Alex H. Barnett

Tenure File

November 2010

Alex H. Barnett — Tenure File, November 2010

Contents

Professional statement	7 pages
Curriculum vitae	6 pages
Journal publications (reverse chronological order):	
• Submitted journal articles in review	1 item
• Refereed journal articles authored during current position at Dartmouth	11 items
• Refereed journal articles authored prior to current position at Dartmouth	7 items
Published reviews of journal articles (MathSciNet and Zentralblatt Math)	3 items
Software product website: `MPSpack` front page and download page	2 pages
Cover image produced, and "About the Cover", *Notices of the Americal Mathematical Society*	2 pages
Education / outreach article: "convolution: *son et lumière*"	1 item

Professional statement — Alex H. Barnett, November 2010

As an applied mathematician, I create numerical and analytical tools that enable the solution of challenging problems in diverse areas of science and technology. This requires cross-disciplinary collaboration with scientists as well as other mathematicians, and the pursuit of excellence in developing innovative computational methods. In turn, questions arising in these areas have led me to interesting mathematics. As an educator, I aim to share both the joy that comes when a theorem has been understood deeply, and the satisfaction of using the power of computational mathematics to help reliably solve problems we face in the wider world.

1 Scholarship

My work focuses on solving partial differential equations (PDEs), which are the mathematical objects that describe much of the physical world around us: the flow of heat, the propagation of sound, light and radio signals, and the quantum wavefunctions that lie at the root of all matter. For example, we may represent an electric field throughout a region of space by a function; the PDE is then an equation relating various *rates of change* of this function. An example at the core of my work is the *Helmholtz equation*, which states essentially that if a function is positive its curvature must be negative, and vice versa, leading to oscillations and waves. Solving the PDE then means finding a function which obeys the equation, yet also behaves in prescribed ways at the *boundaries* of geometric objects such as reflectors or other devices with which the waves interact.

It is possible to write down a formula for the exact solution only for a few simple geometric boundary shapes. Thus finding a *numerical approximation* to the solution when the boundary is more complex has been crucial to progress in science and technology. With the invention of electronic computers, the size and complexity of problems we routinely solve has grown exponentially. An interesting consequence is that the choice of a good solution method matters now more than ever before, since it may turn a problem requiring several days of computer time into one requiring mere minutes. I create and apply new algorithms for the efficient and accurate solution of challenging PDE problems—ones that would be prohibitively expensive to solve with current conventional tools. In many cases I also analyze these methods mathematically, so that their understanding is given a more rigorous foundation, and their computational speed and accuracy is guaranteed.

My research achievements and interests are in the following areas:

- Numerical analysis and scientific computing, particularly computational waves at high frequency and/or high accuracy: the Helmholtz equation; eigenvalue, scattering and periodic problems; high-order and spectral methods, corner singularities, boundary integral equations; the heat equation.
- Mathematical physics: numerical studies in quantum chaos, semiclassical (high frequency asymptotic) properties of the spectrum and eigenfunctions of the Laplacian; billiards, and scarring.

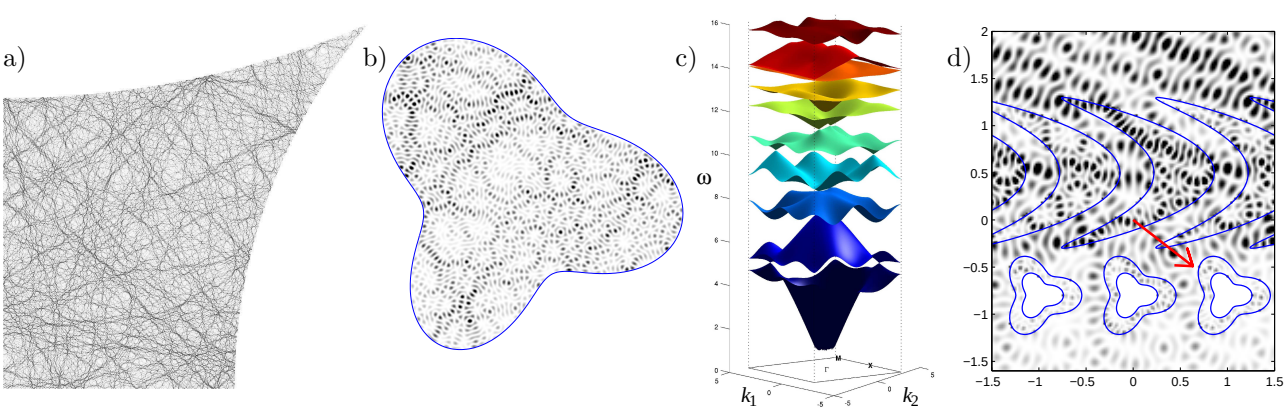


Figure 1: a) 10^5th Dirichlet eigenfunction of a chaotic 2D domain (size 320 wavelengths; see Section 1.2.) b) Eigenfunction of 2D domain, 10 digit accuracy (Section 1.1.2.) c) Band structure of 2D photonic crystal (Section 1.1.1.) d) Scattering from infinite periodic array of dielectric obstacles, 11 digit accuracy (Section 1.1.1.)

- Applied mathematics: mathematical ecology (animal home range modeling), inverse problems in medical imaging (diffuse optical tomography), the mathematics of music.

I have given 46 research talks, all but four of which were by invitation. I have 18 refereed journal publications that have appeared or are in press, and a further preprint in review, all in top-level international journals. I have also authored and co-authored freely-available and documented software, as part of my goal to enable other researchers to take advantage of the tools I have developed. My work has been supported over the last five years by two single-PI 3-year research awards from the Division of Mathematical Sciences of the National Science Foundation, on high-frequency wave eigenvalue and scattering problems, totalling $413,037. My research trajectory reflects my origins in a Ph.D in physics, followed by a postdoc in medical imaging, before finding what I find to be a truly satisfying home in applied mathematics. A significant part of my work remains interdisciplinary, collaborating with biologists, engineers, pure mathematicians, and recently musicologists. Grappling with challenges and modes of thought outside of mathematics is a key part of what excites me.

1.1 Numerical analysis for wave computations

Advances in numerical methods have been of paramount importance for scientific progress and the optimization of physical devices in today's age of computers and communications. Society has been transformed in the last century by technologies that use waves: one need only mention television, cell-phones, radar, ultrasound, and fiber optics. Finite elements and other standard numerical tools have enabled a large variety of devices to be modeled, at least to low accuracy, but there remain many situations which demand high frequency (where the many wavelengths across the system renders it a challenging multiscale problem) and/or high accuracy. These include acoustic and radar scattering problems, resonant devices such as filters, and devices containing periodic arrays or lattices. There is a growing range of modern applications to nano-optics, lasers, optical signal processing, meta-materials, communications, remote sensing, and affordable solar photovoltaics. Efficiency and robustness are paramount because modeling is a bottleneck: predicting real-world device performance often requires thousands of solutions at different parameters.

A key theme of my work has been to develop more efficient representations for waves that allow solutions to the Helmholtz equation to be approximated with high order (often spectral) accuracy. Often these use *particular solutions* (for example, plane waves) as global basis functions, turning the PDE problem into the simpler problem of matching on boundaries via least-squares or other collocation methods. With an efficient basis (often needing only three unknowns per wavelength on boundaries), large-scale problems several hundred wavelengths in size become tractable with, and indeed best handled by, dense linear algebra. For instance, with collaborator Timo Betcke (Reading, UK) I have designed such Fourier-Bessel basis sets to handle the singularities occurring at each corner of a polygonal domain, solving the two-dimensional (2D) scattering problem with proven exponential convergence [15] [1]. By matching on a few artificial boundaries we achieve typical errors of 10^{-10} in only a few seconds of computation, for Dirichlet and Neumann problems that are challenging for both standard discretization and integral equation methods.

Another efficient global basis is a set of fundamental solutions ('point sources') lying on a curve outside the physical domain. This idea is popular with engineers, yet had not been analyzed for interior Helmholtz boundary value problems before our work [13]. We provided a complete convergence analysis in the disc, and demonstrated spectral accuracy in a variety of analytic planar domains. Crucially, it is not the condition number (which is very high), but rather the norm of the coefficient vector that limits the accuracy in finite-precision arithmetic. Keeping this norm small in turn requires understanding the singularities in the analytic continuation of the solution field via Vekua's analytic theory of elliptic PDEs, a realization important for future progress.

I have released a user-friendly, object-oriented and documented MATLAB toolbox (`MPSpack`, co-authored with Betcke) which implements the above methods, as well as the periodic methods described below, allowing the efficient solution of 2D wave scattering and eigenvalue problems on piecewise-homogeneous domains. Codes are human-readable and have a simple interface. For example, `s = segment(20, [0 1])` creates a line segment from 0 to 1 with 20 Clenshaw-Curtis quadrature points; from this, domains, basis sets, and problems may be set up and solved. The 70+ pages of tutorial and user manual include many code examples and figures. Publications [15] and [18] include `MPSpack` codes of a few lines which solve complete example problems with the

[1] Numbered citations are to the works enumerated in the publications section of my CV

new algorithms, allowing other researchers to validate and make use of my work.

One future goal is to exploit analytic and contour integral formulae for the fundamental solution in graded index media in 2D and 3D, continuing work arising with former Ph.D. student Matt Mahoney. This will enable integral equation methods to be applied to continuously-varying material problems for the first time.

1.1.1 Periodic problems: diffraction gratings and photonic crystal band structure

Many devices designed to guide and control waves rely on periodic structures on the wavelength scale: these include antennae, diffraction gratings (used to squeeze multiple signals onto a single optical fiber, and in our highest-powered lasers), photonic crystals (the most promising route to energy-efficient ultra-fast optical computation on a chip), meta-materials (allowing the control of waves in ways impossible in naturally-occurring media), sound absorbers, and solar cells. In recent work with Leslie Greengard (NYU) I have devised new efficient approaches to solving such periodic frequency-domain problems using integral equations, which, unlike existing methods, are robust for all problem parameters.

Integral equations (as with the global bases discussed above) reduce the dimensionality of the problem: a 2D wave problem becomes a 1D integral equation involving the Green's function kernel on the material boundaries. To solve the problem of an infinite array of objects, one must apply quasi-periodic boundary conditions to the solver for a single object. The conventional route is via the *quasi-periodic Green's function*. This turns out to have major drawbacks: i) it breaks down near certain parameters (the so-called Wood's anomalies) even though the boundary value problem remains well-posed, and ii) in large-scale problems its computation via lattice sums is cumbersome, and inefficient in geometries with high aspect ratio. We bypass both problems by using the free space Green's function alone, adding auxiliary layer potentials on the unit cell boundary, while imposing quasi-periodicity as a new linear condition.

With this new approach, in [16] we demonstrated a new spectrally-accurate solver for 2D photonic crystal band structure (see Fig. 1c), that is, the eigenvalue problem for the frequencies of traveling waves with given quasi-periodicity. We periodized using only around 100 extra degrees of freedom, achieving errors approaching machine precision using only small systems that are set up and solved in less than 1 second of computer time. The scheme is shown to be robust at parameters where the conventional quasi-periodic Green's function blows up (the resonances of the empty unit cell). Along the way we provided the first proof (via Calderón projectors) that the conventional scheme is equivalent to the band-structure problem.

We proceeded to the problem of scattering from a grating (infinite 1D array) of smooth 2D dielectric obstacles in [18]: a complication arises in evaluating layer potentials on unit cell walls since they now are infinite in extent. We tackle this using spectral quadrature in the Fourier domain, explicitly handling poles in the complex plane near Wood's anomalies. Thus, as above, our method is robust at parameters where the standard approach fails. Because of careful quadrature in the Fourier variable, we achieve 10^{-14} errors with a few hundred unknowns (see Fig. 1d), whilst maintaining the 2nd kind nature of the system (essential for a low condition number in iterative solvers).

We designed the above schemes to scale well to large-scale problems (10^4 or more unknowns), and allow fast multipole acceleration to be included without fuss. My future plans are to adapt the method to multi-layer dielectrics with embedded inclusions, to solve the scalar (and eventually, full Maxwell) band structure for 3D crystals, and to the technologically-important case of doubly-periodic layered scattering problems in 3D.

1.1.2 Laplacian eigenvalue problems

Eigenvalue problems for the Laplace operator ('modes of the drum') have a wealth of applications in engineering and physics, such as acoustic and electromagnetic resonators and waveguides, vibrations of membranes, quantum mechanics, and, more recently, the design of higher-power micro-cavity lasers for communications applications. This classical problem also has growing mathematical interest in the fields of spectral geometry and quantum chaos (see Section 1.2). Many of these areas demand efficient numerical solution at high frequencies, that is eigenvalue numbers from 10^2 up to 10^6, where conventional discretization methods fail. I have developed, and analyzed rigorously, new particular solution based methods—for instance the scaling method—which solve this problem in regimes where all other known methods are impractical.

In [14], motivated by my numerical observations, I proved that for large Dirichlet eigenvalues, that is, high

frequency applications, the method of particular solutions (MPS) in fact has error bounds a factor of the square-root of the eigenvalue *better* than was previously known (in exact arithmetic). In some applications this guarantees 3 extra digits of accuracy with no extra effort. The tools needed were Kato-Rellich perturbation theory and Cauchy interlacing, applied to a certain boundary operator that depends analytically on an 'energy' (or trial eigenvalue) parameter E. This result applies for E in neighborhoods of unknown size around each eigenvalue.

Then, in collaboration with Andrew Hassell (ANU), in [19] we improved this result to apply for all parameters E, and also for the L^2 error bounds of the corresponding eigenfunctions. We gave a numerical example at around the 2500th eigenvalue where the relative error in eigenvalue is of order 10^{-14}, and eigenfunction error 10^{-10}, for a smooth planar domain (Fig. 1b). A central tool in our analysis was the proof of a new form of boundary *quasi-orthogonality* of a domain's Dirichlet eigenfunctions (more specifically, their normal derivative functions), a PDE result with potential impact beyond numerical analysis. Our work in progress extends the above sharp bounds to large Neumann eigenvalues (with applications to acoustic resonators); it turns out that, for this boundary condition, microlocal analysis becomes essential if similar sharp bounds are to be achieved.

In recent work that we are currently writing up, we give the first rigorous analysis of the *scaling method* of Vergini–Saraceno. The latter is little-known accelerated variant of the MPS which is $O(\sqrt{E})$ times more efficient; in high-frequency applications the speed-up can be a factor of 10^3. Its mechanism deserves a brief explanation. To find Dirichlet eigenvalues with the MPS one must search along the E parameter axis to locate minima of a function (a Rayleigh quotient); the search is laborious and unreliable since the function 'zigzags' in an unpredictable fashion. Using only the effort required above for a single function evaluation, the scaling method instead approximates directly the set of *all* eigenpairs in a spectral window of size $\epsilon = O(\sqrt{E})$. We study this using the linearization of the E-dependence of the eigenvalues of a weighted Neumann-to-Dirichlet map for the interior Helmholtz boundary value problem. We prove that the errors in the computed eigenvalues are $O(\epsilon^2/E + \epsilon^3/E)$, its first known error bound. This is important for the reliable use of the method.

It is worth noting that such a collaboration between pure and applied mathematicians is quite rare, and in this case is opening avenues that neither researcher could tackle alone. My study of the scaling method originated in my physics thesis work [2], however at a less rigorous level. I believe that the method has the potential to be useful in a broad range of engineering and other eigenmode applications, and I plan to continue to advance that goal. For instance, I aim to publish (this time in collaboration with Betcke) side-by-side comparisons of this against other numerical eigenmode tools (including integral equation and finite elements), for a range of test cases relevant to applications in 2D and 3D.

1.2 Numerical study of quantum chaos

Quantum chaos is the study of how the allowed vibrational patterns and vibrational frequencies of a region of space (imagine for instance a drum-head) depend on its shape. Mathematically, these patterns are given by eigenfunctions of the Laplace operator, and the frequencies by the corresponding eigenvalues. It turns out that the statistical behavior of the patterns as one tends to very high frequencies (hence rapid spatial oscillations) depends on the type of motion that a "billiard ball" (or ray of light reflecting off the walls) would undergo in that same domain. If the billiard ball motion is regular (such as in a rectangular billiard table), one gets a distribution of eigenvalues and eigenfunctions that is very different from the case of *chaotic* billiard motion (for instance a table with curved walls). This idea has deep connections to quantum physics, geometry, and number theory, and involves many open questions at the forefront of contemporary mathematics.

More formally, the asymptotics as eigenvalue tends to infinity depends on the integrability of the geodesic flow on the manifold. For ergodic flow, we have a partial understanding in the form of rigorous results such as the Quantum Ergodicity Theorem, which states that *almost all* eigenfunctions become asymptotically equidistributed (spatially, and also in phase space). The Quantum Unique Ergodicity (QUE) conjecture of Rudnick–Sarnak is that, for uniformly-hyperbolic flow, this holds for *every* eigenfunction. It is believed that we are far from a proof of this conjecture (apart from arithmetic cases, for which it has only recently been proven).

This motivated my large-scale numerical study [9] in a Sinai-type billiard (flat manifold with boundary) with uniformly-hyperbolic flow; see Fig. 1a. I analyzed 30000 eigenfunctions up to eigenvalue number 7×10^5, and found strong evidence for QUE, as well as for conjectures arising in the physics literature on the power-law decay of the fluctuations in the spatial equidistribution of eigenfunctions. In a subsequent study with

collaborator Timo Betcke [11], we computed 16000 eigenfunctions of the 'mushroom' billiard, a manifold whose flow is *mixed*: it has two invariant regions in phase space, one integrable and one ergodic. We verified (to an accuracy of 1%) Percival's conjecture on the localization of eigenfunctions, and proposed and tested a model for so-called dynamical tunneling from the integrable into the ergodic region.

In both of the above works, developing specialized numerical tools enabled me to reach eigenvalues 10–100 times higher than previously attained, in very reasonable computation times. In the case of the Sinai-type domain, a fundamental solutions global basis was developed. For the mushroom, a Fourier-Bessel basis set adapted to the corner singularity was required. Only by adapting the scaling method (see Section 1.1.2) to these new basis sets could high accuracy and high efficiency be achieved. The mathematical significance of the work [9] is highlighted by its appearance as images on the cover of *Notices of the American Mathematical Society* (January 2008). Related eigenmode images of mine have appeared in four recent review articles, including one by Peter Sarnak (arguably the world leader in this field).

I am excited to continue the study of chaotic eigenfunctions, since it is an area of 'experimental mathematics' where advanced numerical methods can make an impact. In particular, I am keen to study the so-called *bouncing-ball modes* that occur between two parallel flat walls (in collaboration with Andrew Hassell), and to measure the statistics of the numbers of *nodal domains* (regions of positive or negative sign) in eigenfunctions to test the percolation conjectures of Bogomolny–Schmit. The latter project will involve undergraduates, including the proposed senior thesis of Kyle Konrad '12.

1.3 Other applied mathematics research avenues

1.3.1 Ecology and home range modeling

The way animals use their habitat is controlled by factors such as food, water, terrain, predators, their 'home ranges', and social interactions; revealing these effects is key for conservation efforts. In a collaboration with Paul Moorcroft (Harvard), we have developed a new animal movement Markov model [10], [12]. This unifies two standard approaches: resource selection analysis (a spatially implicit approach), and mechanistic home range analysis (an explicit probabilistic kinetic model). Our model's kernel may be factorized to derive an analytic steady state solution, yet has the flexibility to model variation in habitat via a general spatial *preference function* w. As a result, such models could be fit to data much more rapidly than before, and the time-evolution (probability density functions) of moving animals can be simulated quickly using fast Fourier transform methods. The analytic solution allowed us to predict that steady state animal density scales like w on small scales but like w^2 on large scales, with consequences for current practices of real-world telemetry data analysis and model-fitting.

1.3.2 Diffuse optical tomography

Near-infrared light is a promising tool for non-invasive medical imaging, yet requires the solution of a complex inverse problem to reconstruct the images. Following on earlier work on Bayesian statistical solution of the inverse problem [6], and the development of a rapid solver for the PDE forward problem (heat equation) in multi-layer geometries [7], I have continued to contribute in this area with work validating the diffusion approximation against the more accurate, but much more computationally intensive, radiative transport model, in 3D models of the human head [8]. One conclusion is that the diffusion approximation may often be used, allowing a much faster solution of the inverse problem for brain imaging.

1.3.3 Mathematics of music

In a new collaboration [17] with Larry Polanksy (Dartmouth) and Michael Winter (UCSB), we have analyzed a framework for the random generation of musical melodies with varying amounts of temporal correlation. This simplifies and generalizes methods of James Tenney, one of the first composers to apply computer code to the creation of music in the 1960s. Our cross-disciplinary approach combines stylistic analysis, numerical experimentation, and rigorous mathematical analysis.

2 Teaching

In the classroom I make learning an interactive discovery process: most of my class periods involve a 10-15 minute session where students work in groups of two or three to solve problems on worksheets that I have devised. I am then able to move around the classroom, asking probing questions or giving hints, and learning how each individual thinks. This builds their communication skills (the strong students often teach the weaker ones better than I could), while giving me instant feedback on their areas of difficulty, and helping me steer the remainder of lecture. The classroom buzzes with discussion, and the energy level is high. Students take an active rather than passive role: by the end of lecture they have solved one or more core problems themselves, rather than merely copied down solutions from the board. I scan and post all worksheets and solutions online for later study (see any of my course webpages). Student feedback on the use of worksheets is consistently enthusiastic. In developing this idea I was influenced by my time as a teaching assistant with Eric Mazur (Harvard), a pioneer in such peer instruction techniques. Peer instruction has been proven to be successful with students of different levels and personalities, and to increase female enrollment and retention in math and the sciences.

Another feature of many of my lectures is a 'live demo' involving some real world object that illustrates or brings a concept to life. Examples include: waving around huge colored 'vectors' to explain cross product, using masses on springs or torsional wave machines for differential equations, casting shadows of solids with a flashlight (or penetrating them with knitting needles) for triple integrals, sounding coupled tuning forks to illustrate resonance, hitting drums (or whirling flexible plastic tubing above my head!) while measuring their frequency spectrum to illustrate eigenvalue problems, and taking students outside to measure the speed of sound themselves using clapping and echoes. I also use computer demos, for instance showing students applets that let them see vector fields using fluid flow. I find that visualization is key to student understanding and engagement.

Since computation plays an important and growing role in applied mathematics and the sciences, I have woven a computer component (mostly MATLAB/Octave, but also online applets and audio analysis software) into almost every course I teach. Thus students coming out of my courses have (to varying degrees) experience with computational tools that modern scientists use—a research and career advantage. At the major and graduate level I emphasize programming skills more strongly, so that students learn how to debug their own mistakes, and build more complex routines reliably.

In my five years at Dartmouth so far I have developed from scratch four new courses, each of which attracts healthy enrollments (I have been fortunate to be able to teach each two or three times). These courses greatly expand the applied offerings at Dartmouth, and the first two now form a core part of our applied math major:

- Math 46: *Introduction to Applied Mathematics*, a math- and science-major course covering core analytical tools such as dimensional analysis, perturbation series, integral equations, Green's functions, Fourier transforms, and analysis of PDEs.
- Math 53: *Chaos!*, a math- and science-major course introducing phenomena in dynamical systems (maps and flows), their rigorous analysis, numerical simulation, and applications.
- Math 5: *The Mathematics of Music and Sound*, a non-science major course that fulfills the QDS (quantitative distribution) requirement, covering musical acoustics, functions and signals, Fourier analysis, waves, musical instruments, human hearing, and digital sound technology (all without calculus!)
- Math 116: *Numerical Methods for PDEs and Waves*, a graduate-level course on numerical analysis, scientific computing, and integral equations, incorporating material from my research.

In Math 5, 53, and 116, I make independent project work a key part of the assessment (in Math 53 this replaces the final exam); students respond very enthusiastically, since now they feel ownership of the work they are doing, and experience the excitement of research. In fact, I have seen Dartmouth students 'pull out all the stops' when finishing projects like at no other time, and experience great satisfaction when completed. Undergraduate research projects have included computing the Lyapunov exponents in chemical chaos, testing new numerical methods for solving PDEs in the plane, and measuring the temporal sensitivity of the human ear. In Math 5 and Math 50 (*Mathematical Statistics*) I also set up an online wiki/comments page where students can post examples of each week's material from everyday life and culture, connecting the content to applications, and providing discussion material relevant to their lives. In all of the above classes I grade students on the quality of

their communication skills, and work with them to create engaging and clear class presentations and write-ups. I often have a final presentation night with pizza—a mini math conference—where students learn about each other's projects, and learn how to communicate in a microcosm of a professional science community.

My future goals include developing the numerical analysis and computational sides of the undergraduate and graduate curricula, and continuing to create and refine interactive teaching tools in all of my classes.

3 Service to Department and College

During my time at Dartmouth I have mentored nine undergraduate research projects on topics such as numerical methods, waves, imaging and chaos. Four of these were senior theses (of which three received *high honors*): examples include Yong Su '09, on spectral methods for computing the capacity of the unit square and Laplacian eigenmodes on the sphere (Yong is now in graduate school at Stanford), and Zoe Lawrence '10 on PDE models for the spread of insect-borne disease across farmland (co-advised with Dorothy Wallace). Two more students have been involved in an interdisciplinary project to develop image-processing algorithms to track the genealogy of cell nuclei in 3D microscopy videos, co-advised with Amy Gladfelter (Biology). Preliminary work suggests that we can significantly boost the amount of data the Gladfelter lab can process. Through the *Women in Science Project* I have mentored Vissuta Jiwariyavej '09 on acoustics and number theory, will be mentoring a *WISP* first-year Hahn Nguyen '14 starting Winter 2011, on numerical analysis. Other outreach to undergraduates includes giving four presentations to the Dartmouth Mathematics Society. For the last three years I have served on the Faculty Advisory Board to the *Dartmouth Undergraduate Journal of Science*, and in this time worked with two students to edit their mathematics articles for publication in *DUJS*.

I have served on various department committees including Recruiting, both the Undergraduate and Graduate program committees, and Graduate Admissions (see CV for a full list). I was also heavily involved in revamping the Department's applied math major curriculum during my first couple of years here, as part of an ad-hoc committee with Scott Pauls, Dan Rockmore, and Peter Doyle. In terms of mentoring other faculty, I have given two presentations at the *Dartmouth Center for the Advancement of Learning*, one solo workshop on the use of worksheets to create interaction in science classes, and one group panel on teaching for those new to Dartmouth.

At the graduate level, I advised Matt Mahoney (Ph.D. '09) from 2006–2007 on the method of particular solutions for graded index materials. In conjunction with Scott Pauls and Dan Rockmore, I have developed our Department's new applied math graduate qualifying exam. I then served on the applied qualifying exam committee of Katherine Kinnaird. I have also given three introductory talks to attract prospective graduate students to our Department.

In my first year at Dartmouth I started the Applied and Computational Mathematics Seminar, which has now hosted 34 talks in total (many being invited visitors). This research seminar brings together a community of researchers across campus interested in numerical modeling; its audience includes engineers, physicists and biologists.

Finally, I was one of four organizers of the 8-day-long *International Conference on Spectral Geometry*, which brought around 80 international visitors, including most of the major people in the field, to Dartmouth for a very successful research conference this July. The other organizers were Carolyn Gordon, Peter Perry (Kentucky) and Alejandro Uribe (Michigan). We feel that this has enhanced Dartmouth's already strong reputation in quantum chaos and spectral geometry.

Alex H. Barnett

Department of Mathematics
Kemeny Hall 6188
Dartmouth College
Hanover, NH, 03755
(603) 646-3178

Born: 7th December, 1972
U.S. permanent resident

email: `ahb` at `math.dartmouth.edu`
`http://math.dartmouth.edu/~ahb`

RESEARCH FOCUS

- Numerical analysis, scientific computing, computational partial differential equations: Helmholtz equation, particularly high frequency and/or high accuracy; eigenvalue, wave scattering and periodic problems, photonic crystals; corner singularities, high-order methods, integral equations; heat equation.
- Mathematical physics: quantum chaos and ergodicity, Laplacian eigenfunctions, scarring, billiards.
- Applied mathematics: mathematical ecology (animal home range modeling), inverse problems in imaging (diffuse optical tomography), mathematics of music.

EDUCATION

Harvard University. Ph. D. in Physics: *Dissipation in Deforming Chaotic Billiards* — October 2000
Thesis advisor: Eric J. Heller

Cambridge University, England. B. A. *(First class)* in Theoretical Physics — June 1994
Undergraduate thesis advisor: David J. C. MacKay

POSITIONS

Assistant Professor. — July 2005–present
Department of Mathematics, Dartmouth College, NH

Courant Instructor / Assistant Professor. — 2002–2005
Courant Institute of Mathematical Sciences, New York University, NY

Postdoctoral Research Fellow. — February–August 2002
Photon Migration Imaging Laboratory, Dept. of Radiology, Harvard Medical School, Charlestown, MA

Head Teaching Fellow. — January–June 2001
Department of Physics, Harvard University, Cambridge, MA

Consultant. — November–December 2000
TrueWind Solutions LLC, Lowell, MA

Teaching Fellow. — 1994–1999
Department of Physics, Harvard University, Cambridge, MA

FELLOWSHIPS AND AWARDS

- *National Science Foundation, Grant DMS-1005360*, co-PI, award amount $42,355 — 2010
  Funding support for *International Conference on Spectral Geometry*
- *The Class of 1962 Faculty Fellowship* (competitive award), Dartmouth College — Winter 2010
- *National Science Foundation, Grant DMS-0811005*, PI, award amount $310,517 — 2008–2011
  "Efficient spectrally accurate global basis methods for high frequency wave scattering, eigenmodes, and photonics"
- *National Science Foundation, Grant DMS-0507614*, PI, award amount $102,520 — 2005–2008
  "High frequency cavity eigenmodes: rapid computation methods, applications and asymptotics"
- *Courant Instructorship*, New York University — 2002–2005
- *Postdoctoral Fellowship*, Mathematical Sciences Research Institute (declined) — 2002

- *Harvard University Certificate of Distinction in Teaching* Fall 1997, Spring 2001 and Fall 2001
  ($\geq$4.5 out of 5 in anonymous student evaluations)
- *Harold T White Prizes*, teaching introductory physics, Harvard Physics Department 1996 and 1997
- *Kennedy Scholarship*, Kennedy Memorial Trust, London 1994
  One of 10 chosen annually for funded year of US graduate research
- *Hockin Prize*, *Duchess of Somerset Scholarships*, St John's College, Cambridge 1993 and 1994
- XXI International Physics Olympiad, *First Prize* of 155 entrants, Groningen, The Netherlands 1990

PUBLICATIONS

[19] "Boundary quasi-orthogonality and sharp inclusion bounds for large Dirichlet eigenvalues," A. H. Barnett and A. Hassell, *submitted, SIAM J. Numer. Anal.*, 17 pages (May 2010).

