lesssim \omega_2 \sim \Omega_c - \omega_p \lesssim \Omega_c \sim \omega_1 \tag{6.265}$$

while a realistic temperature might satisfy $T_e \gtrsim \hbar\omega_p$ and $k_p\lambda_e \lesssim 1$.

So while in the case of the atomic vapor the intense pump field can be taken to satisfy the vacuum dispersion relation, in the case of the plasma it is not so far below the cyclotron resonance that its effects of the can be neglected, so the pump carrier oscillation should be taken to satisfy the usual linear dispersion relation for a magnetized rather than unmagnetized plasma.

Also, while both the atomic and plasma media will be subject to Doppler broadening, the plasma will otherwise be more susceptible to the effects of additional thermal damping, relativistic detuning, and non-resonant terms and stronger competing two-photon or other multi-wave terms. Because of the latter, exact cancellation of the cyclotron response does not occur at the bare resonance. We have attempted to account for this difference, with moderate success, by a simple approximation incorporating this shift but otherwise neglecting the effects of the competing scattering terms.

6.5.5 A "Ho Hum" Interpretation

With the same facts regarding the parallels between EIT in atomic vapors and magnetized plasmas, one can arrive at rather different interpretations, depending on the perspective adopted. One could downplay these analogies by reasoning of the following sort: EIT in classical harmonic oscillators can be described classically, without any recourse to quantum mechanics, or else quantum mechanically, if the system is quantized in the usual way. EIT in quantum harmonic oscillators can of course be described quantum mechanically, but most features can also be reproduced classically. EIT behavior simply requires EM fields in a suitable polarizable medium, or, more generally, any suitably-coupled harmonic oscillators, whether classical or quantum, with multiple channels for excitation that can superpose and interfere. Under the assumption of linearity, this superposition/interference can manifest in quantum amplitudes or histories, or in classical fields or their corresponding sources.

Atomic or plasma EIT may be interpreted either in high-brow, quantum terms (dressed operators, dark states, etc.) or low-brow, classical terms (force balance, and "charge-on-a-spring" models). According to this view, when you have seen one harmonic oscillator, you have essentially seen them all: the Hamiltonian/Heisenberg equations of motion are the same quantum mechanically as classically, with c-numbers (classical phase-space observables) replaced by q-numbers (operators), and Poisson brackets replaced by commutators. Looking back, we see that the analysis never actually needed to commit to commutators versus brackets, so we are free to choose a less exotic classical description.

Most features of EIT depend only on the nature of the coupled modes, not on the quantum states of those modes. In "second-quantization" formalism applied to bosons, most truly quantum effects do not really appear in the first stage of quantization (involving normal mode decomposition, dressing, etc.), but only at the second stage, and only then if those modes are excited into some non-classical state such as a Fock (number) state, squeezed state, entangled Bell-type state, or other states which are not Glauber coherent states or classical (i.e., positive and normalizable) statistical mixtures thereof.

6.5.6 A "Gee Whiz" Interpretation

Alternatively, we can take as the moral of this story the fact that any wave interference may be interpreted, ultimately, as a quantum interference. In particular, the EIT phenomenon helps to reveal connections between two strands of Bohr's notion of complemen-

tarity: wave/particle duality, and classical/quantum correspondence, in both its statistical and dynamical aspects.

Note that waves and particles have *opposite*, or *complementary* Correspondence Principle limits. In the classical limit, waves exhibit many of the very properties – coherence, interference, diffraction, spin/polarization, Heisenberg uncertainty for Fourier-conjugate variables – that make quantum particles seem so weird, while in the quantum regime waves may seem to exhibit properties – discreteness effects – which are regarded as normal for classical point particles. Neither quantum entanglement nor ontological chance,[16] as opposed to merely epistemic uncertainty, are ever exhibited classically by either waves or particles, and therefore might be regarded as fundamental defining features of quantum mechanics shared by both field and particle.

Truly fermionic degrees-of-freedom must be particle-like (statistically and dynamically) in the classical limit, while bosonic degrees-of-freedom may appear wave-like or particle-like in the classical limit, depending on circumstances. Conversely, a classical coherent wave is really behaving like a quantum-degenerate bosonic system, while a classical particle system is behaving like a non-degenerate fermionic (or bosonic) system.

A classical wave-limit requires collective Bosonic degrees-of-freedom, although these can be carried by underlying bosons (as with the EM field in vacuum) or fermions (as with plasmons). That is, certain collective excitations of fermionic particles may be bosonic, and exhibit persistent wave-like behavior in the classical limit even as the remaining individual degrees-of-freedom of the very same matter behave as classical, Maxwell-Boltzmann particles. This is an essential feature of plasma physics.

Therefore, we may argue that quintessentially wave-like properties (superposition, interference, coherence), whenever observed, may be attributed to persistent quantum coherence of some collective bosonic degrees-of-freedom. Again, because the Heisenberg equations of motion for quantum harmonic oscillators are formally analogous to the Hamilton's equations of motion for classical oscillators, and because we never needed to commit to c-numbers versus q-numbers, or Poisson brackets versus commutators, we are equally free to use a quantum description, interpretation, or formalism for EIT. Only in the truly classical limit, participating modes will all be described by a positive Glauber-Sudarshan P quasi-distribution function, which are the most classical-looking distributions allowed by quantum mechanics. In this classical case, EIT can adiabatically transfer classical action and information in envelopes between electromagnetic and medium modes, but cannot transfer entanglement or

[16] Also called essential, aleatoric, or Tychist chance, in various contexts.

other non-classical information between fields and matter and back again, which is being pursued in atomic systems.

6.5.7 An Editorial Aside

EIT is but one of many parallels evident between the quantum mechanics of atoms or molecules and the (mostly classical) physics of plasmas and beams, suggesting further opportunities for cross-fertilization. Both types of systems support versions of Raman scattering, and it is obviously by intentional analogy that these processes go by the same name. Stimulated and spontaneous emission in atoms have their analogs in particle beams traveling through wigglers, while the free electron laser (FEL) is analogous in many respects to its more familiar atomic predecessors – in fact, the first analyses of FELs were completely quantum mechanical, before it was realized that a classical description was simpler but still captured the essential notion of gain through stimulated emission in an "inverted" medium. For sufficiently intense EM field strengths, nonlinear solitons of various sorts can propagate in plasmas and in atomic media.

What are called quasi-particles or elementary excitations in atomic or condensed matter physics are known as collective modes in plasma physics. Energy and momentum conservation laws for these quasi-particles correspond to the temporal or spatial resonance conditions in plasma, while conservations laws for certain combinations of the number of quanta or quasi-particles are analogous to the Manley-Rowe relations of classical plasma physics, expressing wave action conservation.[17]

Both types of systems are amenable in certain regimes to WKB-type approximations or other long-wavelength asymptotic techniques. Both systems have adiabatic theorems and invariants, and what are typically called Landau-Zener transitions in atomic physics go by the more general description of mode conversion in plasma physics. Stimulated Raman Adiabatic Passage (STIRAP) in atomic system is closely related to adiabatic passage in wave phenomena in plasmas, while quantum ladder-climbing has its classical counterpart in autoresonance[290].

Recently, profound analogies were pointed out between the longitudinal and transverse dynamics of ultra-old atoms in an annular magnetic trap, and the betatron and synchrotron

[17]Professor Allan Kaufman has recounted that even the great Marshall Rosenbluth, one of the Founding Fathers of plasma physics, initially expressed disbelief that the Manley-Rowe relations, that looked so manifestly quantum mechanical in form and seemed to suggest so compellingly a quantum mechanical interpretation, could nevertheless be valid within a completely classical description of the plasma.

motion of elementary particles in high-energy storage rings, despite differences in characteristic length-scales of six or seven orders-of-magnitude, and in energy-scales of fifteen to twenty orders-of-magnitude[291, 292].

6.6 Conclusions

We have seen how, in a certain idealized limit (i.e., an unsaturated, non-relativistic, low-temperature regime with an initially quiescent medium, and within the rotating wave approximation), EIT in magnetized plasmas and atomic vapors are essentially equivalent, apart from the obvious differences in the relative orderings and scales of certain frequencies, coupling strengths, and other parameters, which in the case of plasma-based EIT leads to more salient signatures of certain scattering terms which in effect compete with the EIT process, resulting in imperfect cancellation of the resonant response precisely at the bare resonance. In both systems, one can use the intense pump field to control the effective transmission coefficient of the probe field, to control the effective speed of the probe, or to adiabatically tune the excitation between an electromagnetic and material form, or even to store the action and information initially contained in the probe inside the medium.

Using a Hamiltonian description for each system, the parallels emerge in a quantum mechanical Heisenberg operator formalism where excitations "look" collective, whether they are associated with free or bound electrons, and whether they involve truly cooperative or just aggregated single-particle effects. However, relativistic and other nonlinearities, thermal and other noise effects are expected to enter into the dynamics rather differently in the two cases, although this should not detract greatly from our conclusion that at its most basic, EIT in the two systems is fundamentally similar. In both cases, the transparency can be attributed to the quantum interference between different pathways of excitation, or equivalently in terms of the pump-induced dressing of the system leading to a dark-state polariton mode which is largely immune to absorption.

The differences between the systems are mostly of degree rather than kind, and in the collective picture do not result from the fact that each electron can participate in an entire ladder of Landau levels or Langmuir levels, while a single atom supports only one excited and one metastable state. Electrons orbiting about their gyro-centers, and oscillating longitudinally in their space charge fields, behave much like conventional atoms.

The differences of scale in parameter regimes typical of the two systems does of course

lead to different quantitative, and possibly even qualitative behavior. In particular, the frequency separations in the plasma case are much smaller than in the atomic case, raising questions about the validity of the RWA, the importance of the direct two-photon $\boldsymbol{A}\cdot\boldsymbol{A}$ terms, the influence of thermal effects, and the consequences of the shift in the transparency maximum away from the unperturbed resonance, resulting from competing scattering processes arising out of these terms. We have begun to explore these questions, but a more thorough investigation is left for future work.