[18] "A new integral representation for quasi-periodic scattering problems in two dimensions," A. H. Barnett and L. Greengard, *in press, BIT Numer. Math.*, 24 pages (submitted Aug 2010)

[17] "A few more words about James Tenney: dissonant counterpoint and statistical feedback," L. Polansky, A. H. Barnett, and M. Winter, *accepted, J. Math. Music*, 30 pages (submitted Oct 2009)

[16] "A new integral representation for quasi-periodic fields and its application to two-dimensional band structure calculations," A. H. Barnett and L. Greengard, *J. Comput. Phys.*, **229** (19), 6898–6914 (2010)

[15] "An exponentially convergent nonpolynomial finite element method for time-harmonic scattering from polygons," A. H. Barnett and T. Betcke, *SIAM J. Sci. Comp.* **32** (3), 1417–1441 (2010)

[14] "Perturbative analysis of the Method of Particular Solutions for improved inclusion of high-lying Dirichlet eigenvalues," A. H. Barnett, *SIAM J. Numer. Anal.* **47**, 1952–1970 (2009)

[13] "Stability and convergence of the Method of Fundamental Solutions for Helmholtz problems on analytic domains," A. H. Barnett and T. Betcke, *J. Comput. Phys.* **227** (14), 7003–7026 (2008)

[12] "Analytic steady-state space use patterns and rapid computations in mechanistic home range analysis," A. H. Barnett and P. R. Moorcroft, *J. Math. Biol.* **57** (1), 139–159 (2008)

[11] "Quantum mushroom billiards," A. H. Barnett and T. Betcke, *CHAOS* **17**, 043123, 13 pages (2007)

[10] "Mechanistic home range models and resource selection analysis: a reconciliation and unification," P. R. Moorcroft and A. H. Barnett, *Ecology* **89** (4), 1112–1119 (2008)

[9] "Asymptotic rate of quantum ergodicity in chaotic Euclidean billiards," A. H. Barnett, *Comm. Pure Appl. Math.* **59**, 1457–1488 (2006)

[8] "Effective scattering coefficient of the cerebral spinal fluid in adult head models for Diffuse Optical Imaging", A. Custo, W. M. Wells III, A. H. Barnett, E. M. C. Hillman, and D. A. Boas, *Applied Optics* **45**, 4747–55 (2006)

[7] "A fast numerical method for time-resolved photon diffusion in general stratified turbid media," A. H. Barnett, *J. Comput. Phys.* **201**, 771–797 (2004)

[6] "Robust inference of baseline optical properties of the human head with 3D segmentation from magnetic resonance imaging," A. H. Barnett, J. P. Culver, A. G. Sorensen, A. M. Dale, and D. A. Boas, *Applied Optics* **42**, 3095–3108 (2003)

[5] "Parametric evolution for a deformed cavity," D. Cohen, A. H. Barnett, and E. J. Heller, *Phys. Rev. E* **63**, 046207, 12 pages (2001)

[4] "Mesoscopic scattering in the half-plane: squeezing conductance through a small hole," A. H. Barnett, M. Blaauboer, A. Mody, and E. J. Heller, *Phys. Rev. B* **63**, 245312/1 (2001)

[3] "Rate of energy absorption for a driven chaotic cavity," A. H. Barnett, D. Cohen, and E. J. Heller, *J. Phys. A* **34**, 413–437 (2001)

[2] "Deformations and dilations of chaotic billiards: dissipation rate, and quasi-orthogonality of the boundary wavefunctions," A. H. Barnett, D. Cohen, and E. J. Heller, *Phys. Rev. Lett.* **85**, 1412–15 (2000)

[1] "Substrate-based atom waveguide using guided two-color evanescent light fields," A. H. Barnett, S. P. Smith, M. Olshanii, K. S. Johnson, A. W. Adams, M. Prentiss, *Phys. Rev. A* **61**, 023608, 11 pages (2000)

Conference proceedings, technical reports:

"Tensor product of kernel models," O. de la Cruz, A. H. Barnett, H. Tang, and S. Holmes, *submitted*, NIPS, 4 pages (2010).

"Accurate and robust computation of photonic crystal band structure using second-kind integral equations," A. H. Barnett and L. Greengard, Proceedings of WAVES2009: The 9th International Conference on Mathematical and Numerical Aspects of Wave Propagation, 2 pages (2009).

"Quasi-orthogonality on the boundary for Euclidean Laplace eigenfunctions," A. H. Barnett, *preprint*, `math-ph/0601006`, available at `http://arxiv.org`, 21 pages (2004).

"Bayesian estimation of optical properties of the human head via 3D structural MRI," A. H. Barnett, J. P. Culver, A. G. Sorensen, A. M. Dale, and D. A. Boas, *Proc. SPIE* **5138**, 9 pages (2003)

"Bayesian Comparison of Models for Images," A. H. Barnett and D. J. C. MacKay, in *Maximum Entropy and Bayesian Methods*, Proceedings of MAXENT94 (Kluwer, 1996), p. 239–248

Software:

`MPSpack` (coauthor T. Betcke, Reading, UK). A MATLAB toolbox to solve Helmholtz and scattering problems with particular solutions and integral equations. Version 1.0 released September 2009 (tutorial 35 pages, manual 37 pages); downloaded at least 200 times. Freely available at `http://code.google.com/p/mpspack`

Other codes for numerical solution of PDEs freely available at `http://math.dartmouth.edu/~ahb/software`

Partial list of researchers using these codes: M. Ganesh (Colorado Mines), D. Boas (Harvard), M. Porter (Oxford), M. Wright (ISVR, UK), P. Doyle (Dartmouth), C. Mansouri (ISAIP/ESAIP, Paris)

Publicity / outreach / education:

Computed eigenmodes for integrable and chaotic billiards, appearing in: P. Sarnak, "Recent progress on QUE" (2009), and D. Mackenzie, *What's Happening in the Mathematical Sciences, Volume 8* (AMS), (Jan 2011).

"convolution: *son et lumière*", A. H. Barnett, *in press*, issue 01 of *Convolution. A journal for experimental criticism*, 4 pages (Oct 2009).

Cover of *Notices of the American Mathematical Society*, January 2008. I created all images (modes $1, 10, 10^2, 10^3, 10^4, 10^5$ of a planar chaotic cavity), and eigenvalue data for article "Quantum Chaos" by Z. Rudnick.

Quantum chaos research featured in "A Decade of Science at Dartmouth", W. Schpero and C. Chiang, *Dartmouth Undergraduate Journal of Science*, Spring 2008.

INVITED RESEARCH TALKS

Boundary value problems for linear elliptic and integrable PDEs, ICMS, Edinburgh, UK	May 2012
Waves in periodic geometries, mini-symposium, ICIAM, Vancouver	July 2011
Numerical Analysis and Scientific Computing Seminar, Courant Institute, NYU	January 2011
Applied Mathematics and Computational Science Colloquium, U. Penn	January 2011
Conference in honor of 65th birthday of Eric Heller, ITAMP, Harvard	October 2010
Integral Equation Methods, Fast Algorithms and Applications, IMA workshop, Minnesota	August 2010
Frontiers in Computational and Applied Mathematics, NJIT, Newark, NJ	May 2010
Numerical solution of the Painlevé equations, ICMS, Edinburgh, UK	May 2010

Applied Analysis Seminar, Louisiana State University	March 2009
Dartmouth Mathematics Colloquium	November 2009
MIT Applied Mathematics Colloquium	November 2009
Computational Optical Sensing and Imaging Seminar, CU Boulder	September 2009
Topological Complexity of Random Sets, AIM workshop, Palo Alto	August 2009
WAVES2009, Pau, France (contributed talk)	June 2009
Colloquium & PDE/Analysis Seminar (two separate talks), ANU, Canberra, Australia	February 2009
Laplacian Eigenvalues & Eigenfunctions: Theory, Computation, Application, IPAM, UCLA	February 2009
Workshop on Numerical and Analytical Methods for Wave Scattering, Manchester, UK	June 2008
Workshop on quantum chaos, CRM (Univ. de Montréal)	June 2008
Frontiers in Computational and Applied Mathematics, NJIT, Newark, NJ (contributed poster)	May 2008
McGill Applied Mathematics Seminar, Montreal	March 2008
Applied Mathematics Seminar, U. Delaware	November 2007
Numerical analysis seminar, Manchester, UK	July 2007
Three separate mini-symposium talks, ICIAM, Zurich	July 2007
PDE/Analysis Seminar, ANU, Canberra, Australia	February 2007
Heller Group Seminar, Physics Department, Harvard	December 2006
Applied Mathematics Seminar, UMass Amherst, MA	November 2006
Dartmouth Physics Colloquium	October 2006
SIAM Annual Meeting, Boston (contributed talk)	July 2006
Joint MIT/Harvard Analysis Seminar	March 2006
Computations in Science Seminar, U. Chicago	November 2005
Institute of Sound and Vibration Research, Southamption University, UK	June 2005
Computational Mathematics and Applications Seminar, Computing Laboratory, Oxford, UK	June 2005
Numerical Analysis and Scientific Computing Seminar, Courant Institute, NYU	December 2004
Applied Mathematics Colloquium, Columbia University	October 2004
Workshop on Semi-classical Theory of Eigenfunctions and PDEs, CRM (Univ. de Montréal)	June 2004
Applied Mathematics Seminar, Yale University	March 2004
Applied Mathematics Laboratory Seminar, Courant Institute, NYU	February 2004
European Conference on Biomedical Optics, Munich, Germany (contributed talk)	June 2003
Theoretical & Computational Biology Seminar Series, Mount Sinai School of Medicine, NY	May 2003
Applied Mathematics Seminar, Courant Institute, NYU	November 2002
Photon Migration Imaging Seminar, Harvard Medical School	May 2002
Applied Mathematics Laboratory Seminar, Courant Institute, NYU	December 2001
Statistics Seminar, University of Toronto	September 2001
Pan-American Advanced Study Institute on Quantum Chaos, Ushuaia, Argentina	October 2000
14th Maximum Entropy Workshop (MAXENT94), St John's College, Cambridge, England	June 1994

TEACHING

Dartmouth College *(the four new courses I created in the period 2006–2008 are shown by asterisks)*

- Math 53: Chaos! (dynamical systems)* — Fall 2007, 2009
- Math 5: The Mathematics of Music and Sound (for non-majors)* — Spring 2007, Fall 2008, 2010
- Math 46: Introduction to Applied Mathematics (for majors)* — Spring 2007, 2008, 2009, 2011
- Math 22: Linear Algebra with Applications — Summer 2006
- Math 11: Multivariable Calculus — Fall 2010
- Math 116: Numerical Methods for PDEs and Waves (graduate level)* — Winter 2006, Fall 2008
- Math 50: Probability and Statistical Inference — Winter 2006
- Math 23: Differential Equations — Fall 2005, 2007

New York University

- Business Calculus — Fall 2004
- Mathematical Wave Dynamics (new honors VIGRE course, with O. Bühler) — Spring 2004
- Linear Algebra — Fall 2003
- Multivariable Calculus — Spring 2003
- Undergraduate Math Lab (new honors VIGRE course, with P. Sarnak) — Fall 2002

Harvard University

- Laboratory Electronics (TA) — Fall 2001
- The Physics of Music and Sound (head TA, created labs and content) — Spring 2001
- Advanced Quantum Mechanics (TA, graduate course) — Fall 1998
- *Microteaching Facilitator* for the Derek Bok Center — Fall 1997
- Mechanics (TA) — Fall 1996, 1997
- Principles of Physics (TA) — Summer 1996
- Introductory Mechanics and Relativity (TA) — Spring 1996
- Mechanics and Special Relativity (TA) — Fall 1995

STUDENTS & MENTORING

Hahn Nguyen '14 (first-year WISP intern, Women in Science Program) — January–June 2011
Accurate evaluation of layer potentials up to the boundary

Zoe Lawrence '10 (senior thesis, with 'high honors', co-supervisor Dorothy Wallace) — Spring 2010
The spatiotemporal dynamics of African Cassava Mosaic Disease

Kathleen Champion '11 (Presidential Scholar and senior thesis, co-supervisor Amy Gladfelter, Biology)
Three-dimensional tracking of nuclear mitosis — January 2010–present

Emmanuel Mensah '09 (independent study) — Spring 2009
The inverse source problem in medical imaging (published in DUJS, November 2009)

Yong Su '09 (senior thesis, with 'high honors') — September 2008–March 2009
Computing the capacitance of the unit cube

Evan Tice '09 (computer science major, co-supervisor Amy Gladfelter, Biology) — January 2008–June 2009
Automated image tracking of cell movement and division (awarded Kemeny Prize, 2008)

Chetan Mehta '08 (senior thesis, with 'high honors') — June 2007–May 2008
Optimal optode location in Diffuse Optical Tomography

Chor Lam '08 (Presidential Scholar) — January–June 2008
Chaos in billiards

Matt Mahoney (Ph. D. '09) — July 2006–July 2007
Global numerical methods for eigenmodes with gravity

Vissuta Jiwariyavej '09 (sophomore WISP intern, Women in Science Program) — January–June 2007
A clap can chirp: waves and echoes in the racquetball court

External: William A. Webb (Caltech, advisor Mason Porter) — Summer 2006
A computational study of the quantization of billiards with mixed dynamics

SERVICE TO DEPARTMENT AND COLLEGE

Conference organizer:

Organizer (one of four), *International Conference on Spectral Geometry*, Dartmouth College — July 2010

Seminar organizer:

Applied and Computational Mathematics Seminar, Dartmouth (hosted 34 talks so far) — May 2006–present

Mathematics education:

Interactive Learning in the Sciences, session given at Dartmouth Center for the Advancement of Learning — October 2007

Cross-disciplinary classroom visits by professional musicians to Math 5, 53 — 2007, 2008, 2010

Committee member, Applied Mathematics Qualifying Examination, Katherine Kinnaird — Spring 2010

Outreach:

Advisory Committee, Leslie Center for the Humanities — Winter 2011–present

Faculty Advisory Board, Dartmouth Undergraduate Journal of Science (DUJS) — Fall 2007–present

Affiliated faculty, Electro-Acoustic Music Program, Dartmouth — 2007–present

Research talks for Dartmouth Mathematics Society — 2006, 2008, 2009, 2010

Mathematical model building, Dartmouth Mathematics Society — October 2006

Research talks for Dartmouth Graduate Recruiting Open House — 2006, 2007, 2009

Guest lecturer, Music 3 (Music and Technology), Dartmouth — November 2006

First-year advising — Fall 2007, 2008, 2010

Department committees:

Undergraduate Program Committee — 2005, 2008–2010

Recruiting Committee — 2007–2008, 2010–2011

Equipment Committee — 2007–2008

Graduate Program Committee — 2006–2007

Mirkil Book Committee — 2006–2007

Graduate Admissions Committee — 2005–2006

Computing Committee — 2005–2006

OTHER PROFESSIONAL ACTIVITIES

Conferences & seminar organizer, outreach:

Scientific Committee, *International Conference on Mathematical and Numerical Aspects of Wave Propagation (WAVES)*, Vancouver, BC — July 2011

Organizer (with L. Demanet), two minisymposia on numerical waves, ICIAM, Vancouver, BC — July 2011

Organizer (with J. Zhang), Applied Mathematics Laboratory Seminar, Courant Institute, NYU — 2002–2003

Organizer, Creating Careers in Physics series, Harvard Physics Department — 1999–2000

Co-organiser, Boston Area Undergraduate Physics Competition — 1998–2001

Referee, review panels:

SIAM Journal on Scientific Computing, Notices of the AMS, Inverse Problems, Proceedings of the Royal Society of London A, Engineering Analysis with Boundary Elements, Numerical Functional Analysis and Optimization, Experimental Mathematics, Applied Optics, Waves in Random and Complex Media, NeuroImage, Journal of Electronic Imaging, Physics in Medicine and Biology, Transactions on Medical Imaging, Journal of the Optical Society of America A, Nonlinear Dynamics, Journal of Physics A, Canadian Journal of Physics, New Journal of Physics; National Science Foundation.

Member: Optical Society of America, Society for Industrial and Applied Mathematics.

BOUNDARY QUASI-ORTHOGONALITY AND SHARP INCLUSION BOUNDS FOR LARGE DIRICHLET EIGENVALUES

A. H. BARNETT AND A. HASSELL

Abstract. We study eigenfunctions ϕ_j and eigenvalues E_j of the Dirichlet Laplacian on a bounded domain $\Omega \subset \mathbb{R}^n$ with piecewise smooth boundary. We bound the distance between an arbitrary parameter $E > 0$ and the spectrum $\{E_j\}$ in terms of the boundary L^2-norm of a normalized trial solution u of the Helmholtz equation $(\Delta + E)u = 0$. We also bound the L^2-norm of the error of this trial solution from an eigenfunction. Both of these results are sharp up to constants, hold for all E greater than a small constant, and improve upon the best-known bounds of Moler–Payne by a factor of the wavenumber $\sqrt{E}$. One application is to the solution of eigenvalue problems at high frequency, via, for example, the method of particular solutions. In the case of planar, strictly star-shaped domains we give an inclusion bound where the constant is also sharp. We give explicit constants in the theorems, and show a numerical example where an eigenvalue around the 2500th is computed to 14 digits of relative accuracy. The proof makes use of a new quasi-orthogonality property of the boundary normal derivatives of the eigenmodes (Theorem 1.3 below), of interest in its own right. Namely, the operator norm of the sum of rank 1 operators $\partial_n\phi_j\langle\partial_n\phi_j,\cdot\rangle$ over all E_j in a spectral window of width $\sqrt{E}$ — a sum with about $E^{(n-1)/2}$ terms — is at most a constant factor (independent of E) larger than the operator norm of any one individual term.

1. Introduction and main results. The computation of eigenvalues and eigenmodes of Euclidean domains is a classical problem (in two dimensions this is the 'drum problem', reviewed in [21, 34]) with a wealth of applications to engineering and physics, including acoustic, electromagnetic and optical cavity and resonator design, micro-lasers [35], and data analysis [30]. It also has continued interest in mathematical community in the areas of quantum chaos [37, 3] and spectral geometry [18]. Let ϕ_j be a sequence of orthonormal eigenfunctions and E_j the respective eigenvalues ($0 < E_1 < E_2 \leq E_3 \leq \cdots$ counting multiplicities) of $-\Delta$, where $\Delta := \sum_{m=1}^n \partial^2/\partial x_m^2$ is the Laplacian in a bounded domain $\Omega \in \mathbb{R}^n$, $n \geq 2$, with Dirichlet boundary condition. That is, ϕ_j satisfies

$$(\Delta + E_j)\phi_j = 0 \qquad \text{in } \Omega \tag{1.1}$$

$$\phi_j = 0 \qquad \text{on } \partial\Omega \tag{1.2}$$

$$\|\phi_j\|_{L^2(\Omega)} = 1. \tag{1.3}$$

We will call the spectrum $\sigma := \{E_j\}_{j=1}^\infty$. Many of the applications mentioned demand high frequencies, that is, mode numbers j from 10^2 to as high as 10^6. Efficient solution of the problem thus requires specialized numerical approaches that scale with wavenumber better than conventional discretization methods.

The goal of this paper is to bound the errors of approximate eigenvalues and eigenfunctions computed using trial functions that satisfy exactly the homogeneous Helmholtz equation in Ω. As we will review below, such computational methods have proven very powerful. Recently one of the authors [4] improved upon the classical eigenvalue bound of Moler–Payne [24] by a factor of the wavenumber; however, this result has limited utility since it applies only to Helmholtz parameters lying in neighborhoods of σ of unknown size. In the present paper we go well beyond this result by giving new theorems, which i) hold for *all* Helmholtz parameters (greater than an $O(1)$ constant), ii) retain the improved high-frequency asymptotic behavior of [4] and show that this behavior is sharp, and iii) improve upon the best-known eigenfunction estimates, again by a factor of the wavenumber. To achieve this we make use of a new form of quasi-orthogonality of the eigenfunctions on the boundary, Theorem 1.3, of independent interest.

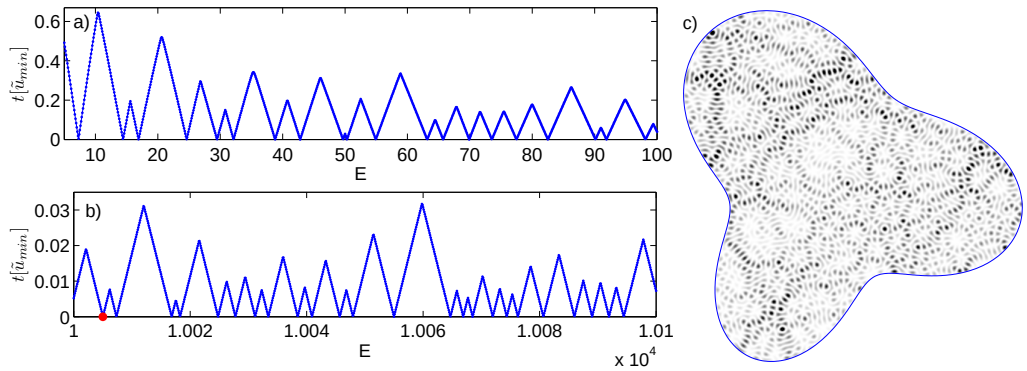


FIG. 1.1. *Tension $t[\tilde{u}_{\min}]$ versus energy E for the domain shown on the right. $\tilde{u}_{\min}$ is the optimal trial Helmholtz solution lying in the span of a numerical basis set (see Section 7). a) Low frequency, showing the minima corresponding to the lowest 20 Dirichlet eigenvalues. b) Medium-high frequency, showing a similar interval starting at eigenvalue number $j \approx 2552$; note the new vertical scale. c) Density plot of eigenfunction $\phi_j \approx \tilde{u}_{\min}$ corresponding to the eigenvalue $E_j = 10005.02135797\cdots$ shown by the dot in b) (black indicates large values of $|\phi_j|^2$, white zero).*

Before presenting our results, we need to review some known inclusion bounds and their importance for applications. Given an energy parameter[1] $E > 0$, let u be a non-trivial solution to the homogeneous Helmholtz equation $(\Delta + E)u = 0$ in Ω with no imposed boundary condition, and define its boundary error norm (or 'tension')

$$t[u] := \frac{\|u\|_{L^2(\partial\Omega)}}{\|u\|_{L^2(\Omega)}} \ . \tag{1.4}$$

Clearly, $t[u] = 0$ implies that E is an eigenvalue. It is reasonable to expect that if $t[u]$ is small for some Helmholtz solution u, then E is close to an eigenvalue. Moler–Payne [24] (building upon [16]) quantified this: there is a constant C_{MP} depending only on the domain, such that

$$d(E,\sigma) \ \leq \ C_{\text{MP}} E\, t[u] \ , \tag{1.5}$$

where $d(E,\sigma) := \min_j |E_j - E|$ denotes the distance of E from the spectrum.

An important application is to solving (1.1)-(1.3) via global approximation methods, including the method of particular solutions (MPS) [9, 4]. One writes a trial eigenmode $u = \sum_{n=1}^{N} c_n \xi_n$ via basis functions ξ_n which are closed-form Helmholtz solutions in Ω but which need not satisfy any particular boundary condition. By adjusting the coefficients c_n (via a generalized eigenvalue [4] or singular value problem [8]) one may minimize $t[u]$ at fixed E; by repeating this in a search for E values where the minimum $t[u]$ is very small, as illustrated by Fig. 1.1a and b, one may then locate approximate eigenvalues whose error is bounded above by (1.5). (This is sometimes called the method of *a priori-a posteriori* inequalities [21, Sec. 16].)

Due to the work of Betcke–Trefethen [9] and others, such methods have enjoyed a recent revival, at least in $n = 2$, due to their high (often spectral) accuracy and their efficiency at high frequency when compared to direct discretization methods such as finite elements. For example, in various domains, 14 digits may be achieved in double

[1] The Helmholtz parameter E may be interpreted as energy, or as the square of frequency, depending on the application.

precision arithmetic [9], and with an MPS variant known as the scaling method, tens of thousands of eigenmodes as high as $j \sim 10^6$ have been computed [3, 5]. (There are also successful variants [14, 15] by Descloux–Tolley, Driscoll, and others, in which subdomains are used, which we will not pursue here.)

If we instead interpret u as the solution error for an interior Helmholtz boundary-value problem (solved, for instance, via MPS or boundary integral methods), then (1.5) states that the interior error is controlled by the boundary error; this aids the numerical analysis of such problems [22, 6]. Similar estimates (which, however, rely on impedance boundary conditions) enable the analysis of least-squares non-polynomial finite element methods [25, Thm 3.1]. Improving such estimates could thus be of general benefit for the numerical solution of Helmholtz problems.

Recently one of the authors [4] observed numerical evidence that (1.5) is not sharp for large E, and showed that there is a constant C_{B} depending only on Ω, such that, for each $\varepsilon > 0$,

$$d(E, \sigma) \;\leq\; C_{\text{B}}(1+\varepsilon)\sqrt{E}\, t[u] \tag{1.6}$$

holds whenever E lies in some open, possibly disconnected, subset of the real axis containing σ. This is an improvement over (1.5) by a factor of the wavenumber $\sqrt{E}$, which in problems of interest can be as high as 10^3. However, since the proof relied on analytic perturbation in the parameter E, there was no knowledge about the *size* of this (ε-dependent) subset, hence no way to know in a given practical situation whether the error bound holds. The point of the present work is then to remedy this problem by removing any restriction to an unknown subset, and also to extend the $\sqrt{E}$ improvement to bounds on approximate eigenfunctions.

We assume the domain $\Omega \subset \mathbb{R}^n$ has unit area (or volume for $n > 2$), and obeys the following rather weak geometric condition.

Condition 1. *The domain $\Omega \subset \mathbb{R}^n$ is bounded, with piecewise smooth boundary in the sense of Zelditch–Zworski [38]. This means that Ω is given by an intersection*

$$\Omega = \bigcap_{i=1}^{N} \{\mathbf{x} \mid f_i(\mathbf{x}) > 0\},$$

where the f_i are smooth functions defined on a neighborhood of $\overline{\Omega}$ such that

- *$\nabla f_i \neq 0$ on the set $\{f_i = 0\}$,*
- *$\{f_i = f_j = 0\}$ is an embedded submanifold of $\mathbb{R}^n$, $1 \leq i < j \leq N$, and*
- *Ω is locally Lipschitz, i.e. for any boundary point $\mathbf{x}_0 \in \partial\Omega$, there is a Euclidean coordinate system $z_1, \ldots, z_n$ and a Lipschitz function k of $n-1$ variables such that in some neighborhood of $\mathbf{x}_0$, we have*

$$\partial\Omega = \{z_n = k(z_1, \ldots, z_{n-1})\}. \tag{1.7}$$

Our main result on eigenvalue inclusion is the following.

Theorem 1.1. *Let $\Omega \subset \mathbb{R}^n$ be a domain satisfying condition 1. Then there are constants C, c depending only on Ω, such that the following holds. Let $E > 1$ and suppose u is a non-trivial solution of $(\Delta + E)u = 0$ in $C^\infty(\Omega)$, with $t[u] := \|u\|_{L^2(\partial\Omega)}/\|u\|_{L^2(\Omega)}$. Then,*

$$d(E, \sigma) \;\leq\; C\sqrt{E}\, t[u]\ , \tag{1.8}$$

and for the normalized Helmholtz solution $u_{\min}$ minimizing $t[u]$ at the given E,

$$c\sqrt{E}\, t[u_{\min}] \;\le\; d(E,\sigma) \;\le\; C\sqrt{E} t[u_{\min}]. \tag{1.9}$$

REMARK 1.1. *The estimate* (1.9) *states that* (1.8) *is sharp, i.e., using $t[u]$ alone one cannot localize the spectrum any more tightly than this, apart from optimizing the constants c and C.*

REMARK 1.2. *The existence of a minimizer for $t[u]$ follows from Lemma 3.1, in the case that E is not a Dirichlet eigenvalue (and is trivial when E is a Dirichlet eigenvalue). The lower bound on the distance to the spectrum in* (1.9) *is of use when the numerical scheme is known to produce a good approximation to $u_{\min}$.*

We will also prove the following corresponding bound on the error of the trial eigenfunction u, which improves by a factor $\sqrt{E}$ the previous best known result (Moler–Payne [24, Thm. 2]).

THEOREM 1.2. *Let Ω be as in Theorem 1.1. Then there is a constant C depending only on Ω, such that the following holds. Let $E > 1$, let E_j be the eigenvalue nearest to E, and let E_k the next nearest distinct eigenvalue. Suppose u is a solution of $(\Delta + E)u = 0$ in $C^\infty(\Omega)$ with $\|u\|_{L^2(\Omega)} = 1$, and let $\hat{u}_j$ be the projection of u onto the E_j eigenspace. Then,*

$$\|u - \hat{u}_j\|_{L^2(\Omega)} \;\le\; C\frac{\sqrt{E}\, t[u]}{|E - E_k|}\ . \tag{1.10}$$

REMARK 1.3. *The left-hand side above is equal to $\sin\theta$, where θ is the subspace angle between u and the E_j eigenspace (this viewpoint is elaborated in [9, Sec. 6]). For example, when E_j is a simple eigenvalue, we may write $\|u - \phi_j\|_{L^2(\Omega)} = 2\sin(\theta/2)$.*

REMARK 1.4. *This result is also sharp, in a certain sense: see Remark 4.1.*

To conclude the introduction, we present some key ingredients of the proofs. Define the boundary functions of the eigenmodes by

$$\psi_j(s) := \partial_n \phi_j(s) \qquad s \in \partial\Omega \tag{1.11}$$

where $\partial_n = \mathbf{n}\cdot\nabla$ is the usual normal derivative. Our main tools will be two theorems stating that boundary functions ψ_j lying close in eigenvalue are almost orthogonal. The first is the following new result which we prove in Section 2.

THEOREM 1.3 (spectral window quasi-orthogonality). *Let $\Omega \subset \mathbb{R}^n$ be a domain satisfying Condition 1. There exists a constant C_Ω depending only on Ω such that the operator norm bound*

$$\Big\| \sum_{|E_j - E| \le E^{1/2}} \psi_j \langle \psi_j, \cdot \rangle \Big\|_{L^2(\partial\Omega)\to L^2(\partial\Omega)} \;\le\; C_\Omega E \tag{1.12}$$

holds for all $E \ge 1$. (Here, $\langle\cdot,\cdot\rangle$ denotes the inner product in $L^2(\partial\Omega)$.)

REMARK 1.5. *By Weyl's Law [17, Ch. 11] there are $O(E^{(n-1)/2})$ terms in the above sum. Since each term already has norm $\ge cE$ [28, 19], the theorem expresses essentially complete mutual orthogonality, up to a constant. Only the scaling of the window width with E is important: the theorem also holds for a window $|E_j - E| \le cE^{1/2}$ for any fixed c (C_Ω will then depend on c as well as Ω). On the other hand, one*

could not expect it to hold over a spectral window of width $O(E^\beta)$ for $\beta > 1/2$, since the boundary functions are approximately band-limited to spatial wavenumber $E^{1/2}$ and thus no more than $O(E^{(n-1)/2})$ of them could be orthogonal on the boundary.

The second result is a pairwise estimate on the inner product of boundary functions lying close in eigenvalue, with respect to a special inner product: (Here, $\mathbf{x}(s)$ refers to the location of boundary point s relative to a fixed origin, which may or may not be inside Ω.)

THEOREM 1.4 (pairwise quasi-orthogonality). *Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain, and let $S := \frac{1}{2}\sup_{\mathbf{x}\in\Omega}\|\mathbf{x}\|$. Then, for all $i, j \geq 1$,*

$$\Big|\int_{\partial\Omega} (\mathbf{x}(s)\cdot\mathbf{n}(s))\,\psi_i(s)\psi_j(s)\,ds - 2E_i\delta_{ij}\Big| \;\leq\; S^2(E_i - E_j)^2 \tag{1.13}$$

REMARK 1.6. *This theorem was proved by the first-named author in [3, Appendix B]. It may be viewed as an off-diagonal generalization of a theorem of Rellich [28] which gives the $i = j$ case. The boundary weight $\mathbf{x}\cdot\mathbf{n}$ (also known as the Morawetz multiplier) is the only one known that gives quadratic growth away the diagonal yet also gives non-zero diagonal elements.*

Note that neither of the above quasi-orthogonality theorems implies the other. We also note that Bäcker et al. derived a completeness property of the boundary functions in a (smoothed) spectral window [2, Eq. (53)], that is closely related to Theorem 1.3.

After proving Theorem 1.3, we combine it with a boundary operator defined in Section 3 to prove the main theorems, in Section 4. In Section 5 we state and prove a variant of Theorem 1.1 for strictly star-shaped planar domains, which has an optimal constant C. This builds on Theorem 1.4 combined with the Cotlar-Stein lemma (see Lemma 5.2). In the main Theorems 1.1, 1.2 and 5.1, the domain-dependent constants are not explicit; we discuss their explicit values in Section 6. We present a high-accuracy numerical example using the MPS, and sketch some of the implementation aspects, in Section 7. Finally, we conclude in Section 8.

2. Quasi-orthogonality in an eigenvalue window. Here we prove Theorem 1.3 using a "TT^* argument". We need the fact that the upper bound $\|\psi_j\|^2_{L^2(\partial\Omega)} \leq CE_j$ on eigenmode normal derivatives, proved for example in [19], generalizes to quasi-modes living in an $O(E^{1/2})$ spectral window. The proof is almost the same as in [19].

LEMMA 2.1. *Let $\Omega \subset \mathbb{R}^n$ satisfy Condition 1. Let $E > 1$, and let*

$$\phi := \sum_{|E_j - E| \leq E^{1/2}} c_j\phi_j \tag{2.1}$$

with real coefficients c_j, and $\sum_j c_j^2 = \|\phi\|^2_{L^2(\Omega)} = 1$. Then,

$$\|\partial_n\phi\|^2_{L^2(\partial\Omega)} \;\leq\; C_\Omega E \tag{2.2}$$

where the constant C_Ω depends only on Ω.

Proof. To prove this we need the following lemma, proved in Appendix A, stating that for any piecewise smooth domain (in the sense of Condition 1) there is a smooth vector field that is outgoing at each boundary point.

LEMMA 2.2. *Let Ω satisfy Condition 1. Then there exists a smooth vector field $\mathbf{a}$, defined on a neighborhood of $\overline{\Omega}$, such that*

$$\mathbf{a}\cdot\mathbf{n} \geq 1 \tag{2.3}$$

almost everywhere on $\partial\Omega$.

The main tool for proving Lemma 2.1 is the identity

$$\int_{\partial\Omega}(D\phi)\partial_n\phi = -\int_\Omega \phi[\Delta, D]\phi + \int_\Omega (D\phi)(\Delta+E)\phi - \int_\Omega \phi D(\Delta+E)\phi \tag{2.4}$$

for any first order differential operator D, which follows from[2] Green's 2nd identity, the definition of the commutator, and $\phi|_{\partial\Omega} = 0$. Choosing $D := \mathbf{a}\cdot\nabla$, where $\mathbf{a}$ is as in Lemma 2.2, we notice that the left-hand side of (2.4) bounds the left-hand side of (2.2), since

$$\int_{\partial\Omega}\psi_j^2 \;\le\; \int_{\partial\Omega}(\mathbf{a}\cdot\mathbf{n})\psi_j^2 \tag{2.5}$$

by Condition 1. We may now bound each of the terms on the right-hand side of (2.4). Defining $C_a = \sup_{\mathbf{x}\in\Omega}|\mathbf{a}(\mathbf{x})|$, we have

$$\|D\phi\|^2_{L^2(\Omega)} \le C_a^2\int_\Omega \|\nabla\phi\|^2 = -C_a^2\int_\Omega \phi\Delta\phi = C_a^2\sum_j |c_j|^2 E_j \le C_a^2 F \tag{2.6}$$

where $F := E + E^{1/2}$ is the upper end of the window. Similarly,

$$\begin{aligned}\|D(\Delta+E)\phi\|^2_{L^2(\Omega)} &\le C_a^2\int_\Omega \|\nabla(\Delta+E)\phi\|^2 = C_a^2\sum_{ij} c_i c_j \int_\Omega (\Delta+E)\phi_i(-\Delta)(\Delta+E)\phi_j \\ &= C_a^2\sum_j c_j^2 E_j(E-E_j)^2 \le C_a^2 EF\ .\end{aligned} \tag{2.7}$$

Using Cauchy-Schwarz, the sum of the last two terms in (2.4) is then bounded by $2C_a\sqrt{EF}$. For the first term on the right of (2.4), we use Einstein notation $[\Delta, D] = \partial_{ii}(a_j\partial_j\cdot) - a_j\partial_{iij}$. After several steps, using integration by parts and $\phi|_{\partial\Omega} = 0$, we get

$$-\int_\Omega \phi[\Delta, D]\phi = 2\int_\Omega (\partial_i a_j)(\partial_i\phi)\partial_j\phi + \int_\Omega (\partial_{ii}a_j)\phi\partial_j\phi \tag{2.8}$$

The constants $C_a' := \sup_{\mathbf{x}\in\Omega}\|\mathbb{A}(\mathbf{x})\|_2$ where the matrix $\mathbb{A}\in\mathbb{R}^{n\times n}$ has entries $\partial_i a_j$, and $C_a'' := \sup_{\mathbf{x}\in\Omega, j=1,\dots,n}|\Delta a_j(\mathbf{x})|$, exist and are finite. Then (2.8) is bounded by $2C_a'F + C_a''F^{1/2}$. Adding all bounds on terms in (2.4) we get

$$\|\partial_n\phi\|^2_{L^2(\partial\Omega)} \le 2(C_a + C_a')F + C_a''\sqrt{F}\ , \tag{2.9}$$

which is bounded by a constant times E for $E > 1$. ∎

Proof of Theorem 1.3. Consider the coefficient vector $\mathbf{c} := \{c_j\}\in\mathbb{R}^N$ appearing in (2.1), where N is the number of eigenvalues (counting multiplicity) in the spectral window. Define the linear operator $T : \mathbb{R}^N \to L^2(\partial\Omega)$ by

$$T\mathbf{c} = \sum_j c_j\psi_j \tag{2.10}$$

Lemma 2.1 states that $\|T\|_{l^2\to L^2(\partial\Omega)} \le (C_\Omega E)^{1/2}$. Thus $\|TT^*\|_{L^2(\partial\Omega)} \le C_\Omega E$. But TT^* is the operator in the statement of Theorem 1.3, which completes its proof.

[2] The computation, involving a total of three derivatives, is justified for our class of domains, since Dirichlet eigenfunctions are in $H^{3/2}(\Omega)$ for any Lipschitz Ω; see [13], Theorem B, p164. Rellich-type computations are also justified on Lipschitz domains in [1].

3. Relating tension to a boundary operator. In this section, we show, following Barnett [4], that the tension $t[u]$ is related to the operator norm of a natural boundary operator.

For E a non-eigenvalue of Ω, let $\mathcal{K}(E) : L^2(\partial\Omega) \to L^2(\Omega)$ be the solution operator (Poisson kernel) for the interior Dirichlet boundary-value problem,

$$(\Delta + E)u = 0 \qquad \text{in } \Omega \tag{3.1}$$

$$u = f \qquad \text{on } \partial\Omega \ , \tag{3.2}$$

that is, $u = \mathcal{K}f$. (For existence and uniqueness for L^2 data on a Lipschitz boundary see for example [23, Thm. 4.25].) Since the eigenbasis is complete in $L^2(\Omega)$, we may write $u = \sum_{j=1}^{\infty} c_j \phi_j$. We evaluate each c_j by applying Green's 2nd identity,

$$(E - E_j)(\phi_j, u)_{L^2(\Omega)} = \int_\Omega (u\Delta\phi_j - \phi_j \Delta u) = \int_{\partial\Omega} (f\psi_j - \phi_j \partial_n u) ds \ , \tag{3.3}$$

thus $c_j = \langle \psi_j, f\rangle/(E - E_j)$. The solution operator may therefore be written as a sum of rank-1 operators,

$$\mathcal{K}(E) = \sum_{j=1}^{\infty} \frac{\phi_j \langle \psi_j, \cdot\rangle}{E - E_j} \ . \tag{3.4}$$

By the definition (1.4) we have, now for any u satisfying $(\Delta + E)u = 0$ in Ω, that $t[u]^{-1} \le \|\mathcal{K}(E)\|$. Since $\|\mathcal{K}^*\mathcal{K}\| = \|\mathcal{K}\|^2$, then by defining the boundary operator in $L^2(\partial\Omega) \to L^2(\partial\Omega)$,

$$A(E) := \mathcal{K}(E)^*\mathcal{K}(E) \ , \tag{3.5}$$

we have an estimate on the tension that will be the main tool in our analysis,

$$t[u]^{-2} \ \le \ \|A(E)\| \ . \tag{3.6}$$

Inserting (3.4) into (3.5) and using orthogonality (or see [4, Sec. 3.1]), we have that A also may be written as the sum of rank-1 operators,

$$A(E) = \sum_{j=1}^{\infty} \frac{\psi_j \langle \psi_j, \cdot\rangle}{(E - E_j)^2} \tag{3.7}$$

This sum is conditionally convergent: the sum of the operator norm of each term diverges. For instance, for $n = 2$, Weyl's law [17, Ch. 11] states that the density of eigenvalues E_j is asymptotically constant, but since $\|\psi_j\|^2 = \Omega(E_j)$ the sum of norms is logarithmically divergent; for $n > 2$ the divergence is worse. Despite this, we have the following, which improves upon the results of [4].

LEMMA 3.1. *Let $\Omega \subset \mathbb{R}^n$, $n \ge 2$, satisfy Condition 1, and let $E > 0$. Then*

$$\lim_{N\to\infty} \ \sum_{j=1}^{N} \frac{\psi_j \langle \psi_j, \cdot\rangle}{(E - E_j)^2} \tag{3.8}$$

converges in the norm operator topology. Furthermore, the limit operator $A(E)$ is compact in $L^2(\partial\Omega)$.

Proof. This follows immediately from (4.8) in the proof of Lemma 4.2 below, which shows that the tail of the sum in (3.7) has vanishing operator norm. A is therefore also the norm limit of a sequence of finite-rank operators. □

4. Proof of Theorems 1.1 and 1.2. In the previous section we related tension to the norm of a boundary operator which itself can be written as a sum involving mode boundary functions. Here we place upper bounds on $\|A(E)\|$ in order to prove Theorems 1.1 and 1.2. Firstly we note that when E is an eigenvalue, Theorem 1.1 is trivially satisfied, since $t[u_{\min}] = 0$. When E is a non-eigenvalue, formula (3.7) enables us to split up contributions from different parts of the Dirichlet spectrum,

$$A(E) = A_{\text{near}}(E) + A_{\text{far}}(E) + A_{\text{tail}}(E) \tag{4.1}$$

where

$$A_{\text{near}}(E) = \sum_{|E_j - E| \le E^{1/2}} \frac{\psi_j \langle \psi_j, \cdot \rangle}{(E - E_j)^2} \tag{4.2}$$

$$A_{\text{far}}(E) = \sum_{E/2 \le E_j \le 2E,\, |E_j - E| > E^{1/2}} \frac{\psi_j \langle \psi_j, \cdot \rangle}{(E - E_j)^2} \tag{4.3}$$

$$A_{\text{tail}}(E) = \sum_{E_j < E/2} \frac{\psi_j \langle \psi_j, \cdot \rangle}{(E - E_j)^2} + \sum_{E_j > 2E} \frac{\psi_j \langle \psi_j, \cdot \rangle}{(E - E_j)^2} \tag{4.4}$$

It is sufficient (due to the operator triangle inequality) to bound the norms of these three terms independently. We first tackle the "far" and "tail" terms.

LEMMA 4.1. *There is a constant C dependent only on Ω such that*

$$\left\| A_{\text{far}}(E) \right\| \;\le\; C \qquad \textit{for all } E > 1 \tag{4.5}$$

Proof. For any $E > 1$, consider the spectral interval $I_m := [E + mE^{1/2}, E + (m+1)E^{1/2}]$. For any such interval lying in $[E/2, 2E]$ we may apply Theorem 1.3, with E replaced by at most $2E$, to bound $\| \sum_{E_j \in I_m} \psi_j \langle \psi_j, \rangle \|$ by $2C_\Omega E$. For terms in (4.3) associated with this interval, the denominators are no less than $m^2 E$. Thus

$$\Big\| \sum_{E_j \in I_m} \frac{\psi_j \langle \psi_j, \cdot \rangle}{(E - E_j)^2} \Big\| \;\le\; \frac{2C_\Omega}{m^2} \tag{4.6}$$

Covering $[E + E^{1/2}, 2E]$ by summing over $m = 1, 2, \ldots$ gives a constant, since $\sum m^{-2} = \pi^2/6$. The same argument applies for intervals covering $[E/2, E - E^{1/2}]$. ∎

LEMMA 4.2. *There is a constant C dependent only on Ω such that*

$$\left\| A_{\text{tail}}(E) \right\| \;\le\; C E^{-1/2} \qquad \textit{for all } E > 1 \tag{4.7}$$

Proof. Consider a spectral interval $I_m := [2^m E, 2^{m+1} E]$. We may cover this with at most $2^{m/2-1} E^{1/2} + 1$ windows of half-width at most $2^{m/2} E^{1/2}$; for each of these windows Theorem 1.3 applies to bound $\| \sum_{E_j \in I_m} \psi_j \langle \psi_j, \rangle \|$ by $C_\Omega 2^{m+1} E$. For each $E_j \in I_m$, the denominator is no smaller than $(2^{m-1} E)^2$. Thus

$$\Big\| \sum_{E_j \in I_m} \frac{\psi_j \langle \psi_j, \cdot \rangle}{(E - E_j)^2} \Big\| \;\le\; (2^{m/2-1} E^{1/2} + 1) \frac{C_\Omega 2^{m+1} E}{(2^{m-1} E)^2} = C_\Omega (2^{-m/2-2} E^{-1/2} + 2^{-m+1} E^{-1}) \tag{4.8}$$

The infinite sum over $m = 1, 2, \ldots$ gives

$$\Big\| \sum_{E_j > 2E} \frac{\psi_j \langle \psi_j, \cdot \rangle}{(E - E_j)^2} \Big\| \le C_\Omega \left(\frac{E^{-1/2}}{4(\sqrt{2}-1)} + 2E^{-1} \right) \le C E^{-1/2} \quad \text{for all } E > 1. \tag{4.9}$$

We treat the interval $(0, E/2)$ similarly, using a sequence of intervals $J_m := [2^{-m-1}E, 2^{-m}E]$. Each such interval may be covered by at most $2^{-(m+3)/2}E^{1/2} + 1$ windows of half-width $2^{-(m+1)/2}E^{1/2}$. For each $E_j \in J_m$, the denominator is no smaller than $E^2/4$. In a similar manner as before, the operator norm of the partial sum associated with J_m is then $O(2^{-m}E^{-1/2})$, thus the infinite sum over m is $O(E^{-1/2})$. Note that Theorem 1.3 does not apply for $E < 1$, but that there are $O(1)$ such E_j values and each contributes $O(E^{-1})$. This proves the Lemma. □

Proof of Theorem 1.1. Examining the "near" term (4.2), we use Theorem 1.3 on the sum of numerators, and get a bound by taking the minimum denominator,

$$\big\| A_{\text{near}}(E) \big\| \;\le\; \frac{C_\Omega E}{d(E,\sigma)^2} \qquad \text{for all } E > 1 \tag{4.10}$$

Using this and the above Lemmas to sum the terms in (4.1) gives

$$\big\| A(E) \big\| \;\le\; \frac{C_\Omega E}{d(E,\sigma)^2} + C \qquad \text{for all } E > 1 \tag{4.11}$$

From Lemma B.1, an upper bound on the distance to the spectrum, we see that the second term is bounded by at most a constant times the first, so may be absorbed into it to give

$$\big\| A(E) \big\| \;\le\; \frac{CE}{d(E,\sigma)^2} \qquad \text{for all } E > 1 \tag{4.12}$$

Combining this with (3.6) proves (1.8), hence also the second inequality in (1.9). The first inequality in (1.9) simply follows from the fact that, since A is a sum of positive operators,

$$t[u_{\min}]^{-2} = \big\| A(E) \big\| \ge \Big\| \frac{\psi_j \langle \psi_j, \cdot \rangle}{(E - E_j)^2} \Big\| = \frac{\|\psi_j\|^2}{d(E,\sigma)^2} \ , \tag{4.13}$$

where E_j is the eigenvalue closest to E. Using the lower bound $\|\psi_j\|^2 \ge cE_j$ from [19] this becomes

$$d(E,\sigma) \ge c\sqrt{E_j}\, t[u_{\min}] \ . \tag{4.14}$$

With a change of constant, E_j may be replaced here by E to give the first inequality in (1.9), since Lemma B.1 insures that E_j is relatively close to E. (The lemma is not useful for E less than some constant and $E_j < E$, but then the ratio E/E_j is still bounded by a constant because $E_j \ge E_1$).

Proof of Theorem 1.2. We next prove the eigenfunction error bound (1.10), first considering E a non-eigenvalue. We denote the boundary data by $U := u|_{\partial\Omega}$. From orthogonality, then using the formula for the c_i coefficients below (3.3), we get,

$$\|u - \hat{u}_j\|^2_{L^2(\Omega)} = \sum_{E_i \ne E_j} |(\phi_i, u)_{L^2(\Omega)}|^2 = \sum_{E_i \ne E_j} \frac{|\langle \psi_i, U \rangle|^2}{(E - E_i)^2} \le \Big\| \sum_{E_i \ne E_j} \frac{\psi_i \langle \psi_i, \cdot \rangle}{(E - E_i)^2} \Big\| \|U\|_2^2. \tag{4.15}$$

The operator in the last expression is identical to (3.7) except with the E_j-eigenspace terms omitted. Therefore, its norm may be bounded in the same way as that of $A(E)$, the only difference being that the $d(E,\sigma)$ introduced in (4.10) is replaced by $\min_{E_i \neq E_j} |E - E_i| = |E - E_k|$. Thus the bound analogous to (4.12) is

$$\| \sum_{E_i \neq E_j} \frac{\psi_j \langle \psi_j, \cdot \rangle}{(E - E_j)^2} \| \ \leq \ \frac{CE}{(E - E_k)^2} \qquad \text{for all } E > 1 \ ,$$

and inserting this and $\|U\|_{L^2(\partial\Omega)} = t[u]\|u\|_{L^2(\Omega)} = t[u]$ into (4.15) gives (1.10). Finally, if E is an eigenvalue, i.e. $E = E_j$, the solution operator (3.4) is undefined, since a solution to (3.1)-(3.2) exists if and only if f is orthogonal to the normal derivative functions in the E-eigenspace. This can be seen by applying Green's 2nd identity to ϕ, any function in the E-eigenspace, and u, giving $\langle \partial_n \phi, U \rangle = 0$. However, the solution coefficients c_i for which $E_i \neq E$ are uniquely defined by the same formula as before. Thus (4.15) and the rest of the proof carries through.

Remark 4.1. *Theorem 1.2 is sharp, as can be seen in the following way: if u is such that $t[u]$ is close to $t[u_{\min}]$ (say, less than $2t[u_{\min}]$), then we have, by combining* (1.9) *and* (1.10),

$$\|u - \hat{u}_j\|_{L^2(\Omega)} \ \leq \ C \frac{|E - E_j|}{|E - E_k|} \ . \tag{4.16}$$

Apart from the value of the constant, one cannot expect to do better than this. For example, if E is midway between E_j and E_k, then the error $\|u - \hat{u}_j\|_{L^2(\Omega)}$ cannot be expected to be better than $1/\sqrt{2}$.

5. Star-shaped planar domains. The purpose of this section is to say something stronger than Theorem 1.1 in the special case of star-shaped domains in $n = 2$. We take weighted boundary functions

$$\psi_j^{(s)}(s) := (\mathbf{x}(s) \cdot \mathbf{n}(s)) \partial_n \phi_j(s) \ , \qquad s \in \partial\Omega \tag{5.1}$$

and our boundary inner product as

$$\langle f, g \rangle_s := \int_{\partial\Omega} (\mathbf{x}(s) \cdot \mathbf{n}(s))^{-1} f(s) g(s) ds \tag{5.2}$$

hence norm $\|f\|_s := \sqrt{\langle f, f \rangle}$, and $t_s[u] := \|U\|_s / \|u\|_{L^2(\Omega)}$. The significance of the weight $(\mathbf{x} \cdot \mathbf{n})$ is twofold: it is strictly positive for strictly star-shaped domains, and also turns the inner product in (1.13) into $\langle \psi_i^{(s)}, \psi_j^{(s)} \rangle_s$, enabling us to benefit from pairwise quasi-orthogonality. The Rellich theorem $\|\psi_j^{(s)}\|_s^2 = 2E_j$ states that, with this special weight, there is no fluctuation in the L^2-norms of the boundary functions. As shown in [4], the function $t_s[u_{\min}]$ vs E has slope $1/\|\psi_j^{(s)}\|_s^2$ in the neighborhood of E_j (this arises from dominance of a single term in (5.5) below). Hence these slopes are predictable *independently* of the particular form of each mode ϕ_j. This enables us to get the following eigenvalue inclusion result analogous to Theorem 1.1.

Theorem 5.1. *Let $\Omega \subset \mathbb{R}^2$ be a strictly star-shaped bounded domain with piecewise smooth boundary. Then there are constants c_1, c_2, c_3 depending only on Ω, such that the following holds. Let $E > 1$, and suppose u is a non-trivial solution of $(\Delta + E)u = 0$ in $C^\infty(\Omega)$, with $c_2 t_s[u]^2 < 1$. Let $F := E + \sqrt{E}$. Then,*

$$d(E, \sigma) \ \leq \ \sqrt{2F}\, t_s[u] \frac{1 + c_1 \sqrt{F} t_s[u]}{1 - c_2 t_s[u]^2} \ . \tag{5.3}$$

For the Helmholtz solution $u_{\min}$ minimizing $t_s[u]$ at the given E,

$$\sqrt{2(E - c_3 E^{1/2})}\, t_s[u_{\min}] \;\le\; d(E,\sigma) \tag{5.4}$$

Remark 5.1. *In the limit of high frequency $E \gg 1$ and small tension $t_s[u] \ll E^{-1/2}$, the right-hand side of* (5.3) *and the left hand side of* (5.4) *are both $\sqrt{2E}(1 + o(1))t_s$. This proves that both the power of E and the constant $\sqrt{2}$ are sharp.*

Remark 5.2. *Notice that this theorem is not applicable for all E since there may be large spectral gaps where $c_2 t_s[u]^2 < 1$ cannot be satisfied. Due to the numerator, it becomes far from optimal when $t_s[u]$ is $O(E^{-1/2})$ or larger. In these respects it is less general than Theorem 1.1, even though it gives better bounds in the small tension limit.*

The main tool used in the proof of Theorem 5.1 is the pairwise quasi-orthogonality result, Theorem 1.4, together with the Cotlar-Stein lemma, which we state here for the special case of self-adjoint operators:

Lemma 5.2 (Cotlar-Stein [12, 32, 11]). *Let $\{T_j\}_{j\in J}$ be a countable set of bounded self-adjoint operators, $J \subset \mathbb{N}$. Then*

$$\Big\|\sum_{j\in J} T_j\Big\| \le \max_{j\in J} \sum_{i\in J} \sqrt{\|T_i T_j\|}\ .$$

Proof of Theorem 5.1. The weighted equivalent of (3.7) is the operator

$$A^{(s)}(E) = \sum_{j=1}^{\infty} \frac{\psi_j^{(s)} \langle \psi_j^{(s)}, \cdot\rangle_s}{(E - E_j)^2} \tag{5.5}$$

which, by analogy with (3.6), satisfies

$$t_s[u]^{-2} \;\le\; \|A^{(s)}(E)\|_s\ . \tag{5.6}$$

The lower bound (5.4) follows by analogy with (4.13)-(4.14), using $\|\psi_j^{(s)}\|_s^2 = 2E_j$, and $E_j \ge E - c_3 E^{1/2}$ from Lemma B.1.

Using the same splitting into "near", "far", and "tail" parts as in Section 4, we can bound the norm of the "near" part in a new way, as follows.

Lemma 5.3. *There is a constant $c_1 > 0$ depending only on Ω such that*

$$\|A^{(s)}_{\text{near}}(E)\|_s \le \frac{2F}{d(E,\sigma)^2} + \frac{\sqrt{2}c_1 F}{d(E,\sigma)} \qquad \textit{for all } E > 1$$

The first term in this bound will arise simply from the single term in the sum (5.5) with E_j nearest to E. The second term requires more work, as we now show.

Proof. Let $J = \{j : |E_j - E| \le E^{1/2}\}$. Using $T_j = \frac{\psi_j \langle \psi_j, \cdot\rangle}{(E-E_j)^2}$ in Lemma 5.2 gives

$$\Big\|\sum_{j\in J} T_j\Big\|_s \;\le\; \max_{j\in J} \frac{\|\psi_j^{(s)}\|_s^{1/2}}{|E - E_j|} \sum_{i\in J} \frac{\big(\langle \psi_i^{(s)}, \psi_j^{(s)}\rangle_s \, \|\psi_i^{(s)}\|_s\big)^{1/2}}{|E - E_i|} \tag{5.7}$$

Applying quasi-orthogonality (Theorem 1.4) for the inner product, and $\|\psi_j^{(s)}\|_s^2 = 2E_j$, and separating diagonal ($i = j$) from off-diagonal terms, we get,

$$\Big\|\sum_{j\in J} \frac{\psi_j \langle \psi_j, \cdot\rangle}{(E - E_j)^2}\Big\| \;\le\; \max_{j\in J} \frac{2E_j}{(E - E_j)^2} + \sqrt{2}S \max_{j\in J} \frac{E_j^{1/4}}{|E - E_j|} \sum_{i\in J} \frac{E_i^{1/4}|E_i - E_j|}{|E - E_i|} \tag{5.8}$$

Here S is as in Theorem 1.4. The first term is bounded by $2F/d(E,\sigma)^2$. Using $|E_i - E_j| \le |E_i - E| + |E - E_j|$ bounds the second term by

$$\begin{aligned} &\sqrt{2}S \max_{j\in J} \frac{E_j^{1/4}}{|E-E_j|} \sum_{i\in J} E_i^{1/4} \left(1 + \frac{|E-E_j|}{|E-E_i|}\right) \\ &\le \sqrt{2F}S\,|J| \max_{j\in J}\Big(\frac{1}{|E-E_j|} + \frac{1}{d(E,\sigma)}\Big) \;\le\; \frac{2\sqrt{2F}S|J|}{d(E,\sigma)} \end{aligned} \tag{5.9}$$

Recall Weyl's law for the asymptotic density of eigenvalues, which states that, for $n = 2$ and $\operatorname{vol}\Omega = 1$,

$$N(E) := \#\{j : E_j < E\} = \frac{1}{4\pi}E + R(E)\ , \tag{5.10}$$

where the remainder is $R(E) = O(\sqrt{E})$ ([29]; for the case of piecewise-smooth boundary see [31, Eq. (0.3)]). Since the remainder is bounded for small E, there is a constant $C_{\rm w}$ such that $|R(E)| \le C_{\rm w}\sqrt{E}$ for all $E > 1$. Thus $|J|$, the number of terms in the "near" window, is bounded by

$$|J| \le \Big(\frac{1}{4\pi} + 2C_{\rm w}\Big)\sqrt{F}\ .$$

Inserting this into (5.9) proves the Lemma, and we may take $c_1 = 2S(1/4\pi + 2C_{\rm w})$. ∎

Completion of the proof of Theorem 5.1. The proofs of analogously weighted versions of Lemmas 4.1 and 4.2 are unchanged. So we may combine them with Lemma 5.3 and (5.6) to get, for some constant c_2,

$$t_s[u]^{-2} \;\le\; \|A^{(s)}(E)\|_s \;\le\; \frac{2F}{d(E,\sigma)^2} + \frac{\sqrt{2}c_1F}{d(E,\sigma)} + c_2 \qquad \text{for all } E > 1$$

Multiplying through by $d(E,\sigma)^2$ we solve the quadratic inequality for $d(E,\sigma)$,

$$d(E,\sigma) \;\le\; \frac{c_1F/\sqrt{2} + \sqrt{c_1^2F^2/2 + 2(t_s[u]^{-2} - c_2)F}}{t_s[u]^{-2} - c_2}$$

Using the subadditivity of the square-root completes the proof of (5.3).