Despite its having a quantum mechanical description almost identical to the atomic case, magnetized plasma EIT nevertheless has a completely satisfactory classical description, suggesting it might be interpreted as a classical analog of what Marlan Scully[243] in the quantum context called a phaseonium, a macroscopic medium supporting long-range phase coherence, raising fundamental questions about the nature of interference and superposition. Plasma EIT offers some conceptual insight, if not easy answers, into these fascinating questions about the relationship between plasma and atomic systems, between collective and individual dynamics, and between the classical and quantum worlds.

Acknowledgements

We would like express gratitude to Ryan Lindberg, for many helpful discussions, Min-Sup Hur, for providing PIC simulations while working as a Post-Doctoral Fellow under Professor Wurtele at the Center for Beam Physics at the LBNL, and Professor Gennady Shvets in the Department of Physics at the University of Texas, Austin, for useful discussions and critiques.

6.7 Mathematical Appendix: Generalized Bogoliubov/Tyablikov Transformations for Many-Degree-of-Freedom Quadratic Hamiltonians

While existence results for so-called normal forms of quadratic Hamiltonians are well known[293, 294, 295, 296, 297], constructive algorithms for actually generating the required linear canonical transformations and for producing the transformed Hamiltonian are surprisingly rare, except in the cases of a one or two DOF problems, where Bogoliubov-Tyablikov (BT) transformations[275, 276, 277] are well known, or where the needed lin-

ear combinations obviously involve Fourier or other familiar transforms. See, however, [298, 299, 300, 301, 302, 303, 304] and references therein for some more general results on this and related problems.[18] In the most general case, it is clear that the linear part of the Hamiltonian cannot always be removed by completing the square, nor can the quadratic part always be diagonalized[19] in terms of true eigenmodes and eigenfrequencies, despite occasional assumptions to the contrary in the literature. We have developed our own techniques for the general problem, involving the use of the Moore-Penrose pseudo-inverse to find and then transform away the maximally-removable component of the linear contribution by adding appropriate offsets to the mode operators, and then invoking a series of similarity transforms to reduce the bilinear part to a Symplectic Jordan Normal Form, leading within each Jordan block to one true eigenmode and possibly a number of generalized eigenmodes of the same eigenfrequency. Degeneracy of eigenfrequencies is therefore a necessary but not sufficient condition for the bilinear part of the Hamiltonian to be defective, or non-diagonalizable.

However, in the case of our EIT model, only bilinear terms appear, and with the RWA imposed the Hamiltonian is necessarily diagonalizable, while even with non-RWA 2-wave terms retained, for any reasonable parameter values the dressed interaction eigenfrequencies seem to remain non-degenerate,[20] and the Hamiltonian remains diagonalizable. So here we summarize results for the simple case where linear terms are absent, and a complete set of eigenmodes is assumed to exist.

To avoid any confusion between operations on quantum mechanical observables or on matrices of operators or numbers, throughout this section we will use the notation a^* to denote the complex-conjugation of any scalar c-number a or the Hermitian conjugation (adjoint) of any scalar q-number a, the notation M^* to denote in-place conjugation of all the elements (q-number or c-number) of some matrix M, M^{T} to denote the usual matrix transposition, and $M^\dagger = M^{*\mathrm{T}}$ to denote the Hermitian transpose of the matrix M.

In particular, consider a Cartesian vector $\boldsymbol{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$ containing a finite number

[18] In the numerical methods community, interest is growing in what are called structured eigenvalue problems, or more generally structure-preserving matrix computations.

[19] In this context, diagonalization refers to the process of dressing the original harmonic oscillator mode operators via linear canonical transformations to yield a set of uncoupled oscillators. It occurs at the level of first quantization, whereas in second quantization we then adopt the Dirac-Weyl canonical quantization procedure on each uncoupled dressed mode separately, and can formally diagonalize the quantum mechanical versions of the Hamiltonian by considering a multi-mode number-state basis.

[20] Because of the Avoided-Crossing Theorem, generically one must simultaneously adjust two parameters (like pump strength and pump frequency, or pump strength and mode wavenumber) to find a degeneracy. Besides, even if a degeneracy in the eigenfrequencies is encountered, diagonalizability tends not to fail until the coupling terms become sufficiently large compared to the diagonal terms (bare frequencies).

n of discrete modes described by annihilation operators a_j, and the corresponding vector $\boldsymbol{a}^* = (a_1^\dagger, \ldots, a_n^\dagger)^{\mathrm{T}}$ of the conjugate $a_j^\dagger$ creation operators. The same results will hold in the classical as the quantum case. Generalizations to a countably infinite set of modes, or a continuum set of modes may be possible, but some subtle existence and uniqueness issues arise which we need not worry about here. The operators are assumed to satisfy the CCRs, which in an obvious matrix notation can be written as

$$\begin{bmatrix} [\boldsymbol{a}\ , \boldsymbol{a}^{\mathrm{T}}] & [\boldsymbol{a}\ , \boldsymbol{a}^\dagger] \\ [\boldsymbol{a}^*, \boldsymbol{a}^{\mathrm{T}}] & [\boldsymbol{a}^*, \boldsymbol{a}^\dagger] \end{bmatrix} = J = J_{2n} = \begin{bmatrix} 0_n & I_n \\ -I_n & 0_n \end{bmatrix} \tag{6.266}$$

where $I_n \in \mathbb{R}^{n\times n}$ is the usual identity matrix, $0_n \in \mathbb{R}^{n\times n}$ is the zero matrix, so $J \in \mathbb{R}^{2n\times 2n}$ is the canonical symplectic tensor, and therefore satisfies

$$\det J = 1, \tag{6.267}$$

and

$$J^{-1} = J^{\mathrm{T}} = J^\dagger = -J, \tag{6.268}$$

so that J is real, unimodular, anti-idempotent, anti-Hermitian, and unitary.

In terms of the bare modes, we consider a Hamiltonian which, apart from a constant, can be written as

$$\mathscr{H} = \tfrac{1}{2} \begin{pmatrix} \boldsymbol{a} & \boldsymbol{a}^* \end{pmatrix} M \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{a}^* \end{pmatrix}, \tag{6.269}$$

where the c-number coupling matrix $M \in \mathbb{C}^{2n\times 2n}$ can be written in partitioned form as

$$M = \begin{bmatrix} B^* & A^* \\ A & B \end{bmatrix}, \tag{6.270}$$

where $A \in \mathbb{C}^{n\times n}$ is Hermitian, i.e., $A^\dagger = A$, and $B \in \mathbb{C}^{n\times n}$ is (complex) symmetric, i.e., $B^{\mathrm{T}} = B$, but are otherwise arbitrary, so that M is complex symmetric, i.e., $M^{\mathrm{T}} = M$, and JM satisfies

$$J(JM)^{\mathrm{T}} J^{-1} = -JM, \tag{6.271}$$

and therefore JM belongs to the set variously known as J-antisymmetric, (complex) Hamiltonian, or infinitesimally symplectic matrices.

We seek linear transformations to dressed operators $\boldsymbol{\chi}$, i.e.

$$\boldsymbol{\chi} = C\boldsymbol{a} + S\boldsymbol{a}^* \tag{6.272}$$

for some c-number dressing matrices $C \in \mathbb{C}^{n\times n}$ and $S \in \mathbb{C}^{n\times n}$ which are invertible, preserve the CCRs, and de-couple the Hamiltonian, such that in the dressed basis

$$\mathscr{H} = \tfrac{1}{2}\begin{pmatrix}\boldsymbol{\chi} & \boldsymbol{\chi}^*\end{pmatrix}\tilde{M}\begin{pmatrix}\boldsymbol{\chi}\\ \boldsymbol{\chi}^*\end{pmatrix}, \tag{6.273}$$

where the dressed c-number coupling matrix $\tilde{M} \in \mathbb{C}^{2n\times 2n}$ can be written in partitioned form as

$$\tilde{M} = \begin{bmatrix} 0_n & \hbar\Omega \\ \hbar\Omega & 0_n \end{bmatrix} \tag{6.274}$$

for some diagonal eigenfrequency matrix $\Omega \in \mathbb{R}^{n\times n}$. If we expand the dressing relations into the doubled form

$$\begin{pmatrix}\boldsymbol{\chi}\\ \boldsymbol{\chi}^*\end{pmatrix} = F\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{a}^*\end{pmatrix}, \tag{6.275}$$

where $F \in \mathbb{C}^{2n\times 2n}$ is the partitioned c-number matrix

$$F = \begin{bmatrix} C & S \\ S^* & C^* \end{bmatrix}, \tag{6.276}$$

then the dressing transformations are both invertible and canonical if and only if F belongs to the group $Sp\,(2n,\mathbb{C})$ of complex J-orthogonal, or symplectic matrices, i.e., the condition

$$F^{\mathrm{T}}JF = J \tag{6.277}$$

holds. The inverse F^{-1} can then be written as

$$F^{-1} = JF^{\mathrm{T}}J^{-1} = -JF^{\mathrm{T}}J, \tag{6.278}$$

which effects the transformation from the dressed back to the bare mode operators.