6. Discussion of explicit constants. For the practical application of Theorems 1.1 and 1.2, it is important to have an explicit value for the constant C (from the discussion after (4.15) we notice that C in the two theorems is the same.)

We now compute an explicit value of this C that holds for all $E > 1$. Examining (2.9) we see that a choice of constant in Lemma 2.1, and hence Theorem 1.3, that holds for all $E > 1$ is $C_\Omega = 4(C_a + C_a') + \sqrt{2}C_a''$. To compute this we need sup norms of the value, and first and second derivative, of a vector field $\mathbf{a}$ as in Lemma 2.2.

The proof of Lemma 2.2 shows such a construction; the values will depend on the size of the vectors $\mathbf{a}_{x_i}$ and the choice of partition of unity used to cover $\overline{\Omega}$. The vectors $\mathbf{a}_{x_i}$ will be large (order $1/\epsilon$) if Ω has corners with angles less than ϵ or greater than $2\pi - \epsilon$. We note that a numerical procedure for this construction could be useful.

In some special cases, a simpler prescription for the vector field can be given:

- For strictly star-shaped domains in $\mathbb{R}^n$, we may choose $\mathbf{a} = \mathbf{x}/\inf_{\partial\Omega}(\mathbf{x}\cdot\mathbf{n})$, which gives $C_a = \sup_{\partial\Omega}(\mathbf{x}\cdot\mathbf{n})/\inf_{\partial\Omega}(\mathbf{x}\cdot\mathbf{n})$, $C_a' = 1/\inf_{\partial\Omega}(\mathbf{x}\cdot\mathbf{n})$, and $C_a'' = 0$.

- For a domain with C^2 boundary, let $\delta > 0$ be the largest number such that for each $\mathbf{x}_0 \in \partial\Omega$, a ball of radius δ can be placed within Ω so as to be tangent to $\partial\Omega$ at $\mathbf{x}_0$. We may then choose $\mathbf{a} = (1 - r/\delta)^2 \mathbf{n}_r$, for $r < \delta$, $\mathbf{a} = \mathbf{0}$ otherwise, where the coordinate r is the distance from $\partial\Omega$, and $\mathbf{n}_r$ is the unit vector in the local decreasing r direction. This gives constants $C_a = 1$ and $C'_a = 2/\delta$. C''_a depends on δ and an upper bound on the rate of change of surface curvature. (Also note that a slight modification of the proof of Theorem 1.3 would allow estimation purely in terms of C_a and C'_a, but with a doubling of the numerical constants).

Summing the terms (4.6) above and below E we have that the constant in Lemma 4.1 is $2\pi^2 C_\Omega/3$. Similarly, using (4.9) and its equivalent for $(0, E/2)$ gives the constant in Lemma 4.2 as $C_\Omega(\frac{1}{4(\sqrt{2}-1)} + \frac{1}{4-\sqrt{2}} + 6) < 7C_\Omega$. Summing these two constants gives a constant C in (4.11) as $14C_\Omega$. A choice of constant in (4.12) is then $C_\Omega + 14C_\Omega \max[E_1^2, C_d^2]$, where from Appendix B we have $C_d = 2\sqrt{E_1}$, and the max accounts for the case $1 < E \leq E_1$. Finally, the constant in (1.8) is the square-root of this, $C = \sqrt{C_\Omega(1 + 14\max[E_1, 4]E_1)}$.

Requiring that the above estimates hold for all $E > 1$ caused non-optimality in the choice of constant. It is more sensible in high frequency applications to use a better constant which is approached for $E \gg 1$, and small tension $t \ll 1$. We now give this explicitly. In this limit, in (2.9), F tends to E, and we drop lower-order terms to get $C_\Omega = 2(C_a + C'_a)$, which in the star-shaped case is

$$C_\Omega = 2\frac{1 + \sup_{\partial\Omega}(\mathbf{x}\cdot\mathbf{n})}{\inf_{\partial\Omega}(\mathbf{x}\cdot\mathbf{n})} \qquad \text{for } E \gg 1,\ \Omega \text{ star-shaped}\ . \tag{6.1}$$

If tension is small (i.e. E is not in a large spectral gap), the second term in (4.12) is negligible, so we may approximate the constant in (1.8) as

$$C = \sqrt{C_\Omega} \qquad \text{for } E \gg 1,\ t \ll 1\ . \tag{6.2}$$

Remark 6.1. *The limiting constant* (6.2) *does not reflect the limiting slopes of the graph* $t[u_{\min}]$ *vs* E *near eigenvalues. These slopes are known [4] to be* $1/\|\psi_j\|^2$*, which is bounded by* $(2C'_a E_j)^{-1}$ *[19].*

We end by discussing the constants c_1 and c_2 in Theorem 5.1. Constant c_2 may be estimated easily, as above, using the weighted versions of Lemmas 4.1 and 4.2. In the proof of Lemma 5.3, c_1 involves the Weyl constant C_{W}; we know of no explicit estimates for C_{W} in the literature (the closest we know are estimates of the form $|R(E)| < C\sqrt{E}\ln E$ with explicit constants [26, 10]). However, these constants are effectively irrelevant for practical purposes, when $E \gg 1$ and $t \ll E^{-1/2}$, since in these limits, one may replace (5.3) by $d(E, \sigma) \leq \sqrt{2E}t_s[u]$ and still have an error bound very close to that given by the full expression.

7. Numerical example. In Fig. 1.1c we show a planar nonconvex domain given by the radial function $r(\theta) = 1 + 0.3\cos[3(\theta + 0.2\sin\theta)]$. The domain is star-shaped and smooth (we will not address numerical issues raised by corners here; see [16, 14, 9, 15, 3, 7].) For high-frequency eigenvalue problems, a convenient computational basis of Helmholtz solutions are 'method of fundamental solutions' basis functions $\xi_n(\mathbf{x}) = Y_0(\sqrt{E}|\mathbf{x} - \mathbf{y}_n|)$, where Y_0 is the irregular Bessel function of order zero, and $\{\mathbf{y}_n\}_{n=1}^N$ are a set of 'charge points' in $\mathbb{R}^2 \backslash \overline{\Omega}$. The latter were chosen by a displacement

of the boundary parametrization $\mathbf{x}(\theta)$, $0 < \theta \leq 2\pi$, in the imaginary direction (see [6]); specifically $\mathbf{y}_n = \mathbf{x}(2\pi n/N - 0.025i)$.

We compute the data plotted in Fig. 1.1a, b as follows. At each E, $t[\tilde{u}_{\min}]$ is given by the square-root of the smallest generalized eigenvalue of a generalized eigenvalue problem (GEVP) involving $N \times N$ symmetric real dense matrices F and G (the basis representations of the boundary and interior norms respectively.) Both matrices are evaluated using M-point periodic trapezoidal quadrature in θ, that is, quadrature points $\mathbf{x}_m = \mathbf{x}(2\pi m/M)$, $m = 1, \ldots, M$, and weights $w_m = 2\pi|\mathbf{x}'(2\pi m/M)|/M$. For instance, $F = P^*P$, where $P \in \mathbb{R}^{M\times N}$ has elements

$$P_{mn} = \sqrt{w_m}\xi_n(\mathbf{x}_m) \ , \tag{7.1}$$

and G is similarly found [4, Sec. 4.1] using P and the matrices $P^{(1)}$ and $P^{(2)}$ whose entries are the x_1- and x_2-derivatives of those in P. Since the GEVP is numerically singular, regularization was first performed, similarly to [8, Sec. 6], by restricting to an orthonormal basis for the numerical column space of $[P; P^{(1)}; P^{(2)}]$ given by the left singular vectors with singular values at least 10^{-14} times the largest singular value.

For low frequencies (Fig. 1.1a), 8-digit accuracy requires $N = 100$ basis functions and $M = 200$ quadrature points. For higher frequencies corresponding to 40 wavelengths across the domain (Fig. 1.1b, c), it requires $N = 400$ and $M = 500$, and the above GEVP procedure takes 3 seconds per E value.[3] Very small ($< 10^{-8}$) tensions cannot be found this way, and instead are best approximated via the GSVD [8]: the optimal tension at a given E is the lowest generalized singular value of the matrix pair (P, Q), where matrix Q has entries

$$Q_{mn} = \sqrt{\frac{\mathbf{x}_m \cdot \mathbf{n}_m}{2E}}\sqrt{w_m}\,\frac{\partial \xi_n}{\partial n}(\mathbf{x}_m) \ , \tag{7.2}$$

where $\mathbf{n}_m$ is the normal at $\mathbf{x}_m$, and regularization as before. Note that $G = Q^*Q$ well approximates the interior norm in the subspace with zero Dirichlet data, due to the Rellich formula (case $i = j$ of Theorem 1.4).

Any single-variable function minimization algorithm may then be used to search for a local minimum of $t[\tilde{u}_{\min}]$ vs E; we prefer iterated fitting of a parabola to $t[\tilde{u}_{\min}]^2$ at three nearby E values, which converges typically in 5 iterations. Using this with the GSVD (with $N = 500$, $M = 700$, i.e. 6 points per wavelength on $\partial\Omega$, and taking 8 seconds per iteration), we find the tension

$$t[\tilde{u}_{\min}] = 2.2 \times 10^{-12} \qquad \text{at } E = 10005.0213579739 \ . \tag{7.3}$$

This is shown by the dot in Fig. 1.1b. The GSVD right singular vector gives the basis coefficients of the corresponding trial function $\tilde{u}_{\min}$, which is plotted in Fig. 1.1c (this took 34 seconds to evaluate on a square grid of size 0.005, i.e. 1.3×10^5 points.)

Armed with datapoint (7.3), what can we deduce about Dirichlet eigenpairs of Ω using our new theorems, and how much better are they than previous results? The constant in the Moler–Payne bound (1.5) is $C_{\text{MP}} = q_1^{-1/2}$ where q_1 is the lowest eigenvalue of a Stekloff eigenproblem on Ω [20, (2.11)]. Since Ω is star-shaped, the bound $q_1 \geq E_1^{1/2} \inf_{\partial\Omega}(\mathbf{x}\cdot\mathbf{n})/2\sup_{\partial\Omega}(\mathbf{x}\cdot\mathbf{n})$ from [20, Table I] applies, giving $C_{\text{MP}} = 1.31$ as a valid choice. Thus (1.5) states that there is an eigenvalue E_j a distance no

[3] All computation times are reported for a laptop with 2GHz Intel Core Duo processor and 2GB RAM, running MATLAB 2008a on a linux kernel.

more than 2.9×10^{-8} from the above E. On the other hand, (6.2) and (6.1) give the constant in Theorem 1.1 as $C = 2.9$. Applying the theorem gives a distance from the spectrum of no more than 6.3×10^{-10}. Taking the small-tension limit of the star-shaped planar result (5.3), and recomputing the weighted tension $t_s[\tilde{u}_{\min}]$ from Section 5, we get an even smaller distance of 3.5×10^{-10}, that is, an error of ± 3 in the last digit of (7.3). The latter is a 80-fold improvement over Moler–Payne. (Also see [4] for an example at higher frequency with 3 digits of improvement).

How good an approximation is $\tilde{u}_{\min}$ to the eigenfunction ϕ_j? Using the observation that the next nearest eigenvalue is $E_k = 10007.339\cdots$, the eigenfunction bound of Moler–Payne [24] gives an L^2-error of 1.2×10^{-8}. With the same data, using the constant C above, Theorem 1.2 gives an L^2-error of 2.7×10^{-10}, a 50-fold improvement over that achievable with previously known theorems.

8. Conclusions. We have improved, by a factor of the wavenumber, the Moler–Payne bounds on Dirichlet eigenvalues and eigenfunctions which have been the standard for the last 40 years. This makes rigorous the conjectures based on numerical observations in [4]. We expect this to be useful since high-frequency wave and eigenvalue calculations are finding more applications in recent years. Of independent interest is a new quasi-orthogonality result in a spectral window (Theorem 1.3).

For numerical utility, throughout we have been explicit with constants, and have specified a lower bound on E for which the estimates hold (this being stronger than merely a 'big-O' asymptotic estimate). For star-shaped domains we strengthened the inclusion bounds (Theorem 5.1), achieving a sharp power of E and sharp constant, in the limit of small tension, when tension is weighted by a special geometric function. This weight allowed pairwise quasi-orthogonality to be used, but since an upper bound for the number of eigenvalues in a $\sqrt{E}$ window is needed, this is only useful for $n = 2$ (planar domains).

We applied our theorems to a numerical example, enabling close to 14 digits accuracy in a high-lying eigenvalue, and 10 digits in the eigenfunction. Both are two digits beyond what could be claimed with previously-known theorems.

Our estimate $C\sqrt{E}t[u]$ on the distance to the spectrum is sharp (up to constants) if the tension $t[u]$ (or $t_s[u]$) is comparable to $t[u_{\min}]$ ($t_s[u_{\min}]$). However, numerically, one generally has access to other properties of u (e.g. its normal derivative) which can give more detailed information about the spectrum. For example, the powerful 'scaling method' [36, 3] is able to locate many eigenvalues using an operator computed at a single energy E. In another direction, Still [33, Thm. 4] obtains improved inclusion bounds when the approximate eigenvalue is equal to the Rayleigh quotient; in this case, the bound is proportional to $t[u]^2$, but scaling as E^2 for large energy.

An open problem with practical benefits is the generalization of these results to Neumann and Robin boundary conditions, and to multiple subdomains with different trial functions on each subdomain and least-square errors on artificial boundaries [14, 15, 7] (these are known as Trefftz or non-polynomial finite element methods).

Acknowledgments. The authors are grateful for discussions with Dana Williams, Timo Betcke, and Chen Hua. The work of AHB was supported by NSF grant DMS-0811005, and a Visiting Fellowship to ANU in February 2009 as part of the *ANU 2009 Special Year on Spectral Theory and Operator Theory*. The work of AH was supported by Australian Research Council Discovery Grant DP0771826.

Appendix A. Proof of Lemma 2.2.

Proof. For every boundary point x, we can find a constant vector field $\mathbf{a}_x$ having the property (2.3) in a neighbourhood U_x of x. (Take a multiple of the vector field ∂_{z_n} in the Euclidean coordinate system used in (1.7).) By compactness we can find a finite number of such neighbourhoods $U_i = U_{x_i}$, $i = 1, \ldots, N$ covering $\partial\Omega$. We can add to this collection of open sets one additional set U_0, whose closure does not meet $\partial\Omega$, yielding an open cover of $\overline{\Omega}$. Let ϕ_i, $i = 0 \ldots N$ be a smooth partition of unity subordinate to this open cover. Then

$$\mathbf{a} = \sum_{i=1}^{N} \phi_i \mathbf{a}_{x_i}$$

is a vector field with the required property. □

Appendix B. Upper bound on distance to spectrum.

LEMMA B.1. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain. Let E_1 be the lowest Dirichlet eigenvalue of Ω. Then for any $E > E_1$,*

$$d(E, \sigma) \ \leq \ C_d E^{1/2} \ , \qquad C_d = 2\sqrt{E_1}.$$

REMARK B.1. *The result becomes interesting only for $E > 2(1+\sqrt{2})E_1$. Bounds on E_1 exist as follows. If Ω contains a Euclidean ball of radius r, then E_1 is less than or equal to $E_1(B(0,r))$, which is equal to $j^2_{n/2-1,1}/r^2$, where $j_{n/2-1,1}$ is the first positive zero of the Bessel function $j_{n/2-1}$. For $n = 2$ we have $j_{0,1} = 2.4048...$ and for $n = 3$ we have $j_{1/2,1} = \pi$. Also, E_1 is greater than or equal to the first eigenvalue of the ball having the same n-volume as Ω, by the Faber-Krahn inequality [27].*

Proof. Choose a wavevector $\mathbf{k} \in \mathbb{R}^n$ with $|\mathbf{k}|^2 = E - E_1$, and consider the trial function $u : \Omega \to \mathbb{C}$ defined by $u(\mathbf{x}) := \phi_1(\mathbf{x})e^{i\mathbf{k}\cdot\mathbf{x}}$, where ϕ_1 is the normalized first Dirichlet eigenmode of Ω with eigenvalue E_1. We calculate,

$$(\Delta + E)u = 2i\mathbf{k} \cdot \nabla\phi_1 e^{i\mathbf{k}\cdot\mathbf{x}} \ .$$

Since u is in the domain of Ω, has norm $\|u\|_{L^2(\Omega)} = 1$, and

$$\|(\Delta + E)u\|^2 = 4\int_\Omega |\mathbf{k} \cdot \nabla\phi_1|^2 d\mathbf{x} \leq 4(E - E_1)E_1 < 4EE_1, \tag{B.1}$$

we see that u is an $O(E^{1/2})$ quasimode. On the other hand, writing $u = \sum_{j=1}^{\infty} a_j\phi_j$ we have

$$\begin{aligned} \|(\Delta + E)u\|^2 = \|\sum_j a_j(E - E_j)\phi_j\|^2 &= \sum_j |a_j|^2(E - E_j)^2 \\ &\geq d(E, \sigma)^2 \sum_j |a_j|^2 = d(E, \sigma)^2. \end{aligned} \tag{B.2}$$

Combining (B.1) and (B.2) completes the proof. □

REFERENCES

[1] A. Ancona, *A note on the Rellich formula in Lipschitz domains*, Pub. Mathemàtiques, 42 (1998), pp. 223–237.

[2] A. Bäcker, S. Fürstberger, R. Schubert, and F. Steiner, *Behaviour of boundary functions for quantum billiards*, J. Phys. A, 35 (2002), pp. 10293–10310.

[3] A. H. Barnett, *Asymptotic rate of quantum ergodicity in chaotic Euclidean billiards*, Comm. Pure Appl. Math., 59 (2006), pp. 1457–88.

[4] ———, *Perturbative analysis of the Method of Particular Solutions for improved inclusion of high-lying Dirichlet eigenvalues*, SIAM J. Numer. Anal., 47 (2009), pp. 1952–1970.

[5] A. H. Barnett and T. Betcke, *Quantum mushroom billiards*, CHAOS, 17 (2007), p. 043123. 13 pages, `nlin.CD/0611059`.

[6] ———, *Stability and convergence of the Method of Fundamental Solutions for Helmholtz problems on analytic domains*, J. Comput. Phys., 227 (2008), pp. 7003–7026.

[7] T. Betcke, *A GSVD formulation of a domain decomposition method for planar eigenvalue problems*, IMA J. Numer. Anal., 27 (2007), pp. 451–478.

[8] ———, *The generalized singular value decomposition and the Method of Particular Solutions*, SIAM J. Sci. Comp., 30 (2008), pp. 1278–1295.

[9] T. Betcke and L. N. Trefethen, *Reviving the method of particular solutions*, SIAM Rev., 47 (2005), pp. 469–491.

[10] H. Chen, *Irregular but non-fractal drums, and n-dimensional Weyl conjecture*, Acta Math. Sinica, New Series, 11 (1995), pp. 168–178.

[11] A. Comech, *Cotlar-Stein almost orthogonality lemma*, preprint, http://www.-math.tamu.edu/∼comech/papers/CotlarStein/CotlarStein.pdf, (2007).

[12] M. Cotlar, *A combinatorial inequality and its applications to l^2-spaces*, Rev. Mat. Cuyana, 1 (1955), pp. 41–55.

[13] C. K. D. Jerison, *The inhomogeneous Dirichlet problem in Lipschitz domains*, J. Funct. Anal., 130 (1995), pp. 161–219.

[14] J. Descloux and M. Tolley, *An accurate algorithm for computing the eigenvalues of a polygonal membrane*, Comput. Methods Appl. Mech. Engrg., 39 (1983), pp. 37–53.

[15] T. A. Driscoll, *Eigenmodes of isospectral drums*, SIAM Rev., 39 (1997), pp. 1–17.

[16] L. Fox, P. Henrici, and C. Moler, *Approximations and bounds for eigenvalues of elliptic operators*, SIAM J. Numer. Anal., 4 (1967), pp. 89–102.

[17] P. R. Garabedian, *Partial differential equations*, John Wiley & Sons Inc., New York, 1964.

[18] C. Gordon, D. Webb, and S. Wolpert, *Isospectral plane domains and surfaces via Riemannian orbifolds*, Invent. Math., 110 (1992), pp. 1–22.

[19] A. Hassell and T. Tao, *Upper and lower bounds for normal derivatives of Dirichlet eigenfunctions*, Math. Res. Lett., 9 (2002), pp. 289–305.

[20] J. R. Kuttler and V. G. Sigillito, *Inequalities for membrane and Stekloff eigenvalues*, J. Math. Anal. Appl., 23 (1968), pp. 148–160.

[21] ———, *Eigenvalues of the Laplacian in two dimensions*, SIAM Rev., 26 (1984), pp. 163–193.

[22] Z. C. Li, *The Trefftz method for the Helmholtz equation with degeneracy*, Applied Numer. Math., 58 (2008), pp. 131–159.

[23] W. C. H. McLean, *Strongly elliptic systems and boundary integral equations*, Cambridge University Press, 2000.

[24] C. B. Moler and L. E. Payne, *Bounds for eigenvalues and eigenvectors of symmetric operators*, SIAM J. Numer. Anal., 5 (1968), pp. 64–70.

[25] P. Monk and D.-Q. Wang, *A least-squares method for the helmholtz equations*, Comput. Meth. Appl. Mech. Engrg., 175 (1999), pp. 121–136.

[26] Y. Netrusov and Y. Safarov, *Weyl asymptotic formula for the Laplacian on domains with rough boundaries*, Commun. Math. Phys., 253 (2005), pp. 481–509.

[27] G. Pólya and G. Szego, *Isoperimetric inequalities in mathematical physics*, Annals of Mathematics Studies, no. 27, Princeton university press, Princeton, NJ, 1951.

[28] F. Rellich, *Darstellung der Eigenwerte von $\Delta u + \lambda u = 0$ durch ein Randintegral*, Math. Z., 46 (1940), pp. 635–636.

[29] Y. Safarov and D. Vassiliev, *The Asymptotic Distribution of Eigenvalues of Partial Differential Operators*, Translations of Mathematical Monographs #155, American Mathematical Society, Providence, RI, 1996.

[30] N. Saito, *Data analysis and representation on a general domain using eigenfunctions of Laplacian*, Applied and Computational Harmonic Analysis, 25 (2008), pp. 68–97.

[31] R. Seeley, *An estimate near the boundary for the spectral function of the Laplace operator*, Amer. J. Math., 102 (1980), pp. 869–902.

[32] E. M. Stein, *Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals*, Monographs in Harmonic Analysis, Princeton university press, Princeton, NJ, 1993. with the assistance of Timothy S. Murphy.

[33] G. Still, *Computable bounds for eigenvalues and eigenfunctions of elliptic differential operators*, Numer. Math., 54 (1988), pp. 201–223.

[34] L. N. Trefethen and T. Betcke, *Computed eigenmodes of planar regions*, vol. 412 of Contemp. Math., Amer. Math. Soc., Providence, RI, 2006, pp. 297–314.

[35] H. E. Tureci, H. G. L. Schwefel, P. Jacquod, and A. D. Stone, *Modes of wave-chaotic dielectric resonators*, Progress in Optics, 47 (2005), pp. 75–137.

[36] E. Vergini and M. Saraceno, *Calculation by scaling of highly excited states of billiards*, Phys. Rev. E, 52 (1995), pp. 2204–2207.

[37] S. Zelditch, *Quantum ergodicity and mixing of eigenfunctions*, in Elsevier Encyclopedia of Mathematical Physics, vol. 1, Academic Press, 2006, pp. 183–196. `arXiv:math-ph/0503026`.

[38] S. Zelditch and M. Zworski, *Ergodicity of eigenfunctions for ergodic billiards*, Comm. Math. Phys., 175 (1996), pp. 673–682.

BIT manuscript No.
(will be inserted by the editor)

A new integral representation for quasi-periodic scattering problems in two dimensions

Alex Barnett · Leslie Greengard

the date of receipt and acceptance should be inserted later

Abstract Boundary integral equations are an important class of methods for acoustic and electromagnetic scattering from periodic arrays of obstacles. For piecewise homogeneous materials, they discretize the interface alone and can achieve high order accuracy in complicated geometries. They also satisfy the radiation condition for the scattered field, avoiding the need for artificial boundary conditions on a truncated computational domain. By using the *quasi-periodic* Green's function, appropriate boundary conditions are automatically satisfied on the boundary of the unit cell. There are two drawbacks to this approach: i) the quasi-periodic Green's function diverges for parameter families known as Wood's anomalies, even though the scattering problem remains well-posed, and ii) the lattice sum representation of the quasi-periodic Green's function converges in a disc, becoming unwieldy when obstacles have high aspect ratio.

In this paper, we bypass both problems by means of a new integral representation that relies on the free-space Green's function alone, adding auxiliary layer potentials on the boundary of the unit cell strip while expanding the linear system to enforce quasi-periodicity. Summing nearby images directly leaves auxiliary densities that are smooth, hence easily represented in the Fourier domain using Sommerfeld integrals. Wood's anomalies are handled analytically by deformation of the Sommerfeld contour. The resulting integral equation is of the second kind and achieves spectral accuracy. Because of our image structure, inclusions which intersect the unit cell walls are handled easily and automatically. We include an implementation and simple code example with a freely-available MATLAB toolbox.

A. H. Barnett
Department of Mathematics, Dartmouth College, Hanover, NH, 03755, USA. tel:+1-603-646-3178. fax:+1-603-646-1312 E-mail: ahb@math.dartmouth.edu

L. Greengard
Courant Institute, New York University, 251 Mercer St, NY, 10012, USA. E-mail: greengard@cims.nyu.edu

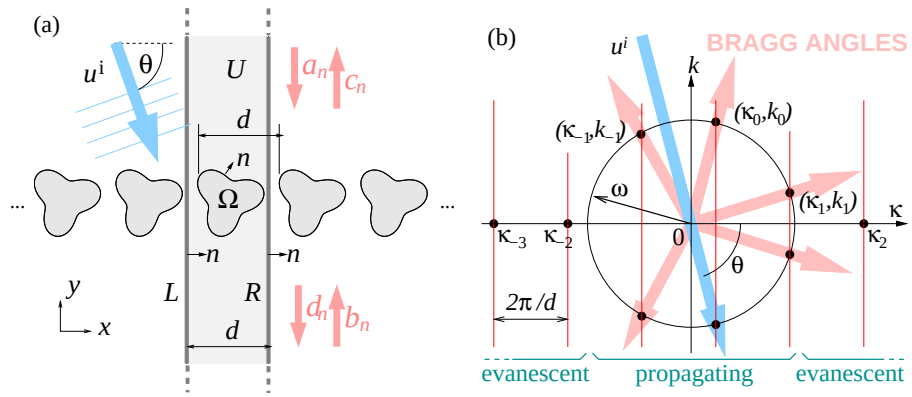


Fig. 1.1 (a) A simple 2D grating scattering geometry: a plane wave incident on a one-dimensional array of obstacles with periodicity d. $\theta \in (-\pi, 0)$ is the angle of incidence with respect to the x-axis. $\Omega \subset \mathbb{R}^2$ denotes the single obstacle lying in the "unit cell" $U := (-d/2, d/2) \times (-\infty, \infty)$. Ω_Λ denotes the set of all obstacles in the array, and $\mathbb{R}^2 \setminus \overline{\Omega_\Lambda}$ denotes the background material. (b) Fourier ($\mathbf{k}$-space) picture showing construction of Rayleigh-Bloch wavevectors, the propagating ones of which give the angles of Bragg diffraction orders.

Keywords scattering · integral equation · periodic · quasi-periodic · grating

Mathematics Subject Classification (2000) 35J25 · 65R20 · 65N38 · 78A45 · 35P25

1 Introduction and problem formulation

A number of problems in wave propagation require the calculation of *quasi-periodic* solutions to the Helmholtz or Maxwell equations in the frequency domain. A typical example in two dimensions involves computing the field scattered by an infinite one-dimensional array (a grating) of dielectric obstacles in response to an incoming plane wave (Fig. 1.1) [37].

For the sake of simplicity, we will restrict our attention here to the analysis of the two-dimensional (locally isotropic) Maxwell equations in what is usually called TM polarization[1] [17,18]. In this case, the full Maxwell equations for the electric and magnetic fields $\mathbf{E}, \mathbf{H}$,

$$\nabla \times \mathbf{E} = i\omega\mu\mathbf{H}$$
$$\nabla \times \mathbf{H} = -i\omega\epsilon\mathbf{E}$$

[1] The reader should be warned that in this setting some authors [5] use TM and TE in the reversed sense.

take the simpler form

$$\mathbf{E}(x,y,z) = \mathbf{E}(x,y) = (0,0,u^{\mathrm{t}}(x,y))$$
$$\mathbf{H}(x,y,z) = \mathbf{H}(x,y) = \frac{1}{i\omega\mu}(u^{\mathrm{t}}_y(x,y), -u^{\mathrm{t}}_x(x,y), 0).$$

where u^{t} is a scalar field defined in $\mathbb{R}^2$. Here, ϵ and μ are the permittivity and permeability, respectively, and we have assumed a time dependence of $e^{-i\omega t}$ at frequency $\omega > 0$. The index of refraction $n = n(x,y)$ is defined to be $n = \sqrt{\epsilon\mu}$. We will concentrate on materials with piecewise constant index of refraction, and assume that units are scaled so that $n = 1$ in the background material and $n \neq 1$ in the obstacles themselves (Fig. 1.1). In order to satisfy the full Maxwell equations, it is straightforward to verify that u^{t} must satisfy the scalar Helmholtz equations

$$(\Delta + n^2\omega^2)u^{\mathrm{t}} = 0 \qquad \text{in } \Omega_\Lambda \tag{1.1}$$
$$(\Delta + \omega^2)u^{\mathrm{t}} = 0 \qquad \text{in } \mathbb{R}^2 \setminus \overline{\Omega_\Lambda} \tag{1.2}$$

Imposing the standard continuity conditions on the tangential components of the electric and magnetic fields [17,18], we have

$$u^{\mathrm{t}+} - u^{\mathrm{t}-} = u^{\mathrm{t}\,+}_n - u^{\mathrm{t}\,-}_n = 0 \quad \text{on } \partial\Omega_\Lambda, \tag{1.3}$$

where the superscripts $+$ $(-)$ denote the limiting values on the surface approaching from outside (inside), and $u^{\mathrm{t}}_n := \partial u^{\mathrm{t}}/\partial n$ is the outward-pointing normal derivative. Writing $\mathbf{x} = (x,y)$, the incoming plane wave with angle θ takes the form

$$u^{\mathrm{i}}(\mathbf{x}) = \begin{cases} 0 & \text{in } \Omega_\Lambda \\ e^{i\mathbf{k}\cdot\mathbf{x}} & \text{in } \mathbb{R}^2 \setminus \overline{\Omega_\Lambda} \end{cases} \tag{1.4}$$

where the wavenumber is $\mathbf{k} = (\kappa^{\mathrm{i}}, k^{\mathrm{i}}) := (\omega\cos\theta, \omega\sin\theta)$. Since $|\mathbf{k}| = \omega$, u^{i} satisfies the appropriate Helmholtz equation in the background material. We will make frequent use of the following.

Definition 1 *A function $f : \mathbb{R}^2 \to \mathbb{C}$ is said to be* quasi-periodic[2] *with phase α, $|\alpha| = 1$, if*

$$f(x+d, y) = \alpha f(x,y) \qquad \forall (x,y) \in \mathbb{R}^2$$

We interpret α as the Bloch phase associated with translation by one period. In particular, the incident wave is quasi-periodic with phase

$$\alpha = e^{i\kappa^{\mathrm{i}} d} = e^{i\omega d\cos\theta}.$$

(This condition reduces to true periodicity only when $\omega d\cos\theta = 2n\pi$ for some integer n, such as when the incoming wave is normally incident, i.e. $\theta = -\pi/2$.) By the translational symmetry of the problem, the resulting total field u^{t} must be similarly quasi-periodic.

[2] also known as pseudoperiodic in some literature.

In the language of scattering theory, we may now write the total field as $u^{\mathrm{t}} = u^{\mathrm{i}} + u$, where u is the scattered field. It is easy to see that u satisfies the homogeneous Helmholtz equations (1.1)-(1.2), and that (1.3) results in the inhomogenous interface conditions

$$u^+ - u^- = -u^{\mathrm{i}}, \quad u_n^+ - u_n^- = -u_n^{\mathrm{i}} \qquad \text{on } \partial\Omega_\Lambda\,. \tag{1.5}$$

Moreover, u is quasi-periodic.

Finally, u satisfies the following radiation condition [9]. Let $y_0 > 0$ be sufficiently large that the rectangle $(-d/2, d/2) \times (-y_0, y_0)$ contains $\overline{\Omega}$. Define $\kappa_n = \kappa^{\mathrm{i}} + 2\pi n/d$, and $k_n = +\sqrt{\omega^2 - \kappa_n^2}$ where the sign is taken as positive real or positive imaginary. Then the plane wave with wavevector (κ_n, k_n) is quasi-periodic for each integer n, and satisfies the Helmholtz equation at frequency ω; a geometric construction for these wavevectors is given in Fig. 1.1b. The radiation condition on u is that it has uniformly-convergent upwards and downwards Rayleigh–Bloch expansions

$$u(x,y) = \sum_{n\in\mathbb{Z}} c_n e^{i\kappa_n x} e^{ik_n(y-y_0)} \qquad y > y_0,\ x \in \mathbb{R} \tag{1.6}$$

$$u(x,y) = \sum_{n\in\mathbb{Z}} d_n e^{i\kappa_n x} e^{ik_n(-y-y_0)} \qquad y < -y_0,\ x \in \mathbb{R} \tag{1.7}$$

The complex coefficients c_n, d_n, for n such that $|\kappa_n| \le \omega$ (propagating waves), are the Bragg diffraction amplitudes at the grating orders. For all other n these give evanescent components which do not contribute to the far field.

The above conditions result in the existence of a solution u for all frequencies ω and incident angles θ [9, Thm. 3.2]. The solution is unique (i.e. the problem well-posed) at all but a discrete set of frequencies ω, for each fixed horizontal wavenumber $\kappa^{\mathrm{i}} = \omega\cos\theta$ [9, Thm. 3.4]. These discrete frequencies correspond to guided modes, which propagate without loss along the grating. The only such modes relevant in the scattering setting are *nonrobust* in the sense that they are embedded in the continuous spectrum, and therefore disappear under generic perturbations [39]. Such modes occur at isolated points in $(\omega, \kappa^{\mathrm{i}})$ parameter space, and may cause anomalous transmission through the grating [40]. (There may also exist robust guided modes with frequencies below the continuous spectrum corresponding to the mode's horizontal wavenumber, which are therefore inaccessible in the scattering setting.) If $n < 1$ and Ω is shaped such that the refractive index is non-decreasing with $|y|$ then no guided modes can exist [9, Thm. 3.5]. Conversely, for arbitrary Ω and $n > 1$, at least one guided mode must exist [9, Thm 4.4 (2)] (although the existence of nonrobust modes is not guaranteed). The above is carefully reviewed (in the three-dimensional setting) by Shipman [39].

There is another entirely different type of "resonance" that may occur, first observed experimentally by Wood [45]: there are parameters where the diffraction intensities may change arbitrarily rapidly with respect to ω or θ. These so-called *Wood's anomalies* are now known to occur when, for some n, $\kappa_n = \pm\omega$ thus $k_n = 0$ and the nth Rayleigh–Bloch mode in (1.6)–(1.7) is

constant in y (a horizontally traveling plane wave). It is also possible to have a *double* Wood's anomaly, when $\kappa_n = -\kappa_m = \omega$ for a pair of integers n, m; clearly this can only occur for integer $\omega d/2\pi$. A thorough analysis of Wood's anomalies in the case of disc obstacles is given by Linton–Thompson [26]. We emphasize that, unlike guided modes discussed above, Wood's anomalies do *not* prevent the boundary value problem for u from having a unique solution. One of the key features of the method we present is that it works at Wood's anomalies.

Applications of the above diffraction problem are widespread, including gratings, antennae, acoustic diffusers, and small-amplitude water waves [36]. Generalizing slightly to more than one inclusion per unit cell, one is able to study finite-thickness photonic crystal slabs [42, 40], with applications to integrated optics and signal processing. Moreover, we expect the tools we present here to generalize well to a much larger class of technologically-important scattering problems, including periodic layered media with inclusions. Modern applications include dielectric gratings [8], process control in semiconductor lithography [29], efficient thin-film photovoltaic cells [4, 19], and integrated nano-scale optical devices [18]. Some of these applications (e.g. photonic crystals, silicon microwire arrays [19]) demand structures that are many periods deep, i.e. high aspect ratio.

A variety of existing numerical methods have been applied to periodic scattering problems. Unfortunately, many of the popular methods are only low order accurate and agree in test problems to only 1 digit of accuracy [44]. Finite difference approximation to the time-domain wave equation (FDTD) [41, 15] is common; it has the advantage of being quite general, but has the disadvantage of requiring large regions of free space to be simulated, with significant dispersion error for electrically large domains. FDTD methods also require non-reflecting boundary conditions and long settling times for resonant structures. Moving to the frequency domain, finite element methods [5, 29] require non-trivial grid generation, especially in three dimensions, and also demand the discretization of substantial regions of free space together with the imposition of non-local radiation conditions constructed from (1.6) and (1.7). The engineering community has developed many other methods—including coupled wave analysis, the waveguide method [44], the C method [23], and cylindrical (multipole) expansions [36, 26]—which are often based on uncontrolled approximations.

In this work we use integral equations (sometimes known as boundary element methods). These have a rigorous mathematical foundation [11, 33], and have the advantages of discretizing the material interfaces alone, of automatically enforcing radiation conditions, and of high-order convergence, which allows high accuracies to be reached very efficiently. The standard way to adapt integral equations to solve periodic scattering problems is by replacing the free-space kernel (see (2.7)) by the quasi-periodic Green's function (the kernel appearing in (3.4) with $P = \infty$); see [42, 34, 35]. Considerable effort has been spent on the tricky task of computing this quasi-periodic Green's function efficiently (see, for example, [24, 27, 30, 3]). The state of the art appears to be

around 10^{-4} to 10^{-3} sec per evaluation at low frequencies [3], and 10^{-1} sec at very high frequencies [22].

A much more serious problem is that the quasi-periodic Green's function blows up at Wood's anomalies with a -1/2 power singularity [24, Eq.(2.13)–(2.14)], causing any method using it to have divergent round-off error near to, and failure at, each such anomaly. In the related case of periodic surface scattering, Zhang and Chandler-Wilde [46] modified the Green's function to that of a half-space, which cures this divergence (this was implemented in [2]); however, this idea fails to help in our case of disconnected obstacles. A final problem is that the quasi-periodic Green's function is often computed using *lattice sums* [24,25], which is natural when using fast multipole acceleration in large-scale scattering problems [35]. However, since this representation converges in discs (or spheres in 3D), it becomes cumbersome for high aspect ratio geometries. We believe that the above reasons highlight the need for the more robust and flexible scheme that we present here.

We propose a periodizing scheme which builds upon the single-obstacle scattering problem *without* changing the kernel, allowing existing code (quadrature, fast multipole schemes, etc) to be used unaltered. Our scheme has much in common with that of [7], except that we only have one dimension of periodicity, requiring Fourier methods to handle the unbounded nature of the other dimension.

We review the non-periodic scattering scheme in Sec. 2, and present our new method in Sec. 3. A simple numerical example, and then a high-aspect ratio one, are demonstrated in Sec. 4, where we also compute Bragg diffraction amplitudes c_n and d_n from (1.6)–(1.7). Slight modifications of the method allow handling of Wood's anomalies in Sec. 5 and obstacles which touch (wrap around) the unit cell walls in Sec. 6. We conclude in Sec. 7, and give a short code example in the Appendix.

2 Transmission scattering from a single obstacle

The approach of this paper builds directly upon the boundary integral method for the scattering problem for a single isolated obstacle Ω, whose formulation and solution we now review. Given the incoming wave u^{i} given by (1.4), one seeks the scattered wave u satisfying the Helmholtz equations

$$(\Delta + n^2\omega^2)u = 0 \qquad \text{in } \Omega \tag{2.1}$$

$$(\Delta + \omega^2)u = 0 \qquad \text{in } \mathbb{R}^2 \setminus \overline{\Omega}\,, \tag{2.2}$$

the inhomogeneous boundary conditions (1.5) on the single boundary $\partial\Omega$, and the free space Sommerfeld radiation condition [10]

$$\frac{\partial u}{\partial r} - i\omega u = o(r^{-1/2}), \qquad r := |\mathbf{x}| \to \infty \tag{2.3}$$

To formulate this as a system of integral equations, we choose the standard Rokhlin–Müller scheme [32,38],

$$u \;=\; \begin{cases} \mathcal{S}^{(n\omega)}\sigma + \mathcal{D}^{(n\omega)}\tau & \text{in } \Omega \\ \mathcal{S}^{(\omega)}\sigma + \mathcal{D}^{(\omega)}\tau & \text{in } \mathbb{R}^2 \setminus \overline{\Omega} \end{cases} \tag{2.4}$$

where the usual single and double layer densities [10] at any wavenumber $\omega > 0$ are defined by

$$(\mathcal{S}^{(\omega)}\sigma)(\mathbf{x}) = \int_{\partial\Omega} G^{(\omega)}(\mathbf{x},\mathbf{y})\sigma(\mathbf{y})ds_{\mathbf{y}} \tag{2.5}$$

$$(\mathcal{D}^{(\omega)}\tau)(\mathbf{x}) = \int_{\partial\Omega} \frac{\partial G^{(\omega)}}{\partial n_{\mathbf{y}}}(\mathbf{x},\mathbf{y})\tau(\mathbf{y})ds_{\mathbf{y}} \tag{2.6}$$

Here the kernel $G^{(\omega)}(\mathbf{x},\mathbf{y}) := G^{(\omega)}(\mathbf{x}-\mathbf{y})$ is the free-space Green's function,

$$G^{(\omega)}(\mathbf{x}) = \frac{i}{4}H_0^{(1)}(\omega|\mathbf{x}|), \qquad \mathbf{x} \in \mathbb{R}^2 \setminus \{\mathbf{0}\}, \tag{2.7}$$

where $H_0^{(1)}$ is the outgoing Hankel function of order zero.

The representation (2.4) already satisfies (2.1)–(2.2) and (2.3). Taking limits on the surface, using the standard jump relations for single and double layer potentials [10,13], we can evaluate the *mismatch* (boundary inhomogeneity),

$$\begin{bmatrix} u^+ - u^- \\ u_n^+ - u_n^- \end{bmatrix} \;=\; \left(\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} D^{(\omega)} - D^{(n\omega)} & S^{(n\omega)} - S^{(\omega)} \\ T^{(\omega)} - T^{(n\omega)} & D^{(n\omega)\,*} - D^{(\omega)\,*} \end{bmatrix} \right) \begin{bmatrix} \tau \\ -\sigma \end{bmatrix} \;=:\; A_0\eta \tag{2.8}$$

Here I is the identity operator, while S and D are defined to be the limiting boundary integral operators (maps from $C(\partial\Omega) \to C(\partial\Omega)$) with the kernels $\mathcal{S}$ and $\mathcal{D}$ interpreted in the principal value sense, and D^* is the adjoint of D. (S is actually weakly singular so the limit is already well defined. A standard calculation [10,13] shows that D is weakly singular as well). The hypersingular operator T has the kernel $\frac{\partial^2 G(\mathbf{x},\mathbf{y})}{\partial n_{\mathbf{x}}\partial n_{\mathbf{y}}}$ and is unbounded as a map from $C(\partial\Omega) \to C(\partial\Omega)$. In these definitions, as in (2.5)-(2.6), it is implied that G inherits the appropriate superscripts from S, D and T. Equating (2.8) with the known mismatch (1.5) gives the following linear equation for the unknown densities $\eta := [\tau; -\sigma]$,

$$A_0\eta = \begin{bmatrix} -u^{\mathrm{i}}|_{\partial\Omega} \\ -u_n^{\mathrm{i}}|_{\partial\Omega} \end{bmatrix} =: b\,. \tag{2.9}$$

Because only the *difference* of hypersingular T operators appears in (2.8), A_0 is a compact perturbation of the identity, and (2.9) is a Fredholm system of integral equations of the second kind. Furthermore, it has a unique solution at all ω [10, Thm 3.41].

We apply the Nyström method to discretize (2.9) with N quadrature points $\{\mathbf{y}_j\}_{j=1}^N$ on $\partial\Omega$, using the periodic trapezoid rule in the curve parametrization, corrected by the Martensen–Kussmaul weight scheme to handle the weak

diagonal singularities, as proposed by Kress [20] (also see [7, Sec. 3.1]). This approximates (2.9) by the $2N$-by-$2N$ linear system

$$\boldsymbol{A}_0\boldsymbol{\eta} = \boldsymbol{b}$$

where $\boldsymbol{\eta}$ is the length-$2N$ column vector of approximations to the densities $[\tau; -\sigma]$ sampled at the quadrature points, and $\boldsymbol{b}$ the similar vector of samples of the right-hand side of (2.9). This scheme gives spectral accuracy, that is, the sampled densities tend to their exact values with error $O(e^{-cN})$ for some $c > 0$.

3 Quasi-periodizing scheme

We restrict attention to the single unbounded strip "unit cell" $U := \{(x, y) \in (-d/2, d/2) \times (-\infty, \infty)\}$, in which we can reformulate the boundary value problem for u (as in, e.g. [9]). Recall that in U, u satisfies the homogeneous partial differential equations (2.1)–(2.2) with inhomogeneous boundary conditions (1.5) on $\partial\Omega$. We introduce the *discrepancy* functions

$$f := u|_L - \alpha^{-1}u|_R \tag{3.1}$$

$$f_n := u_n|_L - \alpha^{-1}u_n|_R \tag{3.2}$$

which measure the failure of the scattered field to satisfy quasi-periodicity between the left (L) and right (R) walls of U. Note that the normals on both walls are defined in the positive x-direction (see Fig. 1.1a). Quasi-periodicity implies vanishing discrepancy,

$$f(y) = f_n(y) = 0\,, \qquad y \in \mathbb{R}\,. \tag{3.3}$$

Moreover, by the unique continuation of Cauchy data for the Helmholtz equation [12], the converse holds. Thus the above conditions on u, combined with the radiation conditions (1.6)–(1.7), are equivalent to that of Sec. 1. We may interpret this as a two-port "channel" or waveguide scattering problem in U, with the radiation conditions being that u has no incoming channel modes from either above or below.

3.1 Scattered field representation and linear system in spatial domain

Our representation builds on (2.4). We first include a direct sum of neighboring phased copies of the obstacle which, as we shall see, will make the subsequent correction to enforce quasi-periodicity very simple. For this, we define

$$(\tilde{\mathcal{S}}^{(\omega)}\sigma)(\mathbf{x}) = \int_{\partial\Omega} \sum_{j=-P}^{P} \alpha^j G^{(\omega)}(\mathbf{x}, \mathbf{y} + j\mathbf{d})\, \sigma(\mathbf{y}) ds_{\mathbf{y}} \tag{3.4}$$

$$(\tilde{\mathcal{D}}^{(\omega)}\tau)(\mathbf{x}) = \int_{\partial\Omega} \sum_{j=-P}^{P} \alpha^j \frac{\partial G^{(\omega)}}{\partial n_{\mathbf{y}}}(\mathbf{x}, \mathbf{y} + j\mathbf{d})\, \tau(\mathbf{y}) ds_{\mathbf{y}} \tag{3.5}$$

where $\mathbf{d} = (d, 0)$ is the lattice translation vector. $P = 0, 1, \ldots$ is a numerical parameter that we may adjust to improve the convergence rate of the scheme. We add auxiliary densities on the unit cell walls as follows. We use the notation $\mathcal{S}_W$ and $\mathcal{D}_W$ to denote the analogs of (2.5)–(2.6) on the curve W, where W is one of the two walls L or $R = L + \mathbf{d}$. We represent the scattered field in U by

$$u \;=\; \begin{cases} \mathcal{S}^{(n\omega)}\sigma + \mathcal{D}^{(n\omega)}\tau & \text{in } \Omega \\ \tilde{\mathcal{S}}^{(\omega)}\sigma + \tilde{\mathcal{D}}^{(\omega)}\tau + (\mathcal{S}_L + \alpha\mathcal{S}_R)\mu + (\mathcal{D}_L + \alpha\mathcal{D}_R)\nu & \text{in } U \setminus \overline{\Omega} \end{cases} \tag{3.6}$$

where μ and ν are unknown single- and double-layer density functions on the real line $\{y \in \mathbb{R}\}$. We will stack these densities into the column vector $\xi := [\nu; -\mu]$.

Imposing the mismatch condition (1.5), and also the new quasi-periodicity condition (3.3) in the form $[f; f_n] = 0$, gives the two block rows of a linear operator system

$$E \begin{bmatrix} \eta \\ \xi \end{bmatrix} := \begin{bmatrix} A & B \\ C & Q \end{bmatrix} \begin{bmatrix} \eta \\ \xi \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix} \tag{3.7}$$

where b is defined as before by (2.9). Notice that the lower part of the right-hand side vector is zero because the incident wave contributes no discrepancy.

The four operators in the 2×2 blocks of E capture the various effects of the obstacle and auxiliary densities on the mismatch on $\partial\Omega$ and the discrepancy. In order to describe them, we will need notation for the free-space boundary integral operators from a source curve W to a (possibly distinct) target curve V,

$$(S_{V,W}\sigma)(\mathbf{x}) = \int_W G^{(\omega)}(\mathbf{x}, \mathbf{y})\sigma(\mathbf{y})ds_{\mathbf{y}} \qquad \mathbf{x} \in V \tag{3.8}$$

$$(D_{V,W}\tau)(\mathbf{x}) = \int_W \frac{\partial G^{(\omega)}}{\partial n_{\mathbf{y}}}(\mathbf{x}, \mathbf{y})\tau(\mathbf{y})ds_{\mathbf{y}} \qquad \mathbf{x} \in V \tag{3.9}$$

$$(D^*_{V,W}\sigma)(\mathbf{x}) = \int_W \frac{\partial G^{(\omega)}}{\partial n_{\mathbf{x}}}(\mathbf{x}, \mathbf{y})\sigma(\mathbf{y})ds_{\mathbf{y}} \qquad \mathbf{x} \in V \tag{3.10}$$

$$(T_{V,W}\tau)(\mathbf{x}) = \int_W \frac{\partial^2 G^{(\omega)}}{\partial n_{\mathbf{x}}\partial n_{\mathbf{y}}}(\mathbf{x}, \mathbf{y})\tau(\mathbf{y})ds_{\mathbf{y}} \qquad \mathbf{x} \in V\ . \tag{3.11}$$

When $V = W$, these operators are to be understood in the principal value sense. When neither V nor W is given, they are both assumed to be $\partial\Omega$, as in Sec. 2. By analogy with (3.4), (3.5), versions of these operators whose kernels include the phased sum over $2P + 1$ images of the source curve will be indicated with a tilde ($\sim$). Note that $*$ no longer signifies operator adjoint, since the adjoint of $D_{V,W}$ is $D^*_{W,V}$.

We now give formulae for the four operator blocks in E. By analogy with (2.8), (3.6) gives

$$A = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} \tilde{D} - D^{(n\omega)} & S^{(n\omega)} - \tilde{S} \\ \tilde{T} - T^{(n\omega)} & D^{(n\omega)*} - \tilde{D}^* \end{bmatrix}\ .$$

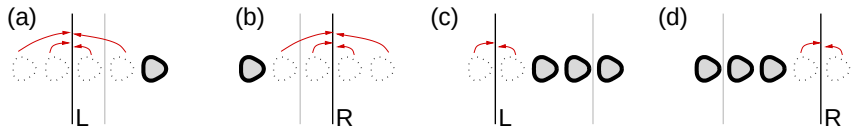


Fig. 3.1 Cancellations in discrepancy occurring in C operator block, for the case $P = 2$. Each arrow represents the influence of a source density on a target segment. For $Z = 0$ (standard method), the effect of the four source images (dotted) on L in (a) cancel the effect of the four source images on R in (b). For $Z = 1$ (which handles intersection of $\partial\Omega$ with unit cell walls) shown in (c) and (d), now only two source images cancel. See Sec. 6 for a discussion of the parameter Z.

Since the $j \neq 0$ terms here involve only non-local interactions, A remains 2nd kind for $P > 0$. The B block describes the auxiliary densities' effect on the mismatch, and follows easily from (3.6),

$$B = \begin{bmatrix} D_{\partial\Omega,L} & -S_{\partial\Omega,L} \\ T_{\partial\Omega,L} & -D^*_{\partial\Omega,L} \end{bmatrix} + \alpha \begin{bmatrix} D_{\partial\Omega,R} & -S_{\partial\Omega,R} \\ T_{\partial\Omega,R} & -D^*_{\partial\Omega,R} \end{bmatrix} . \tag{3.12}$$

The C block describes the obstacle densities' effect on the discrepancy functions $[f; f_n]$ of (3.1)–(3.2). By translational symmetry, as as shown in Fig. 3.1, there are cancellations of $2P$ terms in each subblock of C, leaving only

$$C = \alpha^P \begin{bmatrix} D_{L-P\mathbf{d},\partial\Omega} & -S_{L-P\mathbf{d},\partial\Omega} \\ T_{L-P\mathbf{d},\partial\Omega} & -D^*_{L-P\mathbf{d},\partial\Omega} \end{bmatrix} - \alpha^{-P-1} \begin{bmatrix} D_{L+(P+1)\mathbf{d},\partial\Omega} & -S_{L+(P+1)\mathbf{d},\partial\Omega} \\ T_{L+(P+1)\mathbf{d},\partial\Omega} & -D^*_{L+(P+1)\mathbf{d},\partial\Omega} \end{bmatrix} \tag{3.13}$$

The result is that C only involves interactions of distance at least Pd (this is similar to cancellation in [7]). Thus, increasing P can improve the convergence rate of the quadrature scheme for the solution of the full system (3.7).

Finally, the Q block gives the effect of auxiliary densities on the wall discrepancy itself. Inserting the μ and ν terms of (3.6) into (3.1)–(3.2) one easily gets,

$$Q = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} + \alpha \begin{bmatrix} D_{L,R} & -S_{L,R} \\ T_{L,R} & -D^*_{L,R} \end{bmatrix} - \alpha^{-1} \begin{bmatrix} D_{R,L} & -S_{R,L} \\ T_{R,L} & -D^*_{R,L} \end{bmatrix} \tag{3.14}$$

Now the reason becomes clear for choosing in (3.6) densities on the right wall which are suitably phased copies of those on the left: cancellation of the wall L, L and R, R self-interaction terms occurred in the above evaluation of Q, as shown in Fig. 3.1 (c), (d). This means that Q is the identity plus a convolution operator involving interactions of distance at least d, hence that the linear integral equation system (3.7) is of the 2nd kind. Note that, if densities were placed on L alone, by contrast, no cancellation would occur, and the L, L self-interaction would prevent (3.7) from being 2nd kind. Because the walls L and R are infinitely long, (3.7) is not usable in practice; however, we will presently transform it into a useful form. (We omit a rigorous analysis of solvability. That requires some additional work, since E acts on functions some of which live on unbounded domains and is not a compact perturbation of the identity.)

3.2 Fourier representation of layer potentials and discrepancy

The wall densities μ, ν in a solution of (3.7) represent the field due to the remainder of the infinite line of obstacle densities not accounted for by the direct sum in (3.6), namely the obstacles labelled $\ldots, P-2, P-1$ and $P+1, P+2, \ldots$. Since they must represent Rayleigh-Bloch modes (1.6)–(1.7), these densities generally do not decay as $|y| \to \infty$. Therefore any attempt to spatially truncate densities on L and R would give very slow convergence. To avoid this, we will work instead with the Fourier transform in the y direction, using the usual definitions

$$\hat{g}(k) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iky} g(y)dy\,, \qquad g(y) = \int_{-\infty}^{\infty} e^{iky}\hat{g}(k)dk\,, \tag{3.15}$$

where, as before, k denotes the y-wavenumber. We will also need the Cartesian spectral representation for the free space Green's function [31, Ch. 7.2],

$$G^{(\omega)}(\mathbf{x}) \;=\; \frac{i}{4\pi}\int_{-\infty}^{\infty} e^{iky}\,\frac{e^{i\sqrt{\omega^2-k^2}\,|x|}}{\sqrt{\omega^2-k^2}}\,dk\,, \qquad \mathbf{x} = (x,y) \tag{3.16}$$

Here the square-roots are taken positive real or positive imaginary, which may be achieved by taking the branch cuts of the function $\sqrt{\omega^2-k^2}$ in the k plane as $(-\infty,-\omega)\cup(\omega,+\infty)$ along the real axis, and a Sommerfeld contour passing from the 2nd to 4th quadrants (see Fig. 3.2). Examining (3.16), we see that horizontal separation ($|x|>0$) leads to exponentially decaying tails in the Fourier representation, which we will be able to truncate with spectral accuracy.

We define "Fourier layer potentials" by inserting (3.15) and (3.16) into the definition of layer potentials living on a vertical wall $W = \{(x_0, y) : y \in \mathbb{R}\}$,

$$(\hat{\mathcal{S}}_W\hat{\mu})(\mathbf{x}) = \frac{i}{2}\int_{-\infty}^{\infty} e^{iky}e^{i\sqrt{\omega^2-k^2}\,|x-x_0|}\frac{1}{\sqrt{\omega^2-k^2}}\hat{\mu}(k)\,dk \tag{3.17}$$

$$(\hat{\mathcal{D}}_W\hat{\nu})(\mathbf{x}) = \frac{\operatorname{sign}(x-x_0)}{2}\int_{-\infty}^{\infty} e^{iky}e^{i\sqrt{\omega^2-k^2}\,|x-x_0|}\hat{\nu}(k)\,dk = -\frac{\partial}{\partial x}(\hat{\mathcal{S}}_W\hat{\nu})(\mathbf{x}) \tag{3.18}$$

Here $\hat{\mu}$ and $\hat{\nu}$ are the Fourier transforms of the single- and double-layer densities μ and ν respectively, and $x_0 = -d/2$ or $d/2$ for $W = L$ or R respectively. (3.17)–(3.18) satisfy the usual jump relations on the wall.