Equating the bare and dressed forms of the Hamiltonian, and using the fact that the modal operators themselves are linearly independent as abstract vectors in the algebra of observables, we deduce

$$M = F^{\mathrm{T}}\tilde{M}F \tag{6.279}$$

which is not quite in the form of a similarity transformation, but rather of a congruency transformation, so we cannot yet determine F and Ω directly by solving an eigenvalue problem for M, for which we have some analytical (and, if necessary, numerical) machinery. However, if we multiply both sides by J, we find

$$JM = JF^{\mathrm{T}}\tilde{M}F = JF^{\mathrm{T}}(J^{-1}J)\tilde{M}F = (JF^{\mathrm{T}}J^{-1})(JM)F = F^{-1}(JM)F, \tag{6.280}$$

so that the elements of $\pm\hbar\Omega$ are the eigenvalues of the JM matrix, while the columns of F^{-1} are the corresponding eigenvectors. We see that we have doubled the number of eigenvalues, but because of the symplectic nature of the problem it is easy to verify that the characteristic polynomial $\phi_{JM}(\omega)$ of the matrix JM is a function of ω^2 only, i.e., $\phi_{JM}(\sqrt{x})$ is an nth-degree polynomial in the argument x.

If these eigenfrequencies are non-degenerate, then the eigenvectors are uniquely determined up to overall scalings, which can be chosen to ensure symplecticity. If there are any degeneracies, but the Hamiltonian remains fully diagonalizable, then appropriate eigenvectors can be chosen by a Gram-Schmidt-like procedure, except that the eigenvectors are made J-orthonormal (with respect to the antisymmetric symplectic form) rather than orthonormal with respect to a standard inner product, or symmetric form. If a complete set of eigenvectors cannot be found within each degenerate sub-space, then the Hamiltonian cannot be diagonalized, but can be put into a symplectic Jordan Normal Form where the residual couplings between the degenerate dressed modes take particularly simple form, but such generalizations are not needed here.

Note that the associated eigenproblem can also be derived by demanding that

$$[\chi_j, \mathscr{H}] = \hbar\omega_j \chi_j \tag{6.281}$$

and then taking various further commutators with the a_j and a_j^* to deduce appropriate conditions in terms of the expansion coefficients. The eigenvectors are then normalized so that

$$[\chi_j, \chi_k] = 0 \tag{6.282}$$

and

$$[\chi_j, \chi_k^*] = \delta_{jk} \tag{6.283}$$

hold for all j , $k = 1, \ldots, n$.

In the case of the RWA, the anti-resonant couplings are dropped, i.e., we set $B = 0_n$, and it is clear that the dressing transformations will not mix creation and annihilation operators, so that we can take $S = 0_n$ as well. Instead of diagonalizing a $2n$-dimensional complex Hamiltonian matrix JM in terms of a symplectic matrix of eigenvectors, we need only diagonalize the n-dimensional Hermitian matrix $A = A^\dagger$ by using a unitary dressing matrix C.

Bibliography

[1] P. Maine, D. Strickland, P. Bado, M. Pessot, and G. Mourou. Generation of ultahigh peak power pulses by chirped pulse amplification. *IEEE Journal of Quantum Electronics*, 24(2):398–403, 1988.

[2] D. Strickland and G. Mourou. Compression of amplified chirped optical pulses. *Optics Communications*, 56(3):219–221, 1995.

[3] Eric Esarey, Philip Sprangle, Jonathan Krall, and Antonio Ting. Overview of plasma-based accelerator concepts. *IEEE Transactions on Plasma Science*, 24(2):252–288, 1996. and references cited therein.

[4] T. Tajima and J. M. Dawson. Laser electron accelerator. *Physical Review Letters*, 43(4):267–270, 1979.

[5] E. Esarey, P. Sprangle, J. Krall, A. Ting, and G. Joyce. Optically guided laser wakefield acceleration. *Physics of Fluids B*, 5(7):2690–2697, 1993.

[6] G. Shvets, N. J. Fisch, A. Pukhov, and J. Meyer-ter Vehn. Generation of periodic accelerating strcutures in plasma by colliding laser pulses. *Physical Review E*, 60(2):2218–2223, 1999.

[7] Eric Zeek, Kira Maginnis, Sterling Backus, Ulrich Russek, Margaret Murnane, Gerard Mourou, Henry Kapteyn, and Gleb Vdovin. Pulse compression by use of deformable mirrors. *Optics Letters*, 24(7):493–495, 1999.

[8] P. Chen, A. Spitkovsky, T. Katsouleas, and W. B. Mori. Transformer ratio and pulse shaping in laser wakefield accelerator. *Nuclear Instruemnts and Methods in Physics Research A*, 410(3):488–492, 1998.

[9] Anatoly Spitkovsky and Pisin Chen. Laser shaping and optimization of the laser-plasma interaction. In Patrick L. Colestock and Sandra Kelly, editors, *The NInth Workshop on Advanced Accelerator Concepts*, volume 569, pages 183–194Advanced Accelerat, Santa Fe, New Mexico (USA) 10-16 June 2000, 2000. AIP.

[10] Anatoly Spitkovsky and Pisin Chen. Longitudinal laser shaping in laser wakefield accelerators, 15 April 2000.

[11] Anatoly Spitkovsky and Pisin Chen. Longitudinal laser shaping in laser wakefield accelerators. *Physics Letters A*, 296(2):125–130, 2002.

[12] Pisin Chen, J. J. Su, J. M. Dawson, K. L. F. Bane, and P. B. Wilson. Energy transfer in the plasma wake-field accelerator. *Physical Review Letters*, 56(12):1252–1255, 1986.

[13] Yiton T. Yan and Hudong Chen. Nonlinear solution for optimal shaping of the driving electron beam in the plasma wake-field accelerator. *Physical Review A*, 38(3):1490–1494, 1988.

[14] H. H. Kuehl, C. Y. Zhang, and T. Katsouleas. Interaction of a weakly nonlinear laser pulse in a plasma. *Physical Review E*, 47:1249–1261, 1993.

[15] C. D. Decker and W. B. Mori. Group velocity of large amplitude electromagnetic waves in plasma. *Physical Review Letters*, 72:490–493, 1994.

[16] C. D. Decker and W. B. Mori. Group velocity of large amplitude electromagnetic waves in plasma. *Physical Review E*, 51:1364–1375, 1995.

[17] E. Esarey, P. Sprangle, M. Pilloff, and J. Krall. Theory and group velocity of ultrashort, tightly-focused laser pulses. *Journal of the Optical Society of America B*, 12:1695–1703, 1995. and references cited therein.

[18] P. Sprangle, B. Hafizi, J. R. Peñano, R. F. Hubbard, A. Ting, C. I. Moore, D. F. Gordon, A. Zigler, D. Kaganovich, and T. M. Antonsen. Wakefield generation and gev acceleration in tapered plasma channels. *Physical Review E*, 63(5):0056405, 2001.

[19] Richard F. Hubbard, Philip Sprangle, and Bahman Hafizi. Scaling of accelerating gradients and dephasing effects in channel-guided laser wakefield accelerators. *IEEE Transactions on Plasma Science*, 28(4):1159–1169, 2002.

[20] L. S. Pontryagin, R. V. Boltyanskii, R. V. Gamkrelidze, and Mischenko. E. F. *The Mathematical Theory of Optimal Processes*. WIley, New York, 1962.

[21] Donald A. Pierre. *Optimization Theory with Applications*. John Wiley and Sons, Inc., New York, 1969. see chapter 8.

[22] J. P. Verboncoeur, A. B. Langdon, and N. T. Gladd. An object-oriented electromagnetic pic code. *Computer Physics Communications*, 87(1-2):199–211, 1995.

[23] Cs. Tóth, J. Faure, J. van Tilborg, C. G. R. Geddes, C. B. Schroeder, E. Esarey, and W. P. Leemans. Tuning of laser pulse shapes in grating-based compressors for optimal acceleration in plasmas. *Optics Letters*, 28(19):1823–1825, 2003.

[24] C. B. Schroeder, E. Esarey, B. A. Shadwick, and W. P. Leemans. Raman forward scattering of chirped pulses. *Physics of Plasmas*, 10(1):285–295, 2003.

[25] C. B. Schroeder, E. Esarey, C. G. R. Geddes, Cs. Tóth, B. A. Shadwick, J. van Tilborg, and W. P. Leemans. Frequency chirp and pulse shape effects in self-modulated laser accelerators. *Physics of Plasmas*, 10(5):2039–2046, 2003.

[26] D. Umstadter, E. Esarey, and J. Kim. Nonlinear plasma waves resonantly driven by optimized laser pulse trains. *Physical Review Letters*, 72(8):1224–1227, 1994.

[27] G. Bonnaud, D. Teychenne, and J. L. Bobin. Wake-field effect induced by multiple laser pulses. *Physical Review E*, 50:R36–R39, 1994.

[28] S. Dalla and M. Lontano. Large amplitude plasma wave excitation by means of a sequence of short pulses. *Physical Review E*, 49:R1819–R1822, 1994.

[29] D. Umstadter, J. Kim, E. Esarey, E. Todd, and T. Neubert. Resonant laser-driven plasam waves for electron acceleration. *Physical Review E*, 51:3484–3497, 1994.

[30] G. G. Stokes. *Philosophical Magazine*, 32:52, 1849.

[31] K. J. Kim, K. T. McDonald, G. V. Stupakov, and M. S. Zolotorev. Comment on "coherent acceleration by subcycle laser pulses". *Physical Review Letters*, 84(10):3210, 2000.

[32] R. R. Lindberg, A. E. Charman, and J. S. Wurtele. Comparison of the laser wakefield accelerator and the colliding beam accelerator. In Christopher E. Clayton and Patrick Muggli, editors, *Advanced Accelerator Concepts: Tenth Workshop*, volume 647, pages 727–736, Mandalay Beach, California (USA) 22-28 June 2002, 2002. AIP Conference Proceedings.

[33] G. Shvets, J. S. Wurtele, and B. A. Shadwick. Analysis and simulation of raman backscatter in underdense plasmas. *Physics of Plasmas*, 4(5):1872–1880, 1997.

[34] J. Tajima and J. M. Dawson. Laser electron accelerator. *Physical Review Letters*, 43(4):267–270, 1979.