From now on we will use $\hat{\xi} := [\hat{\nu}, -\hat{\mu}]$ as our unknown auxiliary densities, and furthermore switch to using the Fourier transform of the discrepancy functions $[\hat{f}, \hat{f}']$ to enforce the second block row of our linear system. The obstacle densities and mismatch remain as before. Thus (3.7) becomes,

$$\hat{E}\begin{bmatrix}\eta\\ \hat{\xi}\end{bmatrix} := \begin{bmatrix}A & \hat{B}\\ \hat{C} & \hat{Q}\end{bmatrix}\begin{bmatrix}\eta\\ \hat{\xi}\end{bmatrix} = \begin{bmatrix}b\\ 0\end{bmatrix} \tag{3.19}$$

where A is as before. $\hat{B}$ is the analog of B with (3.17)–(3.18) in place of (2.5)–(2.6). By analogy with (3.8)–(3.9) let $\hat{S}_{V,W}$ and $\hat{D}_{V,W}$ denote (3.17) and

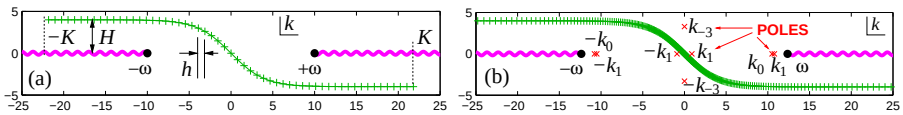


Fig. 3.2 (a) Sommerfeld contour in the complex k (vertical wavenumber) plane, using basic quadrature scheme (3.21)–(3.22) on a tanh curve. The pink wavy lines denote branch cuts towards infinity of $1/\sqrt{\omega^2-k^2}$. (b) Poles at the Rayleigh–Bloch vertical wavenumbers $\pm k_n$, and sinh-bunched nodes (3.25) giving improved convergence.

(3.18) evaluated on a target curve V. Their analytic derivatives in the normal direction at the target are also easy to evaluate, analogous to (3.10)–(3.11),

$$(\hat{D}^*_{V,W}\hat{\mu})(\mathbf{x}) = \frac{-1}{2}\int_{-\infty}^{\infty} e^{iky} e^{i\sqrt{\omega^2-k^2}\,|x-x_0|}\Big(\operatorname{sign}(x-x_0), \frac{k}{\sqrt{\omega^2-k^2}}\Big)\cdot \mathbf{n}(\mathbf{x})\,\hat{\mu}(k)\,dk\,, \quad \mathbf{x}\in V$$

$$(\hat{T}_{V,W}\hat{\nu})(\mathbf{x}) = \frac{i}{2}\int_{-\infty}^{\infty} e^{iky} e^{i\sqrt{\omega^2-k^2}\,|x-x_0|}\big(\sqrt{\omega^2-k^2}, k\operatorname{sign}(x-x_0)\big)\cdot \mathbf{n}(\mathbf{x})\,\hat{\nu}(k)\,dk\,, \quad \mathbf{x}\in V$$

With this notation, $\hat{B}$ is given by (3.12) but with $\wedge$ applied to each entry.

Conversely, it is easy get via (3.16) the Fourier transform of the field (or its normal derivative) on the wall W due to densities on $\partial\Omega$:

$$(S^\wedge_{W,\partial\Omega}\sigma)(k) = \frac{i}{4\pi}\int_{\partial\Omega} e^{-iky}\frac{e^{i\sqrt{\omega^2-k^2}\,|x-x_0|}}{\sqrt{\omega^2-k^2}}\,\sigma(\mathbf{y})ds_{\mathbf{y}}, \qquad \mathbf{y}=(x,y)$$

$$(D^\wedge_{W,\partial\Omega}\tau)(k) = \frac{1}{4\pi}\int_{\partial\Omega} e^{-iky} e^{i\sqrt{\omega^2-k^2}\,|x-x_0|}\Big(-\operatorname{sign}(x-x_0), \frac{k}{\sqrt{\omega^2-k^2}}\Big)\cdot \mathbf{n}(\mathbf{y})\,\tau(\mathbf{y})ds_{\mathbf{y}}$$

$$(D^{*,\wedge}_{W,\partial\Omega}\sigma)(k) = \frac{1}{4\pi}\int_{\partial\Omega} e^{-iky} e^{i\sqrt{\omega^2-k^2}\,|x-x_0|}\operatorname{sign}(x-x_0)\,\sigma(\mathbf{y})ds_{\mathbf{y}}$$

$$(T^\wedge_{W,\partial\Omega}\tau)(k) = \frac{i}{4\pi}\int_{\partial\Omega} e^{-iky} e^{i\sqrt{\omega^2-k^2}\,|x-x_0|}\big(\sqrt{\omega^2-k^2}, -k\operatorname{sign}(x-x_0)\big)\cdot \mathbf{n}(\mathbf{y})\,\tau(\mathbf{y})ds_{\mathbf{y}}$$

Then $\hat{C}$ is given by (3.13) but with $(\,\cdot\,)^\wedge$ applied to each entry.

Finally, $\hat{Q}$ maps from Fourier to Fourier variables. Since each of the 2×2 blocks in Q is a convolution, the corresponding blocks in $\hat{Q}$ are simply multiplication operators. By evaluating (3.17)–(3.18) between L and R, and using (3.15), (3.14) becomes

$$\hat{Q} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} + \frac{e^{i\sqrt{\omega^2-k^2}\,d}}{2}\begin{bmatrix} i(\alpha-\alpha^{-1})/\sqrt{\omega^2-k^2} & -\alpha-\alpha^{-1} \\ \alpha+\alpha^{-1} & i(\alpha-\alpha^{-1})\sqrt{\omega^2-k^2} \end{bmatrix} \tag{3.20}$$

where each block is to be understood as a multiplication operator in the vertical wavenumber variable k.

We now have explicit formulae for all blocks of the linear operator system (3.19) in which Fourier variables have been used to represent the spatially-unbounded y coordinate. We will proceed to truncate this in the k wavenumber domain and numerically approximate its solution.

3.3 Quadrature of the Sommerfeld contour integral

We now discretize (3.19) using the Nyström method; for the densities on $\partial\Omega$ this is done as before in Sec. 2. However, the domain of the Fourier variable k is unbounded, and the integrand has singularities, demanding a more specialized quadrature. (3.16) serves as a model for the type of integrands we need to handle; we expect this since, if our method is successful, the Fourier layer potentials on L and R must represent the effect of sources of the form (3.16) lying on neighboring copies of the obstacles—more precisely, those labelled $\ldots, P-2, P-1$ and $P+1, P+2, \ldots$. A complicating fact is that the infinite nature of the obstacle array sum produces Rayleigh–Bloch waves of the form (1.6)–(1.7), each wave of which introduces poles in the integrand at $\pm k_n$; we postpone addressing this aspect until Sec. 5. Here we present a scheme that we have found to be excellent for generic scattering parameters (i.e., not close to a Wood's anomaly).

The Sommerfeld integral (3.16) could be performed accurately on the real k axis using the trapezoid rule with endpoint corrections [1] appropriate for the $(-1/2)$-power singularities at $\pm\omega$. However, for reasons of superior performance, of the additional poles mentioned above, and of simplicity, we instead prefer the unit-slope hyperbolic tangent contour

$$k(\gamma) = \gamma - iH\tanh(\gamma/H), \qquad \gamma \in (-\infty, \infty) \tag{3.21}$$

shown in Fig. 3.2. H controls the size of the curve in the k plane; for unit period and obstacles of radius no larger than 2, we have found that $H = 4$ is a good choice for a wide range of ω. Our basic quadrature scheme is the M-point periodic trapezoid rule on the interval $-K \le \gamma \le K$, that is,

$$\gamma_j = -K + 2Kj/M, \qquad j = 1, \ldots, M \tag{3.22}$$

which we will see is exponentially convergent in M and K. The quadrature approximation is $\int_{-\infty}^{\infty} g(k)dk \approx \sum_{j=1}^{M} w_j g(k_j)$ where $k_j = k(\gamma_j)$ and the weights are

$$w_j = \frac{2K}{M}k'(\gamma_j) = h\left(1 - i\,\mathrm{sech}^2\,\frac{\gamma_j}{H}\right)$$

When $|x|$ in the integrand (3.16) has a lower bound $x_{\min}$—which we expect due to the horizontal separation of the unit cell walls from the nearest obstacle images—the integrand has exponential tails. We get an estimate of a sufficient K by equating the factor $e^{i\sqrt{\omega^2-K^2}\,x_{\min}}$ with machine precision $\varepsilon_{\mathrm{mach}} \approx 2 \times 10^{-16}$, giving

$$K = \sqrt{\left(\frac{\log \varepsilon_{\mathrm{mach}}}{x_{\min}}\right)^2 + \omega^2}\,. \tag{3.23}$$

In practice this works well; for example, approximating our model problem (3.16) for $\omega = 10$ at the point $x = x_{\min} = 1$, $y = 2$ yields $K = 38$, and gives a relative error 5×10^{-11} with $M = 123$ nodes, which improves to 3×10^{-15} with $M = 179$. (In comparison, to achieve 7×10^{-11} error using endpoint corrections on the real axis, $M = 138$ is needed, and errors for larger M always exceed

2×10^{-14}.) The extremely rapid convergence in M just observed is actually geometric for $2K$-periodic analytic integrands (as a function of γ), with a constant of convergence inversely proportional to the distance from the real axis of the nearest singularity [21, Thm. 9.28]. Despite not being periodic, our integrand is exponentially small beyond $\pm K$, thus may be treated as periodic with negligible error [43, Sec. 8].

When ω is small, the singularities at $\pm\omega$ approach the k-plane contour while K stays large, causing a terrible convergence rate (e.g. for the above case with $\omega = 0.1$, $M = 7000$ points are needed to reach an error of 10^{-14}). This may be cured by bunching the nodes towards the origin as in Fig. 3.2(b) via the conformal transformation,

$$\gamma(\beta) = \lambda\delta \sinh(\beta/\lambda), \qquad \beta \in (-\infty, \infty) \tag{3.24}$$

where δ controls of the width of the central uniform node region, and λ controls the geometric expansion rate outside this region. We apply the truncated periodic trapezoid rule to β,

$$\beta_j = -L + 2Lj/M, \qquad j = 1, \ldots, M, \qquad L = \lambda \sinh^{-1}(K/\delta\lambda),$$

where L was chosen so that $\gamma(\pm L) = \pm K$. Back in the k plane this gives nodes and (via the chain rule) weights

$$k_j = k(\gamma(\beta_j)), \qquad w_j = \frac{2L}{M}\gamma'(\beta_j)k'(\gamma(\beta_j)), \tag{3.25}$$

where $\gamma'(\beta) = \delta\cosh(\beta/\lambda)$. We have not yet fully analyzed the best choice of δ and λ, but find good performance in the range $10^{-15} \leq \omega \leq 10^3$ with $\delta = \min(\omega/4, 1)$ and $\lambda = \max((\omega/2)^{1.2}, 2)$, with not much sensitivity to the particular constants chosen here. For extreme frequencies, the M needed for good accuracy grows: $M = O(\log(1/\omega))$ as $\omega \to 0$, and $M = O(\omega)$ as $\omega \to \infty$ (due to the growth of $K = O(\omega)$). For instance, returning to $\omega = 0.1$, an error of 10^{-15} is reached at $M = 177$, and there is even improvement at $\omega = 10$ to an error of 4×10^{-15} at $M = 137$. This idea of transformations to improve periodic trapezoid rule convergence is not new (for a recent application and review of the literature, see Hale et al. [14]).

In summary, the Nyström method is used to approximate (3.19) by a discretized square linear system of size $2N + 2M$,

$$\hat{\boldsymbol{E}} \begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{\xi} \end{bmatrix} := \begin{bmatrix} \boldsymbol{A} & \hat{\boldsymbol{B}} \\ \hat{\boldsymbol{C}} & \hat{\boldsymbol{Q}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{\xi} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{bmatrix}. \tag{3.26}$$

For the upper and left blocks, the N quadrature nodes and weights on $\partial\Omega$ are as in Sec. 2. For the lower and right blocks we use quadrature in the Fourier variable with nodes and weights (3.25). Although the latter were justified using the model integrand (3.16), we will see in the next section that they perform very well for the full problem.

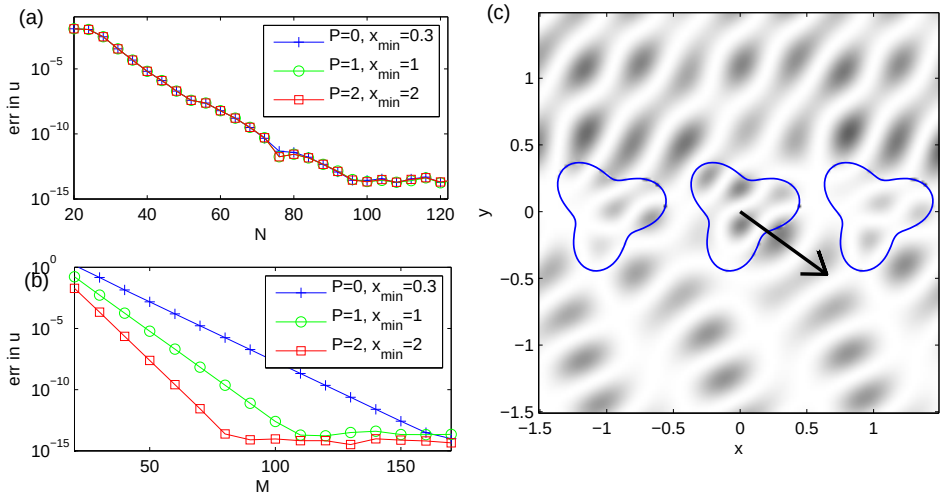


Fig. 4.1 Results for obstacle, index $n = 1.5$, with shape given by the polar function $r(\phi) = 0.35 + 0.105\cos 3(\phi - \pi/12)$ shown in Fig. 1.1(a), with $d = 1$, and incident plane wave $\omega = 10$ and $\theta = -\pi/5$. (a) Convergence in N of absolute error in field value at $(x, y) = (0.2, 1)$, with $M = 170$ fixed at its converged value. Three different P values and $x_{\min}$ values are shown. (b) Convergence in M of the same, with $N = 110$ fixed at its converged value. Errors are measured relative to the converged value, being $0.19914162575818 + 0.24760159732486i$ to 14 digits. (c) Total field $(\mathrm{Re}[u^{\mathrm{t}}])^2$ (i.e. electric energy density at one instant of time), plotted over three periods. Black indicates large values, white zero (dark patches are half a wavelength apart). The arrow shows incident θ.

4 Numerical examples

We now demonstrate the performance of the basic scheme presented above, away from Wood's anomalies and when $\partial\Omega$ does not touch the walls L or R. (Both restrictions will later be removed, in Sections 5 and 6 respectively.) In Fig. 4.1(a)–(b), we show convergence with respect to N (number of obstacle nodes) and M (number of Fourier nodes) of the scattered field value at a single point, for a grating of simple smooth (analytic) trefoil obstacles defined in the caption, and shown in Fig. 1.1(a). The convergence is exponential in both parameters. Increasing P includes more directly-summed neighbors in the A block of the system, hence the fields due to these neighbors no longer have to be represented by the Fourier layer potentials. Thus $x_{\min}$ can be increased, which increases the convergence rate in M, as shown in (b). (Changing P has no effect on convergence in N, as (a) shows). Why then should the user not increase P without limit? The answer is that computing the extra direct sums required to fill the matrix elements in the A block, and evaluate the solution field, takes more time. The optimal P will depend on the trade-off between matrix filling plus field evaluation vs linear system solution time.

With $P = 2$, $x_{\min} = 2$, at parameters $N = 110$, $M = 90$, at which 14 digits of the field value have converged, the CPU time required[3] to fill the matrix is 0.5 sec, and 0.1 sec to solve the complex dense linear system of size $2N + 2M = 400$. The scattered field is evaluated, using the discretized version of (3.6), on a grid (x- and y-spacing 0.02, with 7500 grid points) in 6 sec, giving Fig. 4.1(c). We note that, in contrast to the method of [7], there are no issues with evaluation of the Fourier layer potentials close to the walls: $\mathrm{dist}(\partial\Omega, R)$ is only 0.056. The quadrature scheme is accurate up to the boundary of the unit cell, as long as $|y|$ is not too large (with these parameters, less than about 2). However, as usual with boundary integral equations, full-accuracy evaluation up to the boundary of the obstacle would require specialized methods.

The linear system (3.26) has $\mathrm{cond}(\hat{\boldsymbol{E}}) = 9 \times 10^4$. If we perform simple left and right diagonal preconditioning of the system matrix $\hat{\boldsymbol{E}}$—first scaling rows to have unit 2-norm, then scaling columns to have unit 2-norm—we get a new system with matrix denoted by $\hat{\boldsymbol{E}}_s$ which has a reduced condition number of 4×10^2. This is only a little larger than for the isolated-obstacle problem of Sec. 2, which here has $\mathrm{cond}(\boldsymbol{A}_0) = 31$.

What happens in this example if the frequency is changed? At the low frequency $\omega = 0.1$, 14 digit accuracy is easily reached by increasing M to 180. As $\omega \to 0$, we observe $\mathrm{cond}(\hat{\boldsymbol{E}}_s) = O(1/\omega)$ but note that, since $u = O(\omega)$, an absolute error of 10^{-14} is still achieved. At the high frequency $\omega = 50$ (8 wavelengths per period), one needs $N = 230$, but only $M = 160$, to achieve errors around 10^{-14}; solution time is 2 sec to fill and 0.6 sec to solve, and $\mathrm{cond}(\hat{\boldsymbol{E}}_s)$ is only 7×10^2.

4.1 Extracting Bragg amplitudes and checking flux conservation

In applications it is crucial to know the far-field behavior, i.e., the amplitudes of the propagating Bragg orders c_n in (1.6) and d_n in (1.7). We compute these from the density vector $[\boldsymbol{\eta}; \boldsymbol{\xi}]$, by using (3.6) to evaluate u and u_n on the horizontal lines $-d/2 < x < d/2$, $y = \pm y_0$. (We choose y_0 to leave at least $d/2$ separation from $\partial\Omega$). On the bottom side, for each mode n for which $k_n \neq 0$,

$$\begin{bmatrix} d_n \\ b_n \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & -1/ik_n \\ 1 & 1/ik_n \end{bmatrix} \begin{bmatrix} d^{-1} \int_{-d/2}^{d/2} u(x, -y_0) e^{-i\kappa_n x} dx \\ d^{-1} \int_{-d/2}^{d/2} u_y(x, -y_0) e^{-i\kappa_n x} dx \end{bmatrix} \tag{4.1}$$

where the integrals extract Fourier series coefficients of the d-periodic field $u(x, -y_0)e^{-i\alpha x/d}$ and its y-derivative. The 2×2 matrix projects this into downwards (d_n) and upwards (b_n) propagating amplitudes. Similarly, replacing $-y_0$ by y_0 in (4.1) gives the pair $[a_n; c_n]$, namely the downwards and upwards amplitudes on the top side (see Fig. 1.1(a)). Neither "incoming" amplitude a_n nor b_n is needed here—they are zero for a scattered field obeying the radiation

[3] Here, and elsewhere in this work, computation times are for a 2006-era laptop with 2GHz Intel Core Duo CPU and 2GB of RAM, running MATLAB 2008a and `MPSpack` (see Appendix). There is also significant overhead in the object-oriented code.

conditions—but one will rear its head in Sec. 5. We approximate the integrals numerically via 20-point Gaussian quadrature.

An independent test of the accuracy of the method is then given by the conservation of (energy) flux. For the exact full field it holds that total outgoing flux equals total incoming flux [26, Eq.(40)], [39, p. 12], that is,

$$\sum_{k_n>0} k_n \big(|c_n|^2 + |d_n + \delta_{n,0} e^{-ik^{\mathrm{i}} y_0}|^2\big) = k^{\mathrm{i}} \,. \tag{4.2}$$

where the term with the Kronecker delta accounts for the unit-amplitude incident plane wave in the $n = 0$ mode. In the above $\omega = 10$ example the "flux error" (difference between left and right hand sides of the above) is 3×10^{-14}, consistent with the error estimated from convergence of u at a point; results are similar at the other ω values.

4.2 High aspect ratio example

One motivation for our work is handling high aspect ratio obstacles. Here we present such an example. For tall obstacles many periods in height where $y_{\max} := \max_{(x,y)\in\partial\Omega} |y| \gg 1$ (we assume $d = 1$), exponential growth of the factor e^{iky} in the imaginary k direction causes undesirably large integrands along the Sommerfeld contour, hence large matrix entries and large condition number. This leads us to reduce the scale of the tanh curve (3.21); we choose the height $H = \min(4, 3/y_{\max})$. A side-effect is that a larger M is needed since the contour comes closer to the singularities on the real axis. Thus the choice of H is a trade-off between matrix size and condition number.

In Fig. 4.2 we show the full field for the distorted ellipse of height 10 periods described in the caption, at $\omega = 10$, computed using $P = 2$, $x_{\min} = 2.5$, and $\delta = \omega/4$. Table 4.1 gives numerical parameters for this case, and also for higher wavenumber (at $\omega = 20$ the scatterer is 32 exterior wavelengths long, and 48 interior wavelengths). t_{fill} is the time required to fill the matrix $\hat{\boldsymbol{E}}$, and t_{solve} the time to solve the dense linear system of size $2N + 2M$. In both cases, the N required is almost identical to the smallest N needed to achieve similar accuracy in the isolated obstacle scattering problem (Sec. 2). The plot in Fig. 4.2 took 24 sec to evaluate (grid spacing 0.04, with 7500 points).

It is clear that condition numbers appear to be growing with ω (this is also true for the isolated obstacle, but to a lesser extent). As mentioned above, they can be lowered, by about an order of magnitude, by decreasing H; however, this requires a somewhat larger M. The values shown are a good compromise that achieves 12-digit accuracy.

5 Scattering parameters near to and at Wood's anomalies

So far we have assumed that we are not close to or at a Wood's anomaly, i.e. that no k_n is close to the origin. Let us take a moment to explain why,

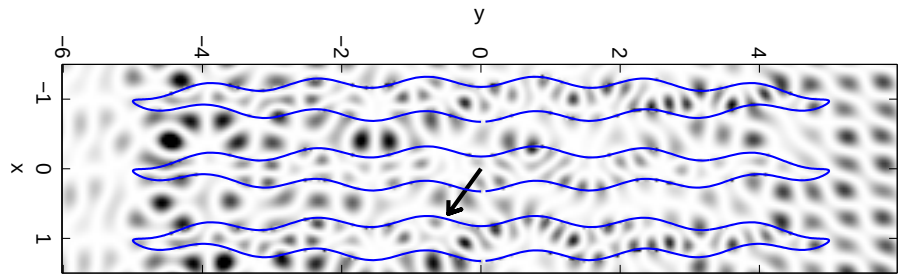


Fig. 4.2 High aspect ratio obstacle: a tall wobbly ellipse given in Cartesians by $\big(0.25\cos\phi + 0.075\cos(5\sin\phi), 0.25\sin\phi\big)$, $0 \le \phi < 2\pi$, with incident plane wave $\omega = 10$ and $\theta = -\pi/5$ (shown by arrow). Total field $(\mathrm{Re}[u^{\mathrm{t}}])^2$ is plotted over three periods. The figure has been rotated by $-\pi/2$ to fit on the page.

ω	N	M	flux error	conv. error	t_{fill}	t_{solve}	cond($\hat{\boldsymbol{E}}$)	cond($\hat{\boldsymbol{E}}_s$)
10	500	330	3e-13	1e-13	9.5 sec	4.5 sec	6e6	1.6e4
20	900	450	2e-12	1e-12	29 sec	18 sec	1.6e8	1.3e5

Table 4.1 Results for scattering from the wobbly ellipse grating of height 10 periods. "conv. error" is the approximate error estimated from convergence of a field value $u^{\mathrm{t}}(0.2, 5.5)$.

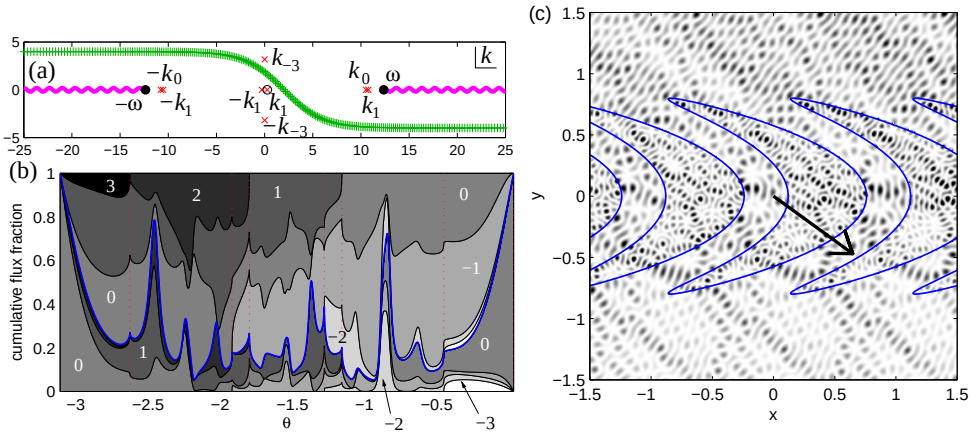


Fig. 5.1 (a) Deformed (translated) Sommerfeld contour; compare Fig. 3.2(b). The contour has crossed the circled pole. (b) Outgoing flux fractions for all Bragg modes ($-3 \le n \le 3$, some labelled by n) for the example of Fig. 4.1, as a function of incident angle $\theta \in (-\pi, 0)$. The top group (above blue line) are downwards (transmitted), the bottom group upwards (reflected). Note the square-root type cusps at Wood's anomalies (shown by vertical dotted lines). (c) Total field $(\mathrm{Re}[u^{\mathrm{t}}])^2$ for kite-shaped obstacles of index $n = 1.5$ and shape $(0.32\cos\phi + 0.64\cos 2\phi - 0.2, 0.8\sin\phi)$, which wrap around the unit cell; see Sec. 6.

in the representation of a scattered solution, there is generically a pole in the integrand (i.e. Fourier densities $\hat{\mu}(k)$ and $\hat{\nu}(k)$ in (3.17)–(3.18)) at $k = \pm k_n$, for each n (both propagating and evanescent); see Fig. 3.2(b). Distributionally,

we have the following Fourier transform pair, for $k_p \in \mathbb{C}$,

$$\hat{g}_p(k) = \frac{1}{2\pi i(k-k_p)} \qquad g_p(y) = \begin{cases} e^{ik_p y}, & y>0 \\ 0, & y<0 \end{cases} \quad \text{or} \quad \begin{cases} 0, & y>0 \\ -e^{ik_p y}, & y<0 \end{cases} \tag{5.1}$$

where in the first (second) case the pole k_p lies above (below) the contour. (This is easily checked using Jordan's lemma and the residue theorem.) Since *both* Rayleigh–Bloch expansions (1.6) and (1.7)—thus the densities that produce them—are represented by the *same* y-direction Fourier function, this function must have a pole at each k_n in order to capture the outgoing amplitude for $y>0$ but zero incoming amplitude for $y<0$. This pole must lie above the Sommerfeld contour to generate this first case rather than the second. By a similar argument, there are poles at $-k_n$, which must lie below the contour.

The astute reader will have noticed that these poles $\pm k_n$ are closer to the Sommerfeld contour (3.21) than are $\pm\omega$, thus realize that they (rather than ω) should control the node bunching parameter δ in (3.24). We have tested this idea, choosing e.g. $\delta = \min(|k_n|)/2$; however matrix condition number appears to grow as $O(1/|k_n|)$, losing digits of accuracy—in fact at the same rate as would a conventional quasi-periodic Green's function based method. This approach also requires large M and fails to get more than a few correct digits when $k_n = 0$.

Our proposal is to translate the Sommerfeld contour by $O(1)$ along the positive real axis, so as to move it to be a distance at least 1 away from the trouble at the origin (and in fact from any other k_n pole; see Fig. 5.1(a)). In doing so, the contour crosses a pole, which we label by $n=p$. From the above discussion, this corresponds to solving for a field u with peculiar radiation conditions: the pth mode is *incoming from below* ($c_p = a_p = 0$), instead of the usual outgoing from above ($b_p = a_p = 0$), whilst all other modes $n \neq p$ are as usual. To fix this we add to (3.6), the representation of u in $U \setminus \overline{\Omega}$, the quasi-periodic plane wave $ae^{i\kappa_p x}e^{ik_p y}$, where a is an additional unknown. This accounts for the *difference* between the two cases in (5.1). An additional constraint is then needed: we impose the correct mode incoming radiation condition $b_p = 0$. This adds to the linear system (3.26) a row whose entries are filled using quadrature applied to the expression for b_p in (4.1), giving,

$$\begin{bmatrix} \boldsymbol{A} & \hat{\boldsymbol{B}} & \boldsymbol{a}_p \\ \hat{\boldsymbol{C}} & \hat{\boldsymbol{Q}} & \boldsymbol{0} \\ \boldsymbol{r}_p & \boldsymbol{s}_p & e^{-ik_p y_0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{\xi} \\ a \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{0} \\ 0 \end{bmatrix}, \tag{5.2}$$

a square system of size $2N+2M+1$. The column vector $\boldsymbol{a}_p$ expresses the effect of the new plane wave on the obstacle mismatch $[u^+ - u^-; u_n^+ - u_n^-]$ at the nodes on $\partial\Omega$.

We apply the above whenever $|k_n| < 1$ for some n, and when this occurs for two values of n (close to a double Wood's anomaly), we apply its natural generalization to two extra degrees of freedom. Precisely at a Wood's anomaly, $k_n = 0$ and the projection in (4.1) onto exponentials degenerates into a projection onto the functions $\{1, y\}$, of which only the constant function is a valid

radiation condition, thus we replace the 2×2 matrix in this formula by the identity. Since the contour is now far from any pole, we might hope that the convergence in M while close to or at a Wood's anomaly is as good as it was far from one. This is true: choosing the example of Fig. 4.1, with $\omega = 10$, and $\theta = \cos^{-1}(1 - 2\pi/\omega d)$ which is a single Wood's anomaly for mode $n = 1$, we achieve convergence at $N = 110$, $M = 90$ to 14 digit accuracy to the value $u^{\mathrm{t}}(0.2, 1) = -0.61990437785568 + 0.46197831200416i$. Condition number after left and right diagonal preconditioning is 3.4×10^3, an order of magnitude larger than before, but still very reasonable.

We thus have made our scheme immune to Wood's anomalies. This lets us produce Fig. 5.1(b) which shows the flux fractions (individual terms in (4.2) normalized to unit total) in each of the modes, swept over all incident angles $-\pi < \theta < 0$. The idea of adding extra discrete unknowns and conditions to an integral equation system dates back at least to Mikhlin [28, §31]. A similar idea has been recently used for an array of sound-soft or sound-hard discs, to 2 significant digits [26]. Removal of singularities near a contour to improve convergence rates has also been used to evaluate periodic Green's functions [16].