[35] Norman M. Kroll, Amiram Ron, and Norman Rostoker. Optical mixing as a plasma density probe. *Physical Review Letters*, 13(3):83–86, 1964.

[36] Bruce I. Cohen, Allan N. Kaufman, and Kenneth M. Watson. Beat heating of a plasma. *Physical Review Letters*, 29(9):581–584, 1972.

[37] M. N Rosenbluth and C. S. Liu. Excitation of plasma waves by two laser beams. *Physical Review Letters*, 29(11):701–705, 1972.

[38] C. Joshi, W. B. Mori, T. Katsouleas, J. M. Dawson, J. M. Kindel, and D. W. Forslund. Ultrahigh gradient particle acceleration by intense laser-driven plasma density waves. *Nature*, 311(5986):525–529, 1984.

[39] Behrouz Amini and Francis F. Chen. Thomson-scattering detection of plasma waves excited by two laser beams. *Physical Review Letters*, 53(15):1441–1444, 1984.

[40] C. M. Tang, P. Sprangle, and R. N. Sudan. Excitation of the plasma waves in the laser beat wave accelerator. *Applied Physics Letters*, 45(4):375–377, 1984.

[41] C. M. Tang, P. Sprangle, and R. N. Sudan. Dynamics of space-charge waves in the laser beat wave accelerator. *Physics of Fluids*, 28(6):1974–1983, 1985.

[42] D. W. Forslund and J. M. Kindel. Two-dimensional simulations of single-frequency and beat wave laser-plasma heating. *Physical Review Letters*, 54(6):558–561, 1985.

[43] C.E. Clayton, C. Joshi, C. Darrow, and D. Umstadter. Relativistic plasma-wave excitation by colinear optical mixing. *Physical Review Letters*, 54(21):2343–2346, 1985.

[44] Robert J. Noble. Plasma-wave generation in the beat-wave accelerator. *Physical Review A*, 32(1):460–471, 1985.

[45] W. B. Mori. On beat wave excitation of relativistic plasma waves. *IEEE Transactions on Plasma Science*, PS-15(2):88–106, 1987.

[46] P. Mora, D. Pesme, A. Héron, G. Laval, and N. Silvestre. Modulational instability and its consequences for the beat-wave accelerator. *Physical Review Letters*, 61(14):1611–1614, 1988.

[47] A. E. Dangor, A. K. L. Dynoke-Bradshaw, and A. E. Dyson. Observation of relativistic plasma-waves generated by the beat-wave with 1 micron lasers. *Physica Scripta*, T30:107–109, 1990.

[48] Y. Kitigawa, T. Matsumoto, T. Minamihata, K. Sawai, K. Matsuo, K. Mima, K. Nishihara, H. Azechi, K. A. Tanaka, H. Takabe, and S. Nakai. Beat-wave excitation of plasma wave and observation of accelerated electrons. *Physical Review Letters*, 68(1):48–51, 1992.

[49] C. E. Clayton, K. A. Marsh, A. Dyson, M. Everett, A. Lal, W. P. Leemans, R. Williams, and C. Joshi. Ultrahigh-gradient acceleration of injected electrons by laser-excited relativistic electron plasma waves. *Physical Review Letters*, 70(1):37–40, 1993.

[50] M. Everett, A. Lal, D Gordon, C. E. Clayton, K. A. Marsh, and C. Joshi. Trapped electron acceleration by laser-driven relativistic plasma wave. *Nature*, 368(6471):527–529, 1994.

[51] C. E. Clayton, C. Joshi, K. A. Marsh, C. Pellegrini, and J. Rosenzwieg. Second generation beatwave experiments at ucla. *Nuclear Instruments and Methods in Physics Research Section A*, 410(3):378–387, 1998.

[52] Eric Esarey, Phillip Sprangle, Jonathan Krall, and Antonio Ting. Overview of plasma-based accelerator concepts. *IEEE Transactions on Plasma Science*, 24(2):252–288, 1996.

[53] A. I. Akhiezer and R. V. Polovin. Theory of wave motion of an electron plasma. *Soviet Physics - JETP*, 3(5):696–705, 1956.

[54] J. M. Dawson. Nonlinear electron oscillations in a cold plasma. *Physical Review*, 113(2):383–387, 1959.

[55] T. Katsouleas and W. B. Mori. Wave-breaking amplitude of relativistic oscillations in a thermal plasma. *Physical Review Letters*, 61(1):90–94, 1988.

[56] J. B. Rosenzweig. Trapping, thermal effects, and wave-breaking in the nonlinear plasma wake-field accelerator. *Physical Review A*, 38(7):3634–3642, 1988.

[57] W. B. Mori and T. Katsouleas. Wave-breaking of longitudinal plasma oscillations. *Physica Scripta*, T30:127–133, 1990.

[58] C. J. McKinstrie and D. W. Forslund. The detuning of relativistic plasma waves. *Physics of Fluids*, 30(3):904–908, 1987.

[59] J. P. Matte, F. Martin, N. A. Ebrahim, P. Brodeur, and H. Pepin. Enhanced beat wave saturation amplitude in an ionizing plasma. *IEEE Transactions on Plasma Science*, PS-15(2):173–178, 1987.

[60] M. Deutsch, B. Meerson, and J. E. Golub. Strong plasma wave excitation by a chirped laser beat wave. *Physics of Fluids B: Plasma Physics*, 3(7):1773–1780, 1991.

[61] C. Darrow, W.B. Mori, T. Katsouleas, C. Joshi, and D. Umstadter. Electrostatic mode coupling of beat-excited electron plasma waves. *IEEE Transactions on Plasma Science*, PS-15(15):107–130, 1987.

[62] A. Loeb and L. Friedland. Autoresonance laser accelerator. *Physical Review E*, 33(3):1828–1835, 1986.

[63] B. Meerson and L. Friedland. Strong autoresonance excitation of rydberg atoms: The rydberg accelerator. *Physical Review A*, 41(9):5233–5236, 1990.

[64] L. Friedland. Spatial autoresonance: Enhancement of mode conversion due to nonlinear phase locking. *Physics of Fluids B: Plasma Physics*, 4(10):3199–3209, 1992.

[65] L. Friedland. Autoresonant excitation and evolution of nonlinear waves: The variational approach. *Physical Review E*, 55(2):1929–1939, 1997.

[66] L. Friedland. Autoresonant solutions of the nonlinear Schrödinger equation. *Physical Review E*, 58(3):3865–3875, 1998.

[67] R. R. Lindberg, A. E. Charman, J. S. Wurtele, and L. Friedland. Robust autoresonant excitation in the plasma beat-wave accelerator. *Physical Review Letters*, 93:055001, 2004.

[68] P. Sprangle, E. Esarey, and A. Ting. Nonlinear theory of intense laser-plasma interactions. *Physical Review Letters*, 64(17):2011–2014, 1990.

[69] D. W. Forslund, J. M. Kindel, and E. L. Lindman. Nonlinear behavior of stimulated brillouin and raman scattering in laser-irradiated plasmas. *Physical Review Letters*, 30(16):739–743, 1973.

[70] Claire Ellen Max, Jonathan Arons, and A. Bruce Langdon. Self-modulation and self-focusing of electromagnetic waves in plasmas. *Physical Review Letters*, 33(4):209–212, 1974.

[71] P. Sprangle, E. Esarey, and A. Ting. Nonlinear theory of intense laser-plasma interactions. *Physical Review Letters*, 64(17):2011–2014, 23 April 1991.

[72] I. S. Gradshteyn and I. M. Ryzhik, editors. *Table of Integrals, Series, and Products*. Academic Press, New York, 1980.

[73] B. V. Chirikov. A universal instability of many-dimensional oscillator systems. *Physics Reports*, 52(5):263–379, 1979.

[74] Frank M. Lewis. Vibration during acceleration through a critical speed. *Transactions ASME*, APM-54(24):253–261, 1932.

[75] E. Grosfeld and L. Friedland. Spatial control of a classical electron state in a rydberg atom by adiabatic synchronization. *Physical Review E*, 65(4):046230/1–8, 2002.

[76] J. Fajans, E. Gilson, and L. Friedland. Autoresonant (nonstationary) excitation of a collective nonlinear mode. *Physics of Plasmas*, 6(12):4497–4503, 1999.

[77] Donna Strickland and Gerard Mourou. Compression of amplified chirped optical pulses. *Optics Communications*, 56(3):219–221, 1988.

[78] P. Maine, D. Strickland, P. Bado, M. Pessot, and G. Mourou. Generation of ultrahigh peak power pulses by chirped-pulse amplification. *IEEE Journal of Quantum Electronics*, QE-24(2):398–403, 1988.

[79] C. V. Filip, S. Ya. Tochitsky, R. Narang, C. E. Clayton, K. A. Marsh, and C. Joshi. Interpretation of resonant and non-resonant beat-wave excitation: experiments and simulations. In C.E. Clayton and P. Muggli, editors, *Advanced Accelerator Concepts (2002)*, volume 647 of *AIP Conference Proceedings*, pages 770–785. AIP, New York, 2002.

[80] C. V. Filip, R. Narang, S. Ya. Tochitsky, C. E. Clayton, P. Musumeci, R. B. Yoder, K. A. Marsh, J. B. Rosenzweig, C. Pellegrini, and C. Joshi. Non-resonant beat-wave excitation of relativistic plasma waves with constant phase-velocity for charged particle-particle acceleration. in preparation, UCLA, November 2003. Manuscript in preparation.

[81] G. Shvets, N.J. Fisch, A. Pukhov, and J. Meyer-ter Vehn. Generation of periodic accelerating structures in plasma by colliding pulses. *Physical Review E*, 60(2):2218–2223, 1999.