6 Obstacle intersecting unit cell walls

When $\partial\Omega$ intersects L or R, the method presented so far fails, crudely speaking because parts of the obstacle see the fields on the wrong side of the wall layer potential. However, we can fix the scheme to handle this case in the following simple way. We redefine L to be the vertical line at location $x_0 = -(1/2+Z)d$, and R at location $x_0 = (1/2+Z)d$, with $Z = 0, 1, \ldots$ a new method parameter. Discrepancies (3.1)–(3.2) are redefined using α^{2Z+1} as the phase instead of α, and thus measure the difference from quasi-periodicity across a strip $2Z + 1$ periods wide. We use the representation (3.6), again replacing α by α^{2Z+1}. We ensure $P \geq Z$ so that obstacle images fill this wider strip. To construct the blocks of the linear system (3.7), we set the mismatch on $\partial\Omega$ as before, i.e. equal to b as defined in (2.9), and impose vanishing discrepancy, as before. Note that $Z = 0$ corresponds to the original method presented in Sec. 3.

Conjecture 2 *The method described above for $Z > 0$ has the same solution field u in the strip $(-d/2, d/2) \times (-\infty, \infty)$ as the method with $Z = 0$.*

This is perhaps surprising since the mismatch is only enforced on *one* of the many copies of $\partial\Omega$ which fall within the wider strip. Pending a formal proof, we explain this (away from Wood's anomalies) as follows. Using the Schur complement of the upper-left block in (3.19) we have

$$A_{\mathrm{QP}}\eta = b \qquad \text{where } A_{\mathrm{QP}} := A - \hat{B}(\hat{Q})^{-1}\hat{C}\,. \tag{6.1}$$

As in [7], it is easy to check that A_{QP} is now A_0 of (2.8) but with quasi-periodic Green's functions ((3.4)–(3.5) with $P = \infty$) in the exterior. This

argument applies for any $Z = 0, 1, \ldots$, thus the solution η, hence the full field, is independent of Z.

Implementing this gives the following changes in the operator system (3.7) and hence (3.19): The phase α becomes α^{2Z+1} in B and Q, i.e. (3.12), (3.14) and (3.20). Note that B now involves larger interaction distances of order Zd. For the C block, less cancellation occurs, as shown in Fig. 3.1(c)–(d), generalizing (3.13) to

$$C = \left(\sum_{j=P-2Z}^{P} - \sum_{j=-P-1}^{-P-1+2Z} \right) \alpha^j \begin{bmatrix} D_{L-j\mathbf{d},\partial\Omega} & -S_{L-j\mathbf{d},\partial\Omega} \\ T_{L-j\mathbf{d},\partial\Omega} & -D^*_{L-j\mathbf{d},\partial\Omega} \end{bmatrix}$$

$Z = 1$ is sufficient for wall-intersecting obstacles that are not much wider than d. It was used, for example, in Fig. 5.1(c), where we compute diffraction from a grating of kite-shaped obstacles which necessarily intersect any artificial vertical unit cell walls. The frequency $\omega = 50$ is quite high, needing $N = 550$, but only $M = 170$, to achieve convergence error of around 10^{-13} and flux error 4×10^{-14}. Computation times are 11.5 sec to fill, 3 sec to solve, and 170 sec to evaluate u at 30000 points for the plot, and $\mathrm{cond}(\hat{\boldsymbol{E}}_s)$ is 8.8×10^3.

In summary, by widening the scheme to one which quasi-periodizes across more than one period, we have effectively made the artificial walls "invisible" to the obstacle, without increasing the computational effort.

7 Conclusions

We have quasi-periodized a scheme for dielectric obstacle scattering, converting it into a scheme for an infinite grating of such obstacles, using only the free-space Green's function and auxiliary layer potentials based on Sommerfeld integrals on the unit cell walls. For typical frequencies $\omega d/2\pi = O(1)$, independently of the complexity of the obstacle, this adds only around 200 extra unknowns to the linear system, keeps condition number reasonable (the scheme is 2nd kind), and achieves errors close to machine precision, uniformly in incident angle.

As far as we are aware, this is the first time an integral equation approach to this problem has been formulated that works at Wood's anomalies (for surfaces see [2]). This is a promising alternative to the use of quasi-periodic Green's functions, which fail around Wood's anomalies, can be complicated to evaluate, and are complicated to use in high aspect ratio cells when based on lattice sums. We expect this advantage to be more pronounced for doubly-periodic problems in 3D, and when the obstacle scattering is accelerated by fast-multipole methods (FMM).

Some issues remain open for study, such as: i) ill-conditioning for very low ω and high aspect ratios, ii) better automated choice of method parameters H, λ, δ, P and Z, and iii) a proof of Conjecture 2 for all scattering parameters. There are several acceleration ideas that we have not presented here. Since $\hat{\boldsymbol{E}}$ is a matrix polynomial in α, the coefficients can be pre-computed and used to fill

the matrix quickly for multiple incident angles. We have only presented results using dense linear algebra (not even using the fact that $\hat{Q}$ is diagonal), so assessing GMRES convergence in large-scale examples with FMM acceleration is a next step.

We also note that the discretized version of the Schur complement (6.1) seems like a good $O(N^2M)$ way to compute N^2 evaluations of the quasi-periodic Green's function, should they be required in a more conventional solver. In the case of Fig. 4.1, for example, this allows each entry to be filled in around 10^{-5} sec, which is competitive with the 10^{-4} to 10^{-3} sec recently reported [3].

The algorithm we present is available through a free toolbox for MATLAB, as shown in the Appendix. Barring a few subtleties, we expect the method to generalize to multi-layered media with periodic inclusions, opening up a wealth of technological applications; we will study this in a future paper.

Acknowledgements We thank Stephen Shipman and Simon Chandler-Wilde for invaluable discussions. The work of AHB was supported by NSF grant DMS-0811005, and by the Class of 1962 Fellowship at Dartmouth College. The work of LG was supported by the Department of Energy under contract DEFG0288ER25053 and by AFOSR under MURI grant FA9550-06-1-0337.

A Simple MATLAB code example in `MPSpack`

`MPSpack` is an object-oriented MATLAB toolbox for solving 2D Helmholtz PDE problems recently developed by Betcke and one of the authors [6]. The methods of the present paper have been incorporated into version 1.1 beta of `MPSpack`. The package is freely available at `http://code.google.com/p/mpspack`. The download section of that Web page also includes an extensive tutorial and a technical manual. Below we give a short, 12-line, human-readable code which implements the methods of this paper to achieve 14 digits of accuracy in solving the quasi-periodic scattering problem of Sec. 4. The `qpscatt` class has been written in such a way that it builds upon the layer-potential scheme for the non-periodic scattering problem. The integer parameter P is called `nei`, and Z is called `buf`. Handling more complicated obstacles is quite easy; this, and other details, are explained in the tutorial.

```
% Time-harmonic scattering of plane wave from grating of dielectric obstacles
d = 1.0;                                          % problem period
N = 110; s = scale(segment.smoothstar(N, 0.3, 3), 0.35); % smooth closed curve
di = domain(s, 1); di.setrefractiveindex(1.5);    % obstacle, refractive index
de = domain([], [], s, -1);                       % obstacle's exterior
s.addinoutlayerpots('d'); s.addinoutlayerpots('s'); % add Rokhlin LP scheme
s.setmatch('diel', 'TM');                         % TM dielectric continuity
p = qpscatt(de, di, d, struct('nei',2,'buf',1,'M',90)); % numerical parameters
p.setoverallwavenumber(10);                       % incident wavenumber
p.setincidentwave(-pi/5);                         % incident plane wave angle
p.solvecoeffs;                                    % fill matrices and solve
[u d n] = p.braggpowerfracs(struct('table',1));   % Bragg order efficiencies
p.showfullfield(struct('ymax', 1));               % plot Re part of full field
```

References

1. Alpert, B.K.: Hybrid Gauss-trapezoidal quadrature rules. SIAM J. Sci. Comput. **20**, 1551–1584 (1999)
2. Arens, T., Chandler-Wilde, S.N., DeSanto, J.A.: On integral equation and least squares methods for scattering by diffraction gratings. Commun. Comput. Phys. **1**, 1010–42 (2006)
3. Arens, T., Sandfort, K., Schmitt, S., Lechleiter, A.: Analysing Ewald's method for the evaluation of Green's functions for periodic media (2010). Preprint, submitted to IMA J. Numer. Anal.; available at `http://digbib.ubka.uni-karlsruhe.de/volltexte/1000019136`
4. Atwater, H.A., Polman, A.: Plasmonics for improved photovoltaic devices. Nature Materials **9**(3), 205–213 (2010)
5. Bao, G., Dobson, D.C.: Modeling and optimal design of diffractive optical structures. Surv. Math. Ind. **8**, 37–62 (1998)
6. Barnett, A.H., Betcke, T.: An exponentially convergent non-polynomial finite element method for time-harmonic scattering from polygons. SIAM J. Sci. Comp. **32**(3), 1417–1441 (2010)
7. Barnett, A.H., Greengard, L.: A new integral representation for quasi-periodic fields and its application to two-dimensional band structure calculations. J. Comput. Phys. **229**, 6898–6914 (2010)
8. Barty, et al., C.P.J.: An overview of LLNL high-energy short-pulse technology for advanced radiography of laser fusion experiments. Nuclear Fusion **44**(12), S266 (2004)
9. Bonnet-BenDhia, A.-S., Starling, F.: Guided waves by electromagnetic gratings and non-uniqueness examples for the diffraction problem. Math. Meth. Appl. Sci. **17**, 305–338 (1994)
10. Colton, D., Kress, R.: Integral equation methods in scattering theory. Wiley (1983)
11. Colton, D., Kress, R.: Inverse acoustic and electromagnetic scattering theory, *Applied Mathematical Sciences*, vol. 93, second edn. Springer-Verlag, Berlin (1998)
12. Garabedian, P.R.: Partial differential equations. John Wiley & Sons Inc., New York (1964)
13. Guenther, R.B., Lee, J.W.: Partial differential equations of mathematical physics and integral equations. Prentice Hall, Englewood Cliffs, New Jersey (1988)
14. Hale, N., Higham, N.J., Trefethen, L.N.: Computing A, $\log(A)$, and related matrix functions by contour integrals. SIAM J. Numer. Anal. **46**(5), 2505–2523 (2008)
15. Holter, H., Steyskal, H.: Some experiences from FDTD analysis of infinite and finite multi-octave phased arrays. IEEE Trans. Antennae Prop. **50**(12), 1725–1731 (2002)
16. Horoshenkov, K.V., Chandler-Wilde, S.N.: Efficient calculation of two-dimensional periodic and waveguide acoustic Green's functions. J. Acoust. Soc. Amer. **111**, 1610–1622 (2002)
17. Jackson, J.D.: Classical Electrodynamics, 3rd edn. Wiley (1998)
18. Joannopoulos, J.D., Johnson, S.G., Meade, R.D., Winn, J.N.: Photonic Crystals: Molding the Flow of Light, 2nd edn. Princeton Univ. Press, Princeton, NJ (2008)
19. Kelzenberg, M.D., Boettcher, S.W., Petykiewicz, J.A., Turner-Evans, D.B., Putnam, M.C., Warren, E.L., Spurgeon, J.M., Briggs, R.M., Lewis, N.S., Atwater, H.A.: Enhanced absorption and carrier collection in Si wire arrays for photovoltaic applications. Nature Materials **9**(3), 239–244 (2010)
20. Kress, R.: Boundary integral equations in time-harmonic acoustic scattering. Mathl. Comput. Modelling **15**, 229–243 (1991)
21. Kress, R.: Numerical Analysis. Graduate Texts in Mathematics #181. Springer-Verlag (1998)
22. Kurkcu, H., Reitich, F.: Stable and efficient evaluation of periodized Green's functions for the Helmholtz equation at high frequencies. J. Comput. Phys. **228**, 75–95 (2009)
23. Li, L., Chandezon, J., Granet, G., Plumey, J.P.: Rigorous and efficient grating-analysis method made easy for optical engineers. Appl. Optics **38**(2), 304–313 (1999)
24. Linton, C.M.: The Green's function for the two-dimensional Helmholtz equation in periodic domains. J. Eng. Math. **33**, 377–402 (1998)

25. Linton, C.M.: Lattice sums for the Helmholtz equation. SIAM Review **52**(4), 603–674 (2010)
26. Linton, C.M., Thompson, I.: Resonant effects in scattering by periodic arrays. Wave Motion **44**, 165–175 (2007)
27. McPhedran, R.C., Nicorovici, N.A., Botten, L.C., Grubits, K.A.: Lattice sums for gratings and arrays. J. Math. Phys. **41**, 7808–7816 (2000)
28. Mikhlin, S.G.: Integral equations, 2nd edn. MacMillan, New York (1964)
29. Model, R., Rathsfeld, A., Gross, H., Wurm, M., Bodermann, B.: A scatterometry inverse problem in optical mask metrology. J. Phys.: Conf. Ser. **135**, 012,071 (2008)
30. Moroz, A.: Exponentially convergent lattice sums. Opt. Lett. **26**, 1119–21 (2001)
31. Morse, P., Feshbach, H.: Methods of theoretical physics, volume 1. McGraw-Hill (1953)
32. Müller, C.: Foundations of the Mathematical Theory of Electromagnetic Waves. Springer-Verlag, Berlin, New York (1969)
33. Nédélec, J.C., Starling, F.: Integral equation methods in a quasi-periodic diffraction problem for the time-harmonic Maxwell's equations. SIAM J. Math. Anal. **22**(6), 1679–1701 (1991)
34. Nicholas, M.J.: A higher order numerical method for 3-D doubly periodic electromagnetic scattering problems. Commun. Math. Sci. **6**(3), 669–694 (2008)
35. Otani, Y., Nishimura, N.: A periodic FMM for Maxwell's equations in 3D and its applications to problems related to photonic crystals. J. Comput. Phys. **227**, 4630–52 (2008)
36. Peter, M.A., Meylan, M.H., Linton, C.M.: Water-wave scattering by a periodic array of arbitrary bodies. J. Fluid Mech. **548**, 237–256 (2006)
37. Petit, R. (ed.): Electromagnetic Theory of Gratings, *Topics in Current Physics*, vol. 22. Springer-Verlag,Heidelberg (1980)
38. Rokhlin, V.: Solution of acoustic scattering problems by means of second kind integral equations. Wave Motion **5**, 257–272 (1983)
39. Shipman, S.: Resonant scattering by open periodic waveguides, *Progress in Computational Physics (PiCP)*, vol. 1, pp. 7–50. Bentham Science Publishers (2010)
40. Shipman, S., Venakides, S.: Resonance and bound states in photonic crystal slabs. SIAM J. Appl. Math. **64**, 322–342 (2003)
41. Taflove, A.: Computational Electrodynamics: The Finite-Difference Time-Domain Method. Artech House, Norwood, MA (1995)
42. Venakides, S., Haider, M.A., Papanicolaou, V.: Boundary integral calculations of two-dimensional electromagnetic scattering by photonic crystal Fabry-Perot structures. SIAM J. Appl. Math. **60**, 1686–1706 (2000)
43. Weideman, J.A.C.: Numerical integration of periodic functions: A few examples. Amer. Math. Monthly **109**(1), 21–36 (2002)
44. Wojcik, G.L., Jr., J.M., Marx, E., Davidson, M.P.: Numerical reference models for optical metrology simulation. In: SPIE Microlithography 92: IC Metrology, Inspection, and Process Control VI, #1673-06 (1992)
45. Wood, R.W.: On a remarkable case of uneven distribution of light in a diffraction grating spectrum. Philos. Mag. **4**, 396–408 (1902)
46. Zhang, B., Chandler-Wilde, S.N.: Integral equation methods for scattering by infinite rough surfaces. Math. Meth. Appl. Sci. **26**, 463–488 (2003)

A Few More Words About James Tenney: Dissonant Counterpoint and Statistical Feedback

Larry Polansky*, Alex Barnett*, and Michael Winter†

∗ Dartmouth College
† University of California Santa Barbara

November 6, 2010

Abstract

This paper discusses a compositional algorithm, important in many of the works of James Tenney, which models a melodic principle known as *dissonant counterpoint*. The technique synthesizes two apparently disparate musical ideas—dissonant counterpoint and statistical feedback—and has a broad range of applications to music which employs non-deterministic (i.e. randomized) methods. First, we describe the historical context of Tenney's interest in dissonant counterpoint, noting its connection to composer/theorist Charles Ames' ideas of statistical feedback in computer-aided composition. Next, we describe the algorithm in both intuitive and mathematical terms, and analyze its behavior and limiting cases via numerical simulations and rigorous proof. Finally, we describe specific examples and generalizations used in Tenney's music, and provide simple computer code for further experimentation.

Contents

1 Tenney, Dissonant Counterpoint, and Statistical Feedback 2
1.1 Tenney and Dissonant Counterpoint 2
1.2 Statistical Feedback: Probability vs. statistics 4

2 The Dissonant Counterpoint Algorithm 7
2.1 Informal description 7
2.2 Formal description 9
2.3 Curvature of the growth function 14
2.4 Vanishing power: "max-1" version and the effect of weights . 16

3 Examples from Tenney's Work **19**
3.1 *Seegersongs* 21
3.2 The *Spectrum* pieces 22
3.3 *To Weave (a meditation)* 25
3.4 *Panacousticon* 25
3.5 *Arbor Vitae* 26

4 Conclusion **28**

A Matlab/Octave code examples **29**

1 Tenney, Dissonant Counterpoint, and Statistical Feedback

> "Carl Ruggles has developed a process for himself in writing melodies for polyphonic purposes which embodies a new principle and is more purely contrapuntal than a consideration of harmonic intervals. He finds that if the same note is repeated in a melody before enough notes have intervened to remove the impression of the original note, there is a sense of tautology, because the melody should have proceeded to a fresh note instead of to a note already in the consciousness of the listener. Therefore Ruggles writes at least seven or eight different notes in a melody before allowing himself to repeat the same note, even in the octave."
>
> Henry Cowell, [13] (pp. 41–42)

> "Avoid repetition of any tone until at least six progressions have been made."
>
> Charles Seeger, *"Manual of Dissonant Counterpoint"* [26] (p. 174)

1.1 Tenney and Dissonant Counterpoint

The music of James Tenney often invokes an asynchronous musical community of collaborators past and present. Many of his pieces are dedicated to other composers, and poetically reimagine their ideas. In these works Tenney sometimes expresses connections to another composer's music via sophisticated transformations of the dedicatee's compositional methods. For

example *Bridge*[1] (1984), for two pianos, attempts to resolve the musical and theoretical connections between two composers—Partch and Cage—both of whom influenced Tenney [35, 33]. The resolution, or "reconciliation" [35], in this case the marriage of Cage-influenced chance procedures and Partch-influenced extended rational tunings, is the piece itself. The construction of that particular bridge made possible, for Tenney, a new style with interesting genealogies.

As a composer, performer, and teacher, Tenney took musical genealogy seriously. He seldom published technical descriptions of his own work,[2] but he wrote several innovative theoretical essays on the work of others (see for example [33, 32]). Some of these explications are, in retrospect, transparent theoretical conduits to his own ideas. Many of the historical connections in Tenney's work (to the Seegers, Partch, Cage, Varèse, even Wolpe) appear in or are suggested by his titles. However, Tenney's formal transformations of other composers' ideas are less well understood. The few cases in which he wrote about his own pieces (for example, see [34]) demonstrate the amount of compositional planning that went into each composition.

In Tenney's theoretical essay on the chronological evolution of Ruggles use of dissonance [31], he used simple statistical methods to explain Ruggles' (and by extension, the Seegers') melodic style. For example, he examined how long it took, on average, for specific pitch-classes and intervals to be repeated in a single melody. Tenney visualized these statistics in an unusual way: as sets of functions over (chronological) time—years, not measures. The article consists, for the most part, of graphs illustrating the statistical evolution of Ruggles' atonality along various axes. In general, the x-axis is Ruggles' compositional life itself. Tenney considered the *data*, such as the ways in which the lengths of unrepeated pitch-classes increased over time (and if in fact they did). Focusing on this specific aspect of Ruggles' work allowed Tenney to explain to himself, in part, what was going on in the music.

In the two decades that followed that paper, Tenney widened his theoretical focus to include the music and ideas of what might be called the 1930s "American atonal school," including Ruggles, Henry Cowell, Charles and Ruth Crawford Seeger, and others. These American composers em-

[1] All Tenney scores referenced in this article are available from Smith Publications or Frog Peak Music. Most of the works mentioned are recorded commercially as well.

[2] Tenney's personal papers, however, include many descriptions and notes pertaining to his pieces. Most of his music after about 1980 was written with the aid of a computer (after a long hiatus in that respect), and the software itself is extant. In addition, Tenney frequently described his compositional methods to his students.

ployed their own pre-compositional principles distinct from, but as rigorous as, European 12-tone composers of the same period. In the article on Ruggles, Tenney had described a way to statistically model an aspect of these composers' compositional intuitions. In the 1980s, he began to formally and computationally integrate Seeger's influential ideas of dissonant counterpoint [26, 27] into his own music. What seems at first to be a series of titular homages (in pieces like the *Seegersongs*, *Diaphonic Studies*, *To Weave (a meditation)*, and others) are in fact a complex, computer-based transplantation of dissonant counterpoint into the fertile soil of his own aesthetic.

In each of these pieces, and several others, Tenney employed a probabilistic technique which we call the *dissonant counterpoint algorithm*. Seegerian dissonant counterpoint encompassed a wide range of musical parameters (rhythm, tonality, intervallic use, meter, even form and orchestration). In this paper, we focus on an algorithm Tenney devised to make a certain kind of probabilistic selection (mostly pitch, but sometimes other things as well). This algorithm which was in part motivated by Tenney's interest in the ideas of dissonant counterpoint. To our knowledge he never published more than a cursory description of this technique [34]. One of the goals of our work is to present the algorithm and explore some of its features in a mathematical framework.

1.2 Statistical Feedback: Probability vs. statistics

> "Along with backtracking, statistical feedback is probably the most pervasive technique used by my composing programs. As contrasted with random procedures which seek to create unpredictability or lack of pattern, statistical feedback actively seeks to bring a population of elements into conformity with a prescribed distribution. The basic trick is to maintain statistics describing how much each option has been used in the past and to bias the decisions in favor of those options which currently fall farthest short of their ideal representation."
>
> Charles Ames [3]

The dissonant counterpoint algorithm is a special case of what the composer and theorist Charles Ames calls *statistical feedback*:[3] the current out-

[3]For a good explanation of this see [5]. However, several of Ames' other articles discuss it as well, including [8, 7, 6, 4, 2, 1].

come depends in some non-deterministic way upon previous outcomes.[4] Tenney's algorithm is an elegant, compositionally-motivated solution to this significant, if subtle compositional idea. Statistical feedback is another form of reconciliation—that of compositional method with musical results—and it has ramifications for any kind of computer- or anything-else-aided-composition or art form. We first give a general (and mathematically simple) introduction to this idea.

Imagine that we flip an unbiased coin $N = 1000$ times. We might end up with 579 heads and 421 tails. This is close to the equal statistical mean we might expect, want, or intend. With $N = 10000$ flips, we would do better, in the sense that although the numbers of heads or tails will most likely have larger differences from their expectation of 5000 than before, the *fraction* of heads is likely to be closer to 1/2 than before. This illustrates the law of large numbers: the average of N outcomes from an independent and identically distributed (iid) random process converges (almost surely) to its expected value as N increases. For example, baseball wisdom holds that "any given team on any given day can beat any other given team." But because the number of games played in a season, 162, is a pretty large number of trials, things generally (but not always) work out well for better teams.

But what about a more local observation of a small number of trials, or frame? For example, some run of ten flips might yield:

$$HTHHHHHTHT$$

This statistical frame contains, not surprisingly, something worse: seven Hs, three Ts.[5] Nothing in our method suggests that we want that: the act of

[4] Ames [5], in discussing his early compositional use of this technique, describes it more simply as "the trick of maintaining statistics detailing how much each option has been used in the past, and of instituting decision-making processes which most greatly favor those options whose statistics fall farthest behind their intended distribution." Also see [12], which surveys Ames' work (up until 1992), and contains an alternate mathematical formalization of statistical feedback (p. 35).

[5] And, from a more local, musical perspective, there are five Hs in a row, an example of what Ames refers to as "heterogeneity" or "dispersion" [5], and Polansky refers to as "clumping" [25, 21]. This is a slightly different, though related, problem to that of monitoring the global statistics of a probability distribution, but the solution also may employ statistical feedback. These terms refer to the difference between the following two, equally well-distributed coin toss statistics:

$$HHHHHTTTTT \quad \text{and} \quad HTHTHTHTHT$$

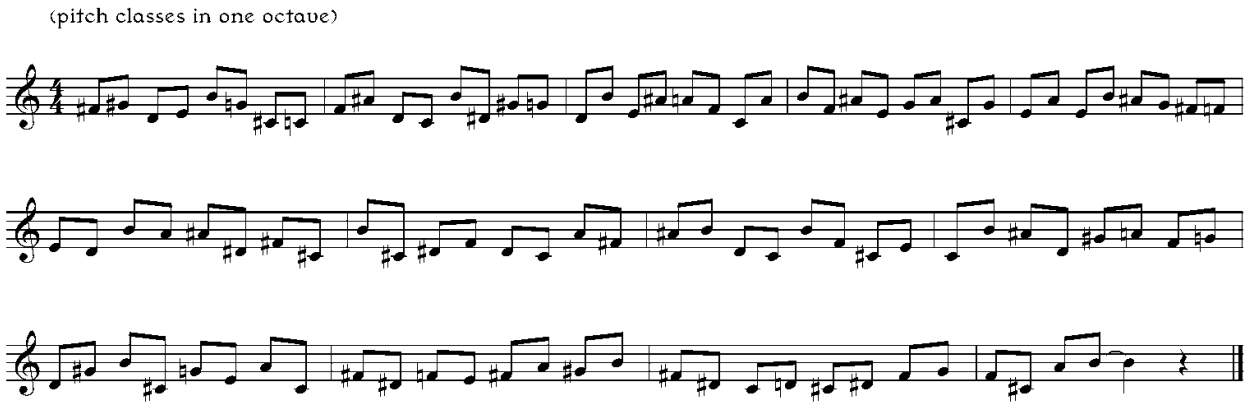


Figure 1: A simple melody generated by the algorithm, using a linear growth function (see next section), with $n = 12$ pitch elements lying in one octave. (Accidentals carry through the measure unless explicitly canceled.)

flipping an unbiased coin most likely (but not unequivocally) suggests that we desire a uniform distribution of outcomes. The random process creates a disjunct between compositional intention and statistical outcome.

Composers have long used probability distributions, but have not often worried about the conformance of observed statistics to probabilistic composition method over short time frames, what Ames calls "balance."[6] This is perhaps due to the typically small populations used in a piece of music, or because of a greater focus on method itself. Ames' work suggests a variety of ways to gain compositional control over this relationship. Statistical feedback "colors" element probabilities so that over shorter time frames, the statistics (results) more closely correspond to the specified probabilities. A scientist might call this *variance reduction*; we will analyze this in Section 2.3.

Let's return to the ten coin flips. We had seven Hs, three Ts. Using statistical feedback we can compensate so that our frame, of, say, twenty trials, is statistically better. The obvious thing to do is positively bias the probabilities of depauperate selections. For instance, we might now use for the eleventh toss $p(H) = .3$ and $p(T) = .7$, favoring the selection of a T. To paraphrase Ames, we use the preceding statistics as an input to the generating probability function.

[6] See, for example [8, 7, 5].

2 The Dissonant Counterpoint Algorithm

2.1 Informal description

Tenney often discussed his interest in "models," and his criteria were in how the brain, the ear, the human accomplished something. His theories and compositional algorithms generally placed a high priority on the success of the theory or algorithm in elucidating some cognitive process.

The algorithm described here reflects this design goal in an efficient and elegant way. A list of n values is maintained, one for each pitch element, which will be interpreted as relative selection probabilities. We initialize these to be all equal, then proceed as follows:

1. Select one element from the list randomly, using the values as relative probabilities for choosing each element
2. Set the selected element's value to zero
3. Increase the values of all the other elements in some deterministic way, for instance by adding a constant
4. Repeat (go back to step 1)

The algorithm is deceptively simple. Note that once selected, an element is temporarily removed from contention (its probability is zero). That element and all other unselected elements become more likely to be picked (their probabilities climb) on successive trials or 'time steps'[7] of the algorithm. The longer an element is *not* picked, the more likely it is that it *will* be picked.[8] Tenney's use of this algorithm is an extension and abstraction of one particular aspect of the compositional technique of Ruggles and/or Crawford Seeger, that of non-deterministic non-repetition of pitch-class or interval-class. Fig. 1 provides the simplest possible example: only pitch-classes are chosen by the algorithm. There is no explicit control of intervallic distribution (which would, of course, be of concern to Ruggles or Crawford Seeger). We give a more complex musical composition using this basic algorithm in Fig. 2.

[7] We use 'time steps' to refer to repetitions of the algorithm; note that this does not necessarily imply regular time intervals in a musical sense.

[8] Note that to fully "avoid repetitions of any tone until at least six progressions have been made" (as suggested by Seeger), an element's probability would have to remain at zero for six trials after its choice. In Section 2.2, we show how this idea is incorporated in the general description of the algorithm, using growth functions with a high power to produce a similar result.