[82] V. M. Malkin, G. Shvets, and N. J. Fisch. Detuned raman amplification of short laser pulses in plasma. *Physical Review Letters*, 84(6):1208–1211, 1999.

[83] G. Shvets and N. J. Fisch. Parametric excitations of fast plasma waves by counter-propagating laser beams. *Physical Review Letters*, 86(15):3328–3331, 2001.

[84] C. E. Clayton and L. Serafini. Generation and transport of ultrashort phase-locked electron bunches to a plasma beatwave accelerator. *IEEE Transactions on Plasma Science*, 24(2):400–408, 1996.

[85] E. Esarey, R. F Hubbard, W. P. Leemans, A. Ting, and P. Sprangle. Electron injection into plasma wakefields by colliding laser pulses. *Physical Review Letters*, 79(14):2682–2685, 1997.

[86] C. B. Schroeder, P. B. Lee, J. S. Wurtele, E. Esarey, and W. P. Leemans. Generation of ultrashort electron bunches by colliding laser pulses. *Physical Review E*, 59(5):6037–6047, 1999.

[87] Edmond B. Treacy. Optical pulse compression with diffraction gratings. *IEEE Journal of Quantum Electronics*, QE-5(9):454–458, 1969.

[88] Robert F. Collin. *Field Theory of Guided Waves*. International Series in Pure and Applied Physics. McGraw-Hill Book Company, New York, 1960.

[89] Roger F. Harrington. *Time-Harmonic Electromagnetic Fields*. McGraw-Hill Electrical and Electronic Enginering Series. McGraw-Hill Book Company, New York, 1961.

[90] S. G. Mikhlin. *Variational Methods in Mathematical Physics*. Maxmillan Company, New York, 1964.

[91] L. Cairo and T. Kahan. *Variational Techniques in Electromagnetism*. Blackie, London, 1965.

[92] J. D. Jackson. *Classical Electrodynamics*. John Wiley and Sons, New York, second edition, 1975.

[93] Jin Au Kong. *Electromagnetic Wave Theory*. John Wiley and Sons, New York, 1986.

[94] Julian L. Davis. *Wave Propgation in Electromagnetic Media*. Springer-Verlag, New York, 1990.

[95] Johnson J. H. Wang. *Generalized Moment Methods in Electromagnetics: Formulation and Computer Solution of Integral Equations*. John Wiley and Sons, New York, 1991.

[96] Wen Zun Zhang. *Engineering Electromagnetism: Functional Methods*. Ellis Horwood, New York, 1991. Chang, Wen-Hsun.

[97] Istvan Vago and Miklos Gyimesi. *Electromagnetic Fields*. Akademiai Kiado, Budapest, 1998.

[98] George W. Hanson and Alexander B. Yakovlev. *Operator Theory for Electromagnetics: an Introduction*. Springer, New York, 2002.

[99] I. B. Bernstein, E. A. Frieman, M. D. Kruskal, and Kulsrud. R. M. An energy principle for hydromagnetic stability problems. *Proceedings of the Royal Society of London A*, 244:17, 1958.

[100] M. D. Kruskal and C. R. Oberman. On the stability of a plasma in static equilibrium. *Physics of Fluids*, 1:275, 1958.

[101] E. A. Frieman and M. Rotenberg. On hydrodynamic stability of stationary equilibria. *Reviews of Modern Physics*, 32:898, 1960.

[102] W. A. Newcomb. Lagrangian and Hamiltonian methods in magnetohydrodynamics. *Nuclear Fusion:*, Suplement(part 2):451, 1962.

[103] C. S. Gardner. Bound on the energy available from a plasma. *Physics of Fluids*, 6:839, 1963.

[104] V. I. Arnold. Variational principle for three dimensional steady-state flows of an ideal fluid. *Journal of Applied Mathematical Mechanics*, 29:1002, 1965.

[105] D. D. Holm, J. E. Marsden, T. Ratiu, and A. Weinstein. Nonlinear stability of fluid and plasma equilibria. *Physics Reports*, 123:1, 1985.

[106] J. M. Finn and G.-Z. Sun. Nonlinear stability and the energy-Casimir method. *Comments Plasma Phys. Controlled Fusion*, 11:7, 1987.

[107] P. J. Morrison and D. Pfirsch. Free-energy expressions for Vlasov equilibria. *Physical Review A*, 40(7):3898–3910, 1989.

[108] R. Evans. Density functionals in the theory of nonuniform fluids. In D. Henderson, editor, *Fundamentals of Inhomogeneous Fluids*. Marcel Dekker, New York, 1992.

[109] Setsuo Ichimaru. *Statistical Plasma Physics*. Westview Press, Bolder, Colorado, revised edition, 2004.

[110] Roger F. Harrington. *Field Computation by Moment Methods*. MacMillan, New York, 1968.

[111] R. Glowinski. *Numerical Methods for Nonlinear Variational Problems*. Springer-Verlag, New York, 1983.

[112] A. R. Mitchell and R. Wait. *The Finite Element Method in Partial Differential Equations*. Wiley, Chichester, 1984.

[113] Phillippe Blanchard and Erwin Bruning. *Variational Methods in Mathematical Physics: a Unified Approach*. Texts and Monographs in Physics. Springer-Verlag, Berlin, 1992.

[114] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, Berlin, 1994.

[115] Charles W. Steele. *Numerical Computation of Electric and Magnetic Fields*. Chapman and Hall, New York, second edition, 1997.

[116] Gregg Penn. FEL performance with harmonic generation. unpublished report, LBNL-CBP, 20 January 2004.

[117] G. Penn, M. Reinsch, and J. S. Wurtele. Analytic model of harmonic generation in the low-gain FEL regime. Technical Report LBNL-56177, LBNL, August 2004.

[118] G. Penn, M. Reinsch, and J. S. Wurtele. Analytic model of harmonic generation in the low-gain FEL regime. In *26th International Free Electron Laser Conference*, 2004.

[119] G Penn, M. Reinsch, J. S. Wurtele, J. N. Corlett, W. M. Fawley, A Zholents, and W. Wan. Harmonic cascade FEL designs for LUX. unpublished report LBNL-56329, LBNL-CBP, July 2004.

[120] G. Penn, M. Reinsch, and J. S. Wurtele. Analytic model of bunched beams for harmonic generation in the low-gain free electron laser regime. *Physical Review Special Topics - Accelerators and Beams*, 9:060702, 2006.

[121] S. J. van Enk and G Nienhuis. Eigenfunction description of laser beams and orbital angular momentum of light. In L. Allen, Stephen M. Barnett, and Miles J. Padgett, editors, *Optical Angular Momentum*, pages 36–47. Institute of Physics Publishing, Bristol, 2003.

[122] G. Nienhuis and L. Allen. Paraxial wave optics and harmonic oscillators. In L. Allen, Stephen M. Barnett, and Miles J. Padgett, editors, *Optical Angular Momentum*, pages 48–57. Institute of Physics Publishing, Bristol, 2003.

[123] L. Allen, Stephen M. Barnett, and Miles J. Padgett, editors. *Optical Angular Momentum*. Institute of Physics Publishing, Bristol, 2003.

[124] Nicholas George and Avshalom Gamliel. Correlation theory of electromagnetic radiation using multipole expansions. In H.N Kritikos and D.L. Jaggard, editors, *Recent Advances in Electromagentic Theory*, pages 144–182. Springer-Verlag, New York, 1990.

[125] James Morehead. Vector spherical harmonics, 2001.

[126] C. T. Tai. *Dyadic Green Functions in Electromagnetic Theory*. IEEE Press, New York, second edition, 1993.

[127] Howard L Weinart, editor. *Reproducing Kernel Hilbert Spaces: Applications in Statistical Signal Processing*, volume 25 of *Benchmark Papers in Electrical Engineering and Computer Science*. Hutchinson Ross Publishing Company, Stroudburg, Pennsylvania, 1982.

[128] R. A. Waldron. *Theory of Guided Electromagentic Waves*. van Nostrand-Reinhold, London, 1970.

[129] J. van Bladel. *Singular Electromagentic Fields and Sources*. Oxford Engineering Science Series. Oxford University Press, Oxford, 1991. 28 in series.

[130] Yahya Rahmat-Samii. On the question of computation of the dyadic Green function at the source region in waveguides and cavities. *IEEE Transactions on Microwave Theory and Techniques*, 23(9):762–765, 1975.

[131] A. Einstein. Zur quantentheorie der strahlung. *Phys. Zeit.*, 18:121, 1917.

[132] G. Bekefi. *Radiation Processes in Plasmas*. Wiley Series in Plasma Physics. John Wiley and Sons, New York, 1966.

[133] V. N. Litvinenko and N. A. Vinokurov. On the classical analog of the Einstein relations between spontaneous emission, induced emission, and absorption. *Nuclear Instruments and Methods in Physics Research A*, 331:440–449, 1993.

[134] A. Friedman, A. Gover, G. Kurizki, S. Ruschin, and A. Yariv. Spontaneous and stimulated emission from quasifree electrons. *Reviews of Modern Physics*, 60(2):471–535, 1988.

[135] S. Krinsky, J. M. Wang, and P. Luchini. Madey's gain-spread theorem for the free electron laser and the theory of stochatic processes. *Journal of Applied Physics*, 53(8):5453–5458, 1982.

[136] Dmitri E. Nikonov, Yuri V. Rostovtsev, and Georg Sussman. Madey's and Liouville's theorems relating to free-electron lasers without inversion. *Physical Review E*, 57(3):3444–3454, 1998.

[137] J. M. J Madey. Relationship between mean radiated energy, mean squared radiated energy, and spontaneous power spectrum in a power series expansion of the equations of motion in a free electron laser. *Nuovo Cimento della Societa Italiana di Fisica B*, 50(ser. 2, no. 1):64–88, 1979.

[138] Kwang-Je Kim. FEL gain taking into account diffraction and beam emittance; generalized Madey's theorem. *Nuclear Instruments and Methods in Physics Research A*, 318:489–494, 1992.

[139] Ming Xie and David A. G. Deacon. Theoretical study of FEL active guiding in the small signal regime. *Nuclear Instruments and Methods in Physics Research A*, 250(1-2):426–431, 1986.

[140] V. H. Rumsey. Reaction concept in electromagnetic theory. *Physical Review*, 94(6):1483–1491, 1954.

[141] A. J. Firth. Propagation of LASER beams through inhomogenous media. *Optics Commuications*, 22(2):226–230, 1977.

[142] A. Anderson and M. Bonnedal. Variational approach to nonlinear focusing of Gaussian laser beams. *Physics of Fluids*, 22(1):105–109, 1979.

[143] Brian J. Duda and Warren B. Mori. Variational principle approach to short-pulse laser-plasma interactions in three dimensions. *Physical Review E*, 61(2):1925–1939, 2000.

[144] Avner Amir and Yuval Greenzweig. Three-dimensional theory of the free-electron laser I: Gain and evolution of optical modes. *Physical Review A*, 34(6):4809–4819, 1986.

[145] P. Luchini and S. Solimeno. Variational solution of the wave equation for a high gain FEL and a finite wiggling radius. *Nuclear Instruments and Methods in Physics Research A*, 272(1-2):311–317, 1988.

[146] Avner Amir. Optical-mode trapping in the free-electron laser. *Physical Review A*, 37(3):780786, 1988.

[147] L. H. Yu, S. Krinsky, and R. L. Gluckstern. Calculation of universal scaling function for free-electron laser gain. *Physical Review Letters*, 64(25):3011–3013, 1990.

[148] B. Hafizi and C. W. Roberson. Effect of emittance and energy spread on a free-electron laser in the gain-focusing regime. *Physical Review Letters*, 68(24):3539–3542, 1992.

[149] Ming Xie. Exact and variational solutions of 3D eigenmodes in high gain FELs. *Nuclear Instruments and Methods in Physics Research A*, 445(1-3):59–66, 2000.

[150] Allan N. Kaufman and Darryl D. Holm. The Lie-transformed Vlasov action principle: Relativistically covariant wave propagation and self-consistent pondermotive effects. *Physics Letters*, 105A(6):277–279, 1984.

[151] D. Pfirsch and P. J. Morrison. Local conservation laws for the Maxwell-Vlasov and collisionless kinetic guiding center theories. *Physical Review A*, 32(3):1714–1721, 1985.

[152] Lukasz A. Turski and Allan N. Kaufman. Canonical-dissipative formulation of relativistic plasma kinetic theory with self-consistent Maxwell field. *Physics Letters A*, 120(7):331–333, 1987.

[153] Allan N. Kaufman, Huanchan Ye, and Yukkei Hui. Variational formulation of covariant eikonal theory for vector waves. *Physics Lettera A*, 120(7):327–330, 1987.

[154] Allain J. Brizard. A new Lagrangian formulation for laser-plasma interactions. *Physics of Plasmas*, 5(4):1110–1117, 1998.

[155] P. Similon and J. S. Wurtele. Nonlinear interaction of a relativistic electron beam with electromagnetic waves. *Physics Letters A*, 154(4,5):224–232, 1991.

[156] M. E. Delanay. *On the averaged Lagrangian technique for nonlinear dispersive waves.* Ph.d., California Institute of Technology, 1971.

[157] G. B. Whitham. *Linear and Nonlinear Waves.* Pure and Applied Mathematics. John Wiley and Sons, Inc., New York, 1974.

[158] P. Luchini and S. Solimeno. Effects of bending on a free electron laser performance. *Nuclear Instruments and Methods in Physics Research A*, 272(1-2):334–339, 1988.

[159] L. Debnath and P. Mikusinski. *Introduction to Hilbert Spaces with Applications.* Academic Press, San Diego, 1990.

[160] E. Zeidler. *Applied Functional Analysis: Applications to Mathematical Physics.* Springer-Verlag, New York, 1995.

[161] S. van der Meer. Stochastic cooling and the accumulation of antiprotons. *Reviews of Modern Physics*, 57(3):689–697, 1985.

[162] D. Mohl. Stochastic cooling for beginners. Technical report, CERN School, 1977.

[163] John Marriner. Stochastic cooling overview. *Nuclear Instruments and Methods in Physics Research A*, 532(1):11–18, 2004.

[164] A. A. Mikhailichenko and M. S. Zolotorev. Optical stochastic cooling. *Physical Review Letters*, 71(25):4146–4149, 1993.

[165] M. S. Zolotorev and A. A. Zholentz. Transit-time method of optical stochastic cooling. *Physical Review E*, 50(4):3087–3091, 1994.

[166] A. Zholents, M. S. Zolotorev, and W. Wan. Optical stochastic cooling of muons. *Physical Review STAB*, 4(3), 2001.

[167] Kwang-Je Kim. Characteristics of synchrotron radiation. In Melvin Month and Margaret Dienes, editors, *Physics of Particle Accelerators*, volume 184-1 of *AIP Conference Proceedings*, pages 565–632, Summer Schools 1987-1988, 1989. American Institute of Physics.

[168] M. Sands. The physics of electron-positron storage rings. Technical Report SLAC-121, Stanford Linear Accelerator Center, 1970.

[169] E. Esarey. Laser cooling of electron beams via Thomson scattering. *Nuclear Instruments and Methods in Physics Research A*, 445(1):7–14, 2000.

[170] Amnon Yariv. *Quantum Electronics.* Wiley, New York, 3rd edition, 1989.

[171] N. F. Mott. *Proceedings Royal Society A*, 124:425–442, 1929.

[172] Stephen Benson and John M. J. Madey. Shot and quantum noise in free electron lasers. *Nuclear Instruments and Methods in Physics Research A*, 237(1-2):55–60, 1985.

[173] Charles A. Brau. *Free-Electron Lasers.* Academic Press, New York, 1990.

[174] W. H. Zurek. Pointer basis of quantum apparatus: into what mixture does the wave packet collapse? *Physical Review D*, 24(6):1516–1525, 1981.

[175] W. H. Zurek. Environment-induced superselection rules. *Physical Review D*, 26(8):1862–1880, 1982.

[176] W. H. Zurek. Preferred states, predictability, classicality, and the environment-induced decoherence. *Progress of Theoretical Physics*, 89(2):281–312, 1993.

[177] Wulf B. Kunkel, editor. *Plasma Physics in Theory and Application*. McGraw-Hill, New York, 1966.

[178] Roy J. Glauber and M. Lewenstein. Quantum optics of dielectric media. *Physical Review A*, 43(1):467–491, 1991.

[179] Bruno Huttner and Stephen M. Barnett. Quantization of the electromagnetic field in dielectrics. *Physical Review A*, 46(7):4306–4322, 1992.

[180] P. W. Milonni. Field quantization and radiative processes in dispersive dielectric media. *Journal of Modern Optics*, 42(10):1991–2004, 1995.

[181] T. Gruner and D.-G. Welsch. Green-function approach to the radiation-field quantization for homogeneous and inhomogeneous Kramers-Kronig dielectrics. *Physical Review A*, 53(3):1818–1829, 1996.

[182] Gediminas Juzeliunas. Microscopic theory of quantization of radiation in molecular dielectrics: Normal mode representation of operators for local and averaged (macroscopic) fields. *Physical Review A*, 53(5):3543–3557, 1996.

[183] B. J. Dalton, E. S. Guerra, and E. L. Knight. Field quantization in dielectric media and the generalized multipolar Hamiltonian. *Physical Review A*, 54(3):2292–2313, 1996.

[184] B. J. Dalton and M. Babiker. Macroscopic quantization in quantum optics and cavity quantum electrodynamics: Interatomic interactions. *Physical Review A*, 56(1):905–911, 1997.

[185] Ho Trung Dung, Ludwig Knoll, and Dirk-Gunnar Welsch. Three-dimensional quantization of the electromagnetic field in dispersive and absorbing inhomogeneous dielectrics. *Physical Review A*, 57(5):3931–3942, 1998.

[186] A. Tip. Linear absorptive dielectrics. *Physical Review A*, 57(6):4818–4841, 1998.

[187] P. D. Drummond and M. Hillery. Quantum theory of dispersive electromagnetic modes. *Physical Review A*, 59(1), 1999.

[188] Michael E. Crenshaw and Charles M. Bowden. Effects of local fields on spontaneous emission in dielectric media. *Physical Review Letters*, 85(9):1851–1854, 2000.

[189] Michael E. Crenshaw. Microscopic foundation of macroscopic quantum optics. *Physical Review A*, 67:033805, 2003.

[190] J. C. Garrison and R. Y. Chiao. Canonical and kinetic forms of the electromagnetic momentum in an ad hoc quantization scheme for dispersive dielctric. *Physical Review A*, 70(5):053826, 2004.

[191] Martijn Wubs, L. G. Suttorp, and A. Lagendijk. Mutliple-scattering approach to interatomic interactions and superradiance in inhomegeneous dielectrics. *Physical Review A*, 70:053823, 2004.

[192] L. G. Suttorp and A. J. van Wonderen. Fano diagonalization of a polariton model for an inhomogeneous absorptive dielectric. *Europhysics Letters*, 67(5):766–772, 2004.

[193] L. G. Suttorp and Martijn Wubs. Field quantization in inhomogeneous absorptive dielectrics. *Physical Review A*, 70:013816, 2004.

[194] Claude Cohen-Tannoudji, Jacques Dupont-Roc, and Gilbert Grynberg. *Photons and Atoms: Introduction to Quantum Electrodynamics*. Wiley Interscience, New York, 1989.

[195] C. M. Caves. Quantum limis on noise in linear amplifiers. *Physical Review D*, 26(8):1817–1839, 1982.

[196] Roy J. Glauber. Photon correlations. *Physical Review Letters*, 10(3):84–86, 1963.

[197] Roy J. Glauber. The quantum theory of optical coherence. *Physical Review*, 130(6):2529–2539, 1963.

[198] Roy J. Glauber. Coherent and incoherent states of the radiation field. *Physical Review*, 131(6):2766–2788, 1963.

[199] U. M. Titulaer and Roy J. Glauber. Correlation functions for coherent fields. *Physical Review*, 140(3B):B676–B682, 1965.

[200] Stephen M. Barnett and Paul M. Radmore. *Methods in Theoretical Quantum Optics*. Clarendon Press, Oxford, 1997.

[201] R. J. Glauber. Classical behavior of systems of quantum oscillators. *Physical Letters*, 21(6):650–652, 1966.

[202] C. L. Mehta and E. C. G. Sudarshan. Time evolution of coherent states. *Physics Letters*, 22(5):574–576, 1966.

[203] C. L. Mehta, P. Chand, E. C. G. Sudarshan, and R. Vedam. Dynamics of coherent states. *Physical Review*, 157(5):1198–1206, 1967.

[204] U. M. Titulaer and R. J. Glauber. Density operators for coherent fields. *Physical Review*, 145(4):1041–1050, 1966.

[205] K. E. Cahill and R. J. Glauber. Density operators and quasiprobability distributions. *Physical Review*, 177(5):1882–1902, 1969.

[206] E. C. G. Sudarshan. Equivalence of semiclassical and quantum mechanical descriptions of statistical light beams. *Physical Review Letters*, 10(7):277–279, 1963.

[207] C. L. Mehta and E. C. G. Sudarshan. Relation between quantum and semiclassical description of optical coherence. *Physical Review*, 138(1B):B274–B280, 1965.

[208] K. E. Cahill and R. J. Galuber. Ordered expansion in boson amplitude operators. *Physical Review*, 177(5):1857–1881, 1969.

[209] K Shimoda, H. Takahasi, and C. H. Townes. Fluctuations in amplification of quanta with applications to maser amplifiers. *Journal of the Physical Society of Japan*, 12(6):686–700, 1957.

[210] A. L. Schawlow and C. H. Townes. Infrared and optical masers. *Physical Review*, 112(6):1940–1949, 1958.

[211] W. H. Louisell, A. Yariv, and A. E. Siegman. Quantum fluctuations and noise in parametric processes I. *Physical Review*, 124(6):1646–1654, 1961.

[212] H. A. Haus and J. A. Mullen. Quantum noise in linear amplifiers. *Physical Review*, 128(5):2407–2413, 1962.

[213] J. P. Gordan, W. H. Louisell, and L. R. Walker. Quantum fluctuations in parametric processes II. *Physical Review*, 129(1):481–485, 1963.

[214] J. P. Gordan, L. R. Walker, and W. H. Louisell. Quantum statistics of masers and attenuators. *Physical Review*, 130(2):806–812, 1963.

[215] H. Kogelnik and A. Yariv. Considerations of noise and schemes for its reduction in laser amplifiers. *Proceedings of the IEEE*, 52(2):165–172, 1964.

[216] Melvin Lax. Quantum noise IV: Quantum theory of noise sources. *Physical Review*, 145(1):110–129, 1966.

[217] B. R. Mollow and R. J. Glauber. Quantum theory of parametric amplification I. *Physical Review*, 160(5):1076–1096, 1967.

[218] B. R. Mollow and R. J. Glauber. Quantum theory of parametric amplification II. *Physical Review*, 160(5):1097–1108, 1967.

[219] Y. Yamamoto and H. A. Haus. Preparation, measurement, and information capacity of optical quantum states. *Reviews of Modern Physics*, 58(4):1001–1020, 1986.

[220] Horace P. Yuen. Quantum amplifiers, quantum duplicators, and quantum cryptography. *Quantum and Semiclassical Optics*, 8:939–949, 1996.

[221] L. Friedland. Correspondence principle in free-electron lasers. *Physical Review A*, 29(3):1310–1314, 1984.

[222] Juan Pablo Paz, Salman Habib, and Wojciech H. Zurek. Reduction of the wave packet: Preferred observable and decoherence time-scale. *Physical Review D*, 47(2):488–501, 1993.

[223] O. Kübler and H. D. Zeh. Dynamics of quantum correlations. *Annals of Physics*, 76(2):405–418, 1973.

[224] Wojciech H. Zurek, Salman Habib, and Juan Pablo Paz. Coherent states via decoherence. *Physical Review Letters*, 70(9):1187–1190, 1993.

[225] U. Gavish, B. Yurke, and Y. Imry. Generalized constraints on quantum amplification. *Physical Review Letters*, 93:250601, 2004.

[226] S. Heifets and M. S. Zolotorev. Quantum theory of optical stochastic cooling. *Physical Review E*, 65:016507, 2001.

[227] S. E. Harris, J. E. Field, and A. Imamoglu. Nonlinear optical processes using electromagnetically induced transparency. *Physical Review Letters*, 64(10):1107–1110, 1990.

[228] K.-J. Boller, A. Imamoglu, and S. E. Harris. Observation of electromagnetically induced transparency. *Physical Review Letters*, 66(20):2593–2595, 1991.

[229] S. E. Harris, J. E. Field, and A. Kasapi. Dispersive properties of electromagnetically induced transparency. *Physical Review A*, 46(1):R29–R32, 1992.

[230] S. E. Harris. Normal modes for electromagnetically induced transparency. *Physical Review Letters*, 72(2):52–55, 1994.

[231] S. E. Harris. Refractive-index control with strong fields. *Optics Letters*, 19(23):2018–2020, 1994.

[232] Min Xiao, Yong-qing Li, Shao-zheng Ji, and Julio Gea-Banacloche. Measurement of dispersive properties of electromagnetically induced transparency in rubidium atoms. *Physical Review Letters*, 74(5):666–669, 1995.

[233] A. Kasapi, Maneesh Jain, G. Y. Yin, and S. E. Harris. Electromagnetically induced transparency: Propagation dynamics. *Physical Review Letters*, 74(13):2447–2450, 1995.

[234] Julio Gea-Banacloche, Yong-qing Li, Shao-zheng Ji, and Min Xiao. Electromagnetically induced transparency in ladder-type inhomogeneously broadened media: theory and experiment. *Physical Review A*, 51(1):576–584, 1995.

[235] Michael Fleischhauer and Aaron S. Manka. Propagation of laser pulses and coherent population transfer in dissipative three-level systems: an adiabatic dressed-state picture. *Physical Review A*, 54(1):794–803, 1996.

[236] S. E. Harris and Lene Vestergaard Hau. Nonlinear optics at low light levels. *Physical Review Letters*, 82(23):4611–4614, 1999.

[237] Lene Vestergaard Hau, S. E. Harris, Zacchary Dutton, and Cyrus H. Behroozi. Light speed reduction to 17 metres per second in an ultracold atomic gas. *Nature*, 397(6720):594–598, 1999.

[238] M. Fleischhauer. Electromagnetically-induced transparency and coherent-state preparation in optically thick media. *Optics Express*, 4(2):107–112, 1999.

[239] M. D. Lukin, S. F. Yelin, and Fleischhauer. Entanglement of atomic ensembles by trapping correlated photon states. *Physical Review Letters*, 84(18):4232–4235, 2000.

[240] M Fleischhauer, S. F. Yelin, and M. D. Lukin. How to trap photons? Storing single-photon quantum states in collective atomic excitations. *Optics Communications*, 179:395–410, 2000.

[241] M. Fleischhauer and M. D. Lukin. Quantum memory for photons: Dark-state polaritons. *Physical Review A*, 65(2):022414, 2002.

[242] Matthew D. Eisaman. *Generation, Storage, and Retrival of Nonclassical States of Light using Atomic Ensembles*. PhD thesis, Harvard University, 2006.

[243] Marlan O. Scully. From lasers and masers to phaseonium and phasers. *Physics Reports*, 219(3-6):191–201, 1992.

[244] Sadaf Sultana and M. Suhail Zubairy. Effect of finite bandwidth on refractive-index enhancement and lasing without inversion. *Physical Review A*, 49(1):438–448, 1994.

[245] Michael Fleischhauer and Marlan O. Scully. Quantum sensitivity limits of an optical magnetometer based on atomic phase coherence. *Physical Review A*, 49(3):1973–1986, 1994.

[246] Jennifer R. Csesznegi and Rainer Grobe. Recall and creation of spatial excitation distributions in dielectric media. *Physical Review Letters*, 79(17):3162–3165, 1997.

[247] K. Bergmann, H. Theuer, and B. W. Shore. Coherent population transfer among quantum states of atoms and molecules. *Reviews of Modern Physics*, 70(3):1003–1025, 1998.

[248] A. André, L.-M. Duan, and M. D. Lukin. Coherent atom interactions mediated by dark-state polaritons. *Physical Review Letters*, 88(24):243602, 2002.

[249] F. Zimmer and M. Fleischhauer. Sagnac interferometer based on ultraslow polaritons in cold atomic vapors. *Physical Review Letters*, 92(25):253201, 2004.

[250] R. K. Bullough and H. M. Gibbs. Information storage and retrieval by stopping pulses of light. *Journal of Modern Optics*, 20(2):255–284, 2004.

[251] Stephen E. Harris. Electromagnetically induced transparency. *Physics Today*, 50(7):36–42, 1997.

[252] J. P. Marangos. Topical review: Electromagnetically induced transparency. *Journal of Modern Optics*, 45(3):471–503, 1998.

[253] M. D. Lukin. Colloquium: Trapping and manipulating photon states in atomic ensembles. *Reviews of Modern Physics*, 75(2):457–472, 2003.

[254] Michael Fleischhauer, Atac Imamoglu, and Jonathan P. Marangos. Electromagnetically induced transparency: Optics in coherent media. *Reviews of Modern Physics*, 77(2):633–673, 2005.

[255] G. Shvets and J. S. Wurtele. Transparency of magnetized plasma at the cyclotron frequency. *Physical Review Letters*, 89(11):115003, 2002.

[256] A. G. Litvak and M. D. Tokman. Electromagnetically induced transparency in ensembles of classical oscillators. *Physical Review Letters*, 88(9):095003, 2002.

[257] M. S. Hur, J. S. Wurtele, and G. Shvets. Simulation of electromagnetically induced transparency in a magnetized plasma. *Physics of Plasmas*, 10(7):3004–3011, 2003.

[258] G. Shvets, M. Tushenstov, M. D. Tokman, and A. Kryachko. Propagation of electromagnetic waves in the plasma near electron cyclotron resonance: Undulator-induced transparency. *Physics of Plasmas*, 12(5):056701, 2005.

[259] Mikhail Tushentsov, Gennady Shvets, Andrey Yu. Kryachko, and Mikhail D. Tokman. Undulator-induced transparency of magnetized plasma: New aproach to electromagnetic energy compression. *IEEE Transactions on Plasma Science*, 33(1):23–31, 2005.

[260] Zhneg-Feng Hu, Chun-Guang Du, Dai-Jun Li, and Shi-Qun Li. Slow light in ultracold Bose gas. *Chinese Physics Letters*, 19(12):1805–1807, 2002.

[261] Le-Man Kuang and Lan Zhou. Generation of atom-photon entangled states in atomic Bose-Einstein condensate via electromagnetically induced transparency. *Physical Review A*, 68(4):043606, 2003.

[262] A. V Prokhorov, A. P. Alodjants, and S. M. Arekelian. Generation of nonclassical states of light in the Bose-Einstein condensate under electromagnetically induced transparency. *JETP Letters*, 80(12):739–742, 2004.

[263] Guang-Ri Jin, Chul Koo Kim, and Kyun Nahm. Electromagnetically induced transparency in an atom-molecule bose-einstein condensate, 2006.

[264] C. P. Sun, Y. Li, and X. F. Liu. Quasi-spin-wave quantum memories with a dynamical symmetry. *Physical Review Letters*, 91(14):147903, 2003.

[265] Jing-Min Hou, Li-Jun Tian, and Shuo Jin. Dark states and coherent control of spin states in molecular magnets. *Physical Review B*, 73(13):134425, 2006.

[266] Yuri Rostovtsev, Andrey B. Matsko, and Marlon O. Scully. Electromagnetic-induced transparency and amplification of electromagnetic waves in photonic band-gap materials. *Physical Review A*, 57(6):4919–4924, 1998.

[267] Zachary Dutton, K. V. R. M. Murali, William D. Oliver, and T. P. Orlando. Electromagnetically induced transparency in superconducting quantum circuits: effects of decoherence, tunneling,and multilevel crosstalk. *Physical Review B*, 73(10):104516, 2006.

[268] S. E. Harris. Electromagnetically induced transparency in an ideal plasma. *Physical Review Letters*, 77(27):5357–5360, 1996.

[269] Andrey B. Matsko and Yuri Rostovtsev. Electromagnetic-wave propagation and amplification in overdense plasmas: Application to free-electron lasers. *Physical Review E*, 58(6):7846–7854, 1998.

[270] D. F. Gordon, W. B. Mori, and C. Joshi. On the possibility of electromagnetically induced transparency in a plasma. i. infinite plasma. *Physics of Plasmas*, 7(8):3145–3155, 2000.

[271] D. F. Gordon, W. B. Mori, and C. Joshi. On the possibility of electromagnetically induced transparency in a plasma. ii. bounded plasma. *Physics of Plasmas*, 7(8):3156–3166, 2000.

[272] Dwight R. Nicholson. *Introduction to Plasma Theory*. Krieger Publishing Company, Malabar, Florida, 1983.

[273] Lee W. Casperson. Rotating-wave approximation in high-gain lasers. *Physical Review A*, 46(1):401–409, 1992.

[274] S. Guérin, R. G. Unanyan, L. P. Yatsenko, and H. R. Jauslin. Floquet perturbative analysis for stirap beyond the rotating wave approximation. *Optics Express*, 4(2):84–90, 1999.

[275] N. N. Bogoliubov. *Ukrainski MAt. Zh.*, 2:3, 1950.

[276] S. V. Tyablikov. *Oviet Physics JETP*, 21:377, 1951.

[277] N. N. Bogoliubov. *Lectures on Quantum Statistics*. Gordon and Breach Publishers, New York, 1970.

[278] S. Krishnan Mangala and Tucker Jr. Carrington. Quantum canonical transformation of the rotational-vibrational hamiltonian to remove a coriolis term. *Journal of Chemical Physics*, 94(1):461–477, 1991.

[279] C. Emary and R. F. Bishop. Bogoliubov transformations and exact isolated solutions for simple nonadiabatic Hamiltonians. *Journal of Mathematical Physics*, 43(8):3916–3926, 2002.

[280] J. H. Eberly, M. L. Pons, and H. R. Haq. Dressed-field pulses in an absorbing medium. *Physical Review Letters*, 72(1):56–59, 1994.

[281] M. Fleischhauer and M. D. Lukin. Dark-state polaritons in electromagnetically induced transparency. *Physical Review Letters*, 84(22):5094–5097, 2000.

[282] G. Juzeliunas and H. J. Carmichael. Systematic formulation of slow polaritons in atomic gases. *Physical Review A*, 65:021601, 2002.

[283] G. Juzeliunas, M. Masalas, and M. Fleischhauer. Storing and releasing light in a gas of moving atoms. *Physical Review A*, 67(2):023809, 2003.

[284] Fatih Yanik, Mehmut and Shui Fan. Stopping and storing light coherently. *Physical Review A*, 71:013803, 2005.

[285] T. Holstein and H. Primakoff. Field dependence of the intrinsic domain magnetization of a ferromagnet. *Physical Review*, 58(12):1098–1113, 1940.

[286] Michael M. Kash, Vladimir A. Sautenov, Alexander S. Zibrov, L. Hollberg, George R. Welch, Mikhail D. Lukin, Yuri Rostovtsev, Edward S. Fry, and Marlan O. Scully. Ultraslow group velocity and enhanced nonlinear optical effects in a coherently driven hot atomic gas. *Physical Review Letters*, 82(26):5229–5232, 1999.

[287] A. Yu. Kryachko, A. G. Litvak, and M. D. Tokman. Electromagnetically induced transparency in high-temperature magneto-active plasma. *Journal of Experimental and Theoretical Physics*, 95(5):697–704, 2002.

[288] L. D. Landau. *J. Phys. (U.S.S.R.)*, 10:25, 1946.

[289] David Pines and David Bohm. A collective description of electron interactions II: Collective versus individual particle aspects of the interactions. *Physical Review*, 85(2):338–353, 1952.

[290] G. Marcus, L. Friedland, and A. Zigler. From quantum ladder climbing to classical autoresonance. *Physical Review A*, 69:013407, 2004.

[291] S. Gupta, K. W. Murch, and D. M. Stamper-Kurn. Bose-Einstein condensation in a circular waveguide. *Physical Review Letters*, 95:143201, 2005.

[292] K. W. Murch, K. L. Moore, S. Gupta, and D. M. Stamper-Kurn. Dispersion management using betatron resonances in an ultracold-atom storage ring. *Physical Review Letters*, 96:013202, 2005.

[293] John Williamson. On the algebraic problem concerning the normal forms of linear dynamical systems. *American Journal of Mathematics*, 58(1):141–163, 1936.

[294] A. Laub and K. Meyer. Canonical forms for symplectic and Hamiltonian matrices. *Celestial Mechanics*, 9:213–238, 1974.

[295] L. A. Pars. *A Treatise on Analytical Mechanics*. Ox Bow Press, Woodbridge, CT, 1981.

[296] D. Z. Djokovic, J. Patera, P. Winterlitz, and H. Zassenhaus. Normal forms of elements of classical real and complex Lie and Jordan algebras. *Journal of Mathematical Physics*, 24:1363–1374, 1983.

[297] V. I. Arnold. *Mathematical Methods of Classical Physics*, volume 60 of *Graduate Texts in Mathemtics*. Springer, New York, 1989.

[298] P. Huguenin and J.-P. Amiet. *Mécaniques Clasique at Quantique dans l'Espace de Phase*. Université de Neuchâtel, Neuchâtel, 1981.

[299] C Van Loan. A symplectic method for approximating the eigenvalues of a Hamiltonian matrix. *Linear Algebra and its Applications*, 16:233–251, 1984.

[300] A. Bunse-Gerstner. Matrix factorizations for symplectic QR-like methods. *Linear Algebra and its Applications*, 83:49–77, 1986.

[301] A. Bunse-Gerstner, R. Byers, and V. Mehrman. A chart of numerical methods fro structured eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 13:419–453, 1992.

[302] Peter Benner and Heike Farssbender. An implicitly re-started symplectic Lanczos method for the symplectic eigenvalue problem. *SIAM Journal on Matrix Analysis and Applications*, 22(3):682–713, 2000.

[303] Erica A. Butcher and S. C. Sinha. On the construction of transformations of linear Hamiltonian systems to real normal forms. *International Journal of Bifurcation and Chaos*, 10(9):2177–2191, 2000.

[304] M. Konstantinov, V. Mehermann, and P. Petkov. Perturbation analysis of Hamiltonian Schur and block-Schur forms. *SIAM Journal on Matrix Analysis and Applications*, 23(2):387–424, 2001.