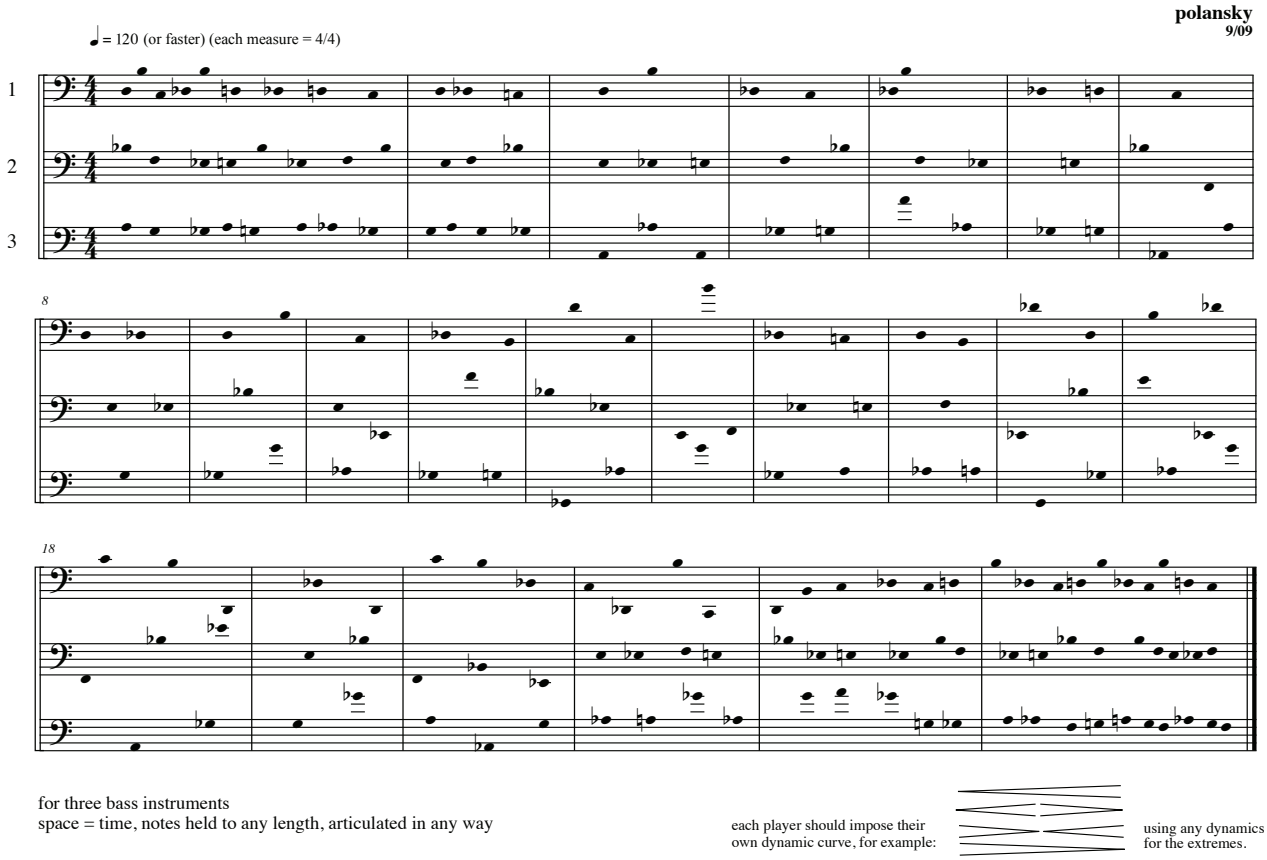


Figure 2: A short example piece, a trio for any three bass instruments. Each instrument plays one of four pitch-classes, which are selected by the linear version of the dissonant counterpoint algorithm. Durations are chosen by a narrow Gaussian function whose varying mean follows a curve which begins quickly, gets slower, and then speeds up again. A simple set of stochastic functions determine the likelihood of octave displacement for each voice.

Let's say we run the algorithm for 100 time steps, and then re-run it from the same initial values again for 100 time steps: due to the randomness in step 1 we will get different sequences (but with similar statistics). There is an important difference between sequences produced in this way and those produced by an 'un-fedback' random number generator with statistically independent (iid) trials. The statistical feedback reduces the sequence-to-sequence *fluctuations* (i.e. variance about the expectation) in the total number of occurrences of any given element over the 100 time steps, when compared to that of the un-fedback case. In Section 2.3 we will also show that, depending on the increment rule, the generated sequences exhibit different kinds of quasi-periodicity and "memory," as measured by decay of so-called *autocorrelation*, over durations much longer than the time to cycle through the n elements.

2.2 Formal description

By constructing a slightly more general model than sketched above, we open up an interesting parameter space varying from random to deterministic processes, that includes some of Tenney's work as special cases. For the remainder of Section 2 we assume more mathematical background.

For each element $i = 1, \ldots, n$ we maintain a *count* c_i describing the number of time steps since that element was chosen. We define a single *growth function* $f : \mathbb{Z}^+ \to \mathbb{R}^+$ which acts on these counts to update the relative selection probabilities following each trial. Usually we have $f(0) = 0$ (which forbids repeated elements), and f non-decreasing. We also *weight* each element $i = 1, \ldots, n$ with a fixed positive number w_i whose effect is to bias the selection probabilities towards certain elements and away from others. We now present a pseudocode which outputs a list $\{a_t\}_{t=1}^T$ containing the element chosen at each timestep $t = 1, \ldots, T$.

Dissonant Counterpoint Algorithm
input parameters: weight vector $\{w_i\}_{i=1}^n$*, growth function* f
initialize: $c_i = 1$, $i = 1, \ldots, n$
for timestep $t = 1, \ldots, T$ **do**
  compute probabilities: $p_i = \dfrac{w_i f(c_i)}{\sum_{k=1}^n w_k f(c_k)}$, $i = 1, \ldots, n$
  randomly choose j from the set $1, \ldots, n$ with probabilities $\{p_i\}_{i=1}^n$
  update counts: $c_j = 0$, and $c_i = c_i + 1$, $i = 1, \ldots, n$, $i \neq j$.
  store chosen note in the output list: $a_t = j$
end for

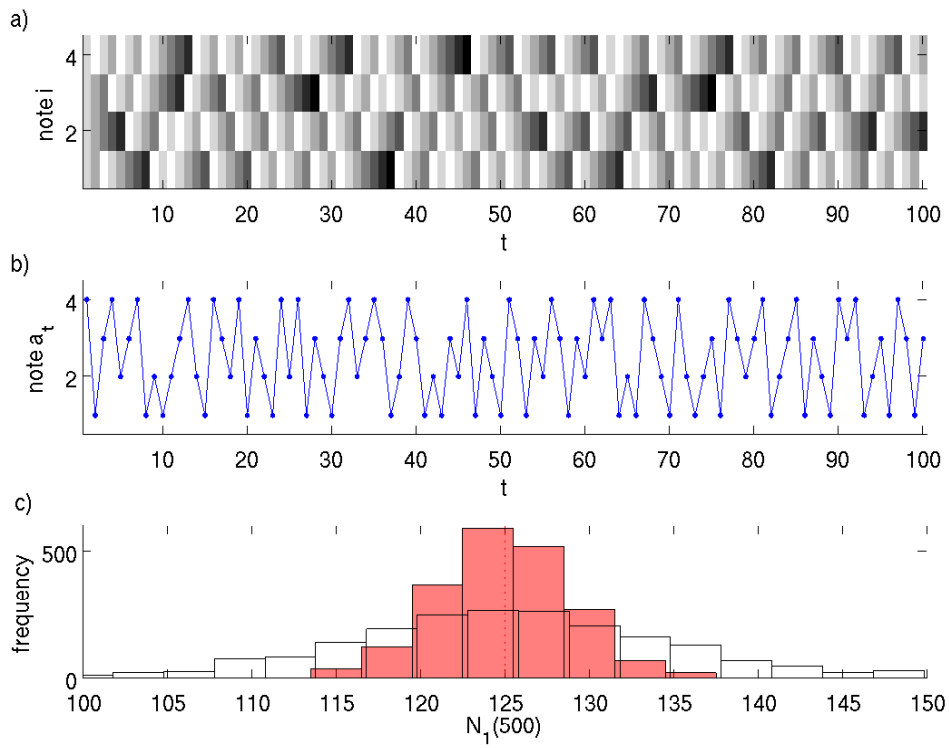


Figure 3: Simulation of the dissonant counterpoint algorithm for the simplest linear case (power $\alpha = 1$), and uniform weights $w_i = 1$ for $i = 1, \ldots, 4$. a) grayscale image showing counts c_i for each element $i = 1, \ldots, 4$ versus time t horizontally, with white indicating zero and darker larger counts. b) graph of elements selected a_t versus time t. c) distribution of $N_1(500)$, the total number of occurrences of element 1 in a run of 500 time steps, shown as a histogram over many such runs; shaded bars are for the dissonant counterpoint algorithm, white bars are for random iid element sequences with uniform distribution $p_i = 1/4$ for $i = 1, \ldots, 4$ (note the much wider histogram).

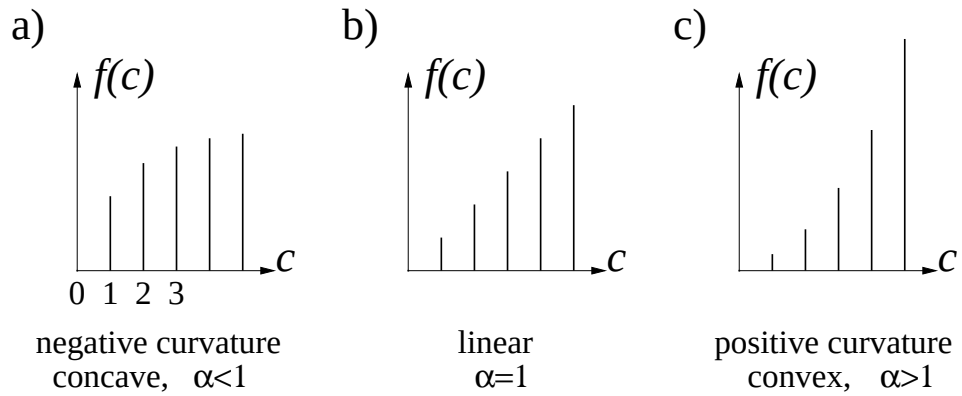


Figure 4: Illustration of three types of growth function f, depending on choice of power law α in (1).

The normalizing sum in the denominator of the expression for p_i merely ensures that the total probability is 1.

One simple but flexible form for the growth function is a *power law*,

$$f(c) = c^{\alpha} \tag{1}$$

for some power $\alpha \geq 0$. For now, we also fix equal weights $w_i = 1$ for all i. Note that $\alpha = 1$ gives the linear case where the relative selection probabilities are the counts themselves. A typical evolution of the counts c_i that form the core of the algorithm, for this linear case, is shown in Fig. 3a. A typical output sequence a_t is shown graphically in Fig. 3b (also see melody Fig. 1). A large reduction in variance of element occurrence statistics, relative to uniform random iid element sequences, is apparent in Fig. 3c.

Large powers, such as $\alpha > 5$, strongly favor choosing the notes with the largest counts, i.e. those which have not been selected for the longest time. Conversely, taking the limit $\alpha \to 0$ from above gives a process which chooses equally among the $n - 1$ notes other than the one just selected; because of its relative simplicity this version allows rigorous mathematical analysis (Section 2.4). Observe that $\alpha < 1$ leads to *concave* functions, i.e. with everywhere negative curvature (extending f to a function on the reals we would have $f'' < 0$), and that $\alpha > 1$ gives *convex* functions, positive curvature ($f'' > 0$). These cases are illustrated by Fig. 4, and their output is compared in the next section. In Fig. 5 we give a more complex musical example in which the power α, and hence the sonic and rhythmic texture, changes slowly during the piece.

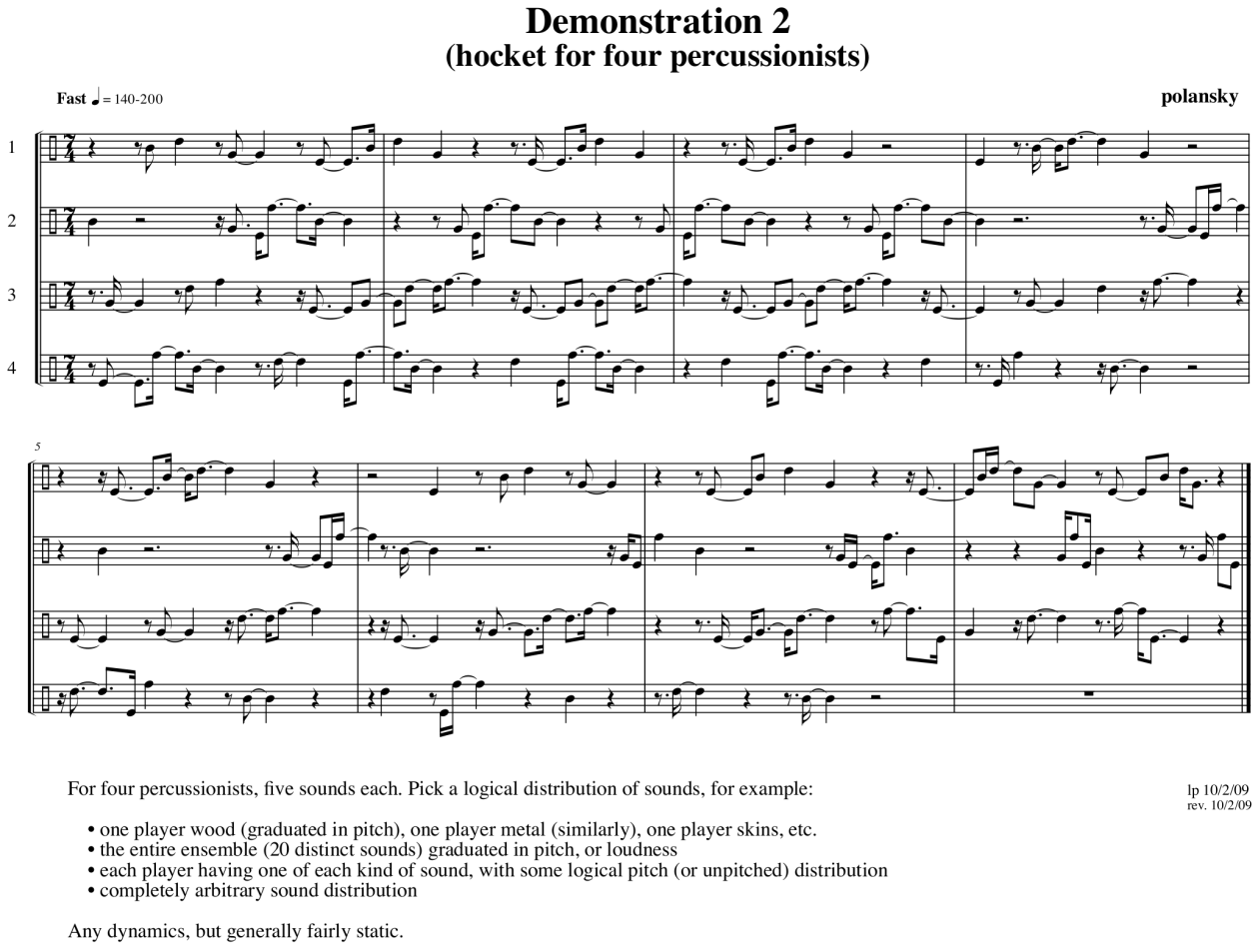


Figure 5: In this short quartet, 20 elements (percussion sounds) are selected by the algorithm, and distributed in hocket fashion to the four instruments. Durations are selected by the algorithm independently from a set of 9 distinct values. Durations and elements are selected independently, by different power functions of the form (1). The power α is interpolated over the course of the piece from 1 to some very high power (or vice versa). In the case of durations, the exponent begins high (maximum correlation) and ends at 1 (little correlation). In the case of the percussion elements, the interpolation goes in the other direction. An exponentially decreasing weight function is used for durations, favoring smaller values.

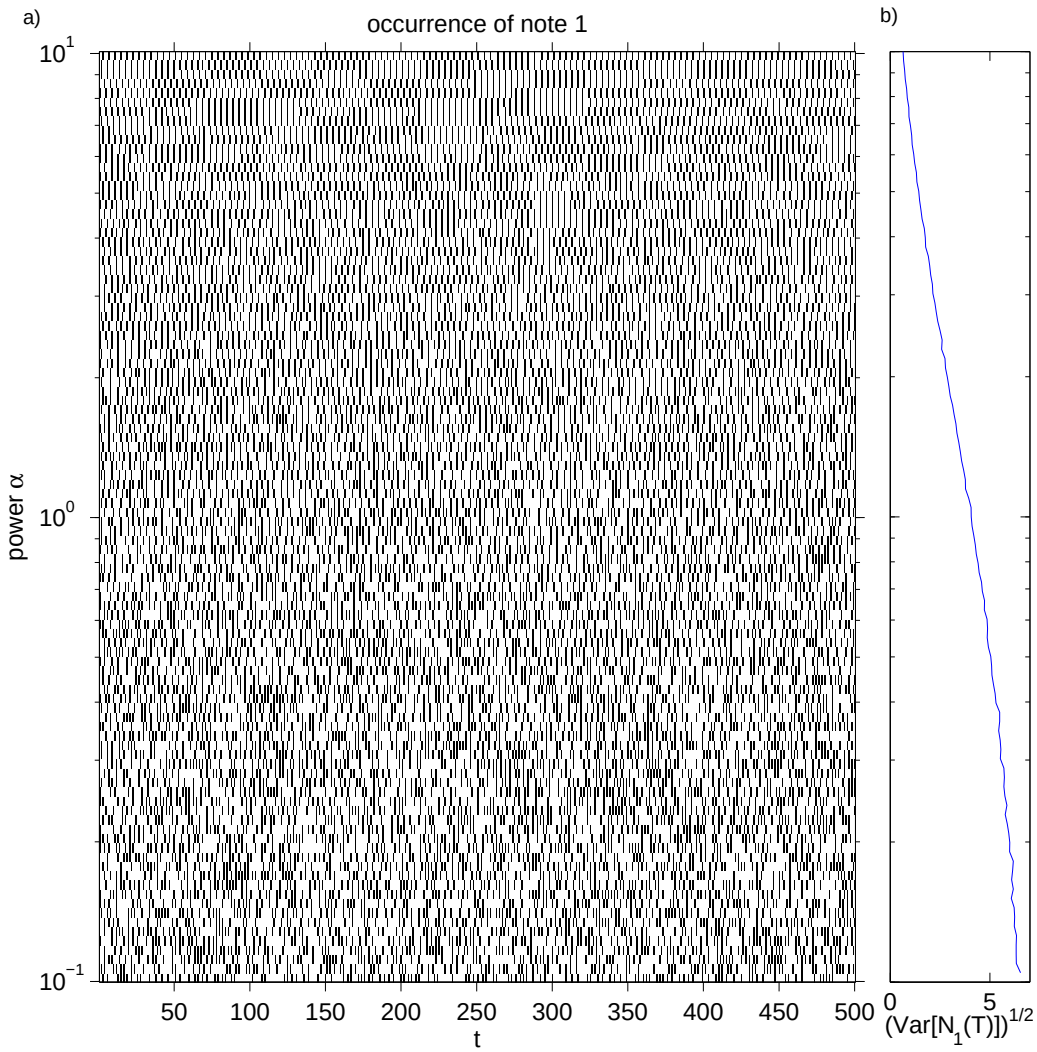


Figure 6: Dependence of periodic correlations on power law α. a) Occurrence of element 1 (black) for $n = 4$ elements, versus time $1 \leq t \leq T = 500$ horizontally, for 100 values of α spanning logarithmically the vertical direction from low (concave) to high (convex). Results are similar for the other elements. b) Standard deviation of the number of occurrences of element 1 in the run of length $T = 500$ (shown on a linear scale), as a function of α on the same vertical axis as in a).

2.3 Curvature of the growth function

In this section we investigate the effect of varying α, and introduce the autocorrelation function. We will take the weights w_i to be all equal. Figure 6a shows, for the case $n = 4$, sequences of algorithm outputs for a logarithmic range of α values. (Behaviors for other numbers of elements are qualitatively very similar.) The linear case, where $\alpha = 1$, lies half-way up the figure. Note that the larger α becomes, and consequently the more positive the curvature of the growth function, the more temporal order (repetitive quasi-periodic structure) there is in the occurrence of a given element. The sequences become highly predictable, locking into one of the $n!$ possible permutations of the n elements for a long period of time, then moving to another closely related permutation for another long period, and so forth. We may estimate[9] this typical locked time period, when it, and n and α, are all large, as $e^{\alpha/n}$. Thus, to achieve the same level of temporal order with more elements, α would need to be increased in proportion to n.

What is the effect of varying α on the resultant statistics of a particular element? The number of times the element i occurs in a time interval (run) of length T is defined as

$$N_i(T) := \#\{t : \ a_t = i, \ 1 \leq t \leq T\} \tag{2}$$

$N_i(T)$ varies for each run of the algorithm, and is therefore a *random variable*, with mean (in the case of equal weights) T/n and variance $\mathrm{Var}[N_i(T)]$. In Fig. 6b we show its standard deviation $(\mathrm{Var}[N_1(T)])^{1/2}$ (indicating fluctuation size) against α, again for $n = 4$ and $T = 500$. This was measured using many thousand runs of this length. The standard deviation is large for low powers (concave case), and decreases by roughly a factor of 10 for $\alpha = 10$ (convex case). By comparison, the standard deviation for an iid uniform random sequence is larger than any of these: since $N_i(T)$ then has a binomial distribution with $p = 1/n$, it has standard deviation $\sqrt{Tp(1-p)} = 9.68\cdots$.

The change of "rhythmic patterning" seen above can be quantified using the *autocorrelation* in time. Consider the autocorrelation of the element signal a_t, defined as,

$$C_{aa}(\tau) := \lim_{T\to\infty} \frac{1}{T} \frac{\sum_{t=1}^{T}(a_t - \bar{a})(a_{t+\tau} - \bar{a})}{\mathrm{Var}[a]} \tag{3}$$

where $\bar{a} = (n+1)/2$ is the mean, and $\mathrm{Var}[a]$ is the variance of the element signal. Fig. 7 shows $C_{aa}(\tau)$ plotted for four types of growth function, now

[9] This is done by estimating the ratio between the selection probabilities of the elements with $n-1$ vs $n-2$ counts as $[(n-1)/(n-2)]^\alpha \approx (1+1/n)^\alpha \approx e^{\alpha/n}$.

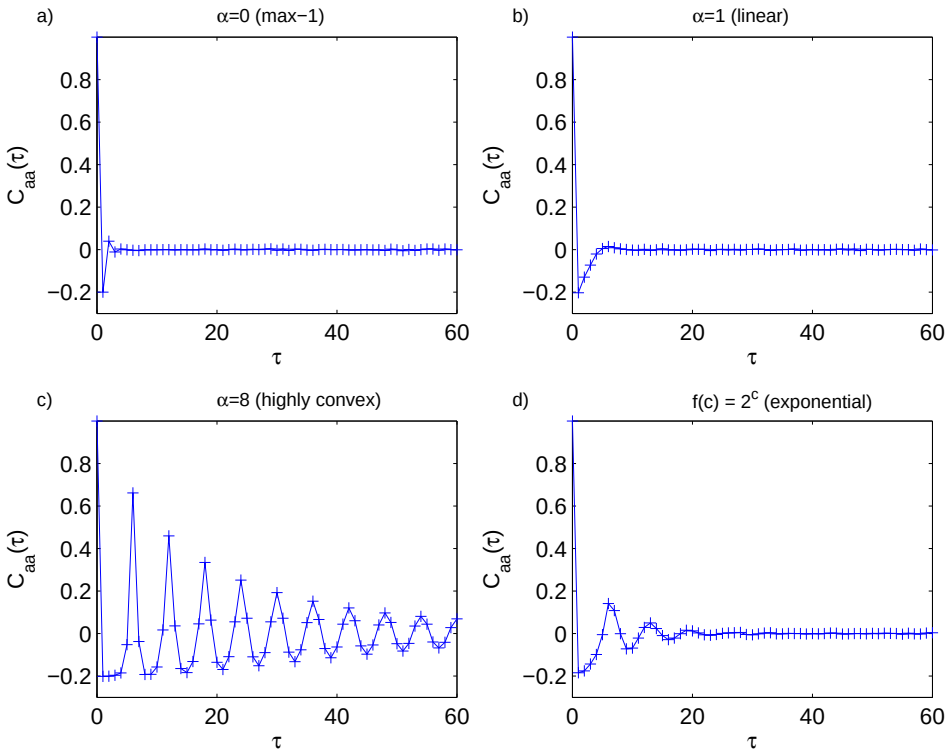


Figure 7: Autocorrelation functions measured from the dissonant counterpoint algorithm (with $n = 6$ elements, and equal weights) for four choices of growth function: three power laws, and an exponential law used in Section 3.1.

for $n = 6$. For $\alpha = 0$ we obtain the *max-1 rule* discussed in the next section: it has almost instantaneous decay of correlations, i.e. its "memory" is very short. For $\alpha = 1$, the linear model shows anti-correlation for the first few time steps, and virtually no memory beyond $\tau = n$. For high α (convex growth function), the tail of the autocorrelation extends much further (longer memory) up to $\tau = 50$ and beyond. Periodic peaks which soften with increasing τ also indicate quasi-periodic structure with a dominant period n. For this n, the exponential function (9) discussed in Section 3.1 gives correlation decay roughly similar to that of a power law model with $\alpha \approx 4$.

On a technical note, a well-known mathematical result from stochastic processes (related to Einstein's fluctuation-dissipation relation in physics [11, Eq. (2.1)]) equates the rate of growth with T of $\mathrm{Var}[N_i(T)]$ to the area under (or sum over τ of) the autocorrelation function graph.[10] It is surprising that the above results on variance of $N_1(T)$ imply that for large α the signed area under the autocorrelation graph is actually *smaller* than for small α, despite the fact that the tails extend to much *longer* times τ. It would be interesting to find an explanation for this.

2.4 Vanishing power: "max-1" version and the effect of weights

When $\alpha = 0$, the growth function (1) becomes the function

$$f(c) = \begin{cases} 0, & c = 0 \\ 1, & c = 1, 2, \ldots \end{cases} \tag{4}$$

This *max-1 rule* is the linear algorithm truncated at a maximum value of 1. As discussed above, its element statistics have a large variance and an almost instant decay in the autocorrelation (i.e. almost no memory). The algorithm chooses between all $n-1$ notes other than the current one, weighted only by their corresponding weights w_i. For convenience, in this section we assume these weights have been normalized, thus

$$\sum_{i=1}^{n} w_i = 1 \tag{5}$$

Since no account is taken of the number of *counts* each eligible note has accumulated, the algorithm becomes what is known as a *Markov chain* [28],

[10] Strictly speaking, the relevant area is $\sum_{\tau \in \mathbb{Z}} C_{ii}(\tau)$ where C_{ii} is the autocorrelation of the binary signal which is 1 when element i is selected, and 0 when any other element is chosen. We have checked that $C_{ii}(\tau)$ and $C_{aa}(\tau)$ look very similar.

with no explicit memory of anything other than the current selected element. Its n-by-n transition matrix M then has nonnegative elements

$$M_{ij} = \begin{cases} \frac{w_i}{1-w_j}, & i \neq j \\ 0, & i = j \end{cases} \tag{6}$$

which give the probability of element i being selected given that the current element is j. By using (5) one may verify the required column sum rule $\sum_i M_{ij} = 1, \forall j$.

Recall that weights w_i were included in the algorithm to give long-term bias towards various elements. So, how do the long-term frequencies of elements depend on these weights? The relationship is not trivial: frequencies are not strictly proportional to weights. The relative frequencies tend to the Markov chain's so-called *steady-state* probability distribution $\mathbf{p} := \{p_i\}_{i=1}^n$ (whose components p_i are normalized $\sum_{i=1}^n p_i = 1$), for which we can solve[11] as follows.

Theorem 1 *The max-1 rule with normalized weights $\{w_i\}_{i=1}^n$ and Markov transition matrix* (6) *has a* unique *steady state distribution given by*

$$p_i = \frac{w_i(1-w_i)}{\sum_{j=1}^n w_j(1-w_j)}, \qquad i = 1, \ldots, n \tag{7}$$

Proof: Consider a candidate distribution vector $\mathbf{v} \in \mathbb{R}^n \setminus \mathbf{0}$. We multiply $\mathbf{v}$ by a non-zero scalar to give it the more convenient weighted normalization $\sum_{i=1}^n v_i/(1-w_i) = 1$. Using this and (6) we compute the ith component of $(M - I)\mathbf{v}$ as follows,

$$\begin{aligned} (M\mathbf{v} - \mathbf{v})_i &= w_i \sum_{j \neq i} \frac{v_j}{1-w_j} - v_i \;=\; w_i \left(1 - \frac{v_i}{1-w_i}\right) - v_i \\ &= w_i - \frac{1}{1-w_i} v_i \end{aligned}$$

The condition that this vanish for all $i = 1, \ldots, n$, in other words $v_i = w_i(1-w_i)$, is equivalent to the statement that $\mathbf{v}$ is an eigenvector of M with eigenvalue 1 and therefore an (unnormalized) steady state vector. Hence the eigenvector is unique up to a scalar multiple, i.e. this eigenvalue is simple.

[11] Some intuition as to why an analytic solution is possible here is that M may be factored as the product of three simple matrices: $M = \mathrm{diag}\{w_i\}(\mathbf{1} - I)\,\mathrm{diag}\{(1-w_i)^{-1}\}$, where $\mathbf{1}$ is the n-by-n matrix with all entries 1.

Finally, normalizing (in the conventional sense) this formula for v_i gives the expression (7). $\square$

In other words, with weighted versions of the algorithm, it turns out that statistical differences in element frequencies are less pronounced than the corresponding differences in weights.

How frequently may one of the n elements occur? Even if we push one of the weights towards 1 at the expense of the others (which must then approach zero), the corresponding element may occur no more than 1/2 of the time. Intuitively, this follows since repeated elements are forbidden, so one element can be chosen at most every other timestep. Rigorously, we have the following.

Theorem 2 *Let $n \geq 3$. Then the max-1 rule with positive weights $\{w_i\}_{i=1}^n$ has a steady state distribution whose components obey $p_i < 1/2$ for $i = 1, \ldots, n$.*

Proof: Fixing i, we have, using the fact that $1 - w_j > w_i$ for all $j \neq i$,

$$\begin{aligned} \sum_{j=1}^{n} w_j(1-w_j) &= w_i(1-w_i) + \sum_{j \neq i} w_j(1-w_j) \\ &> w_i(1-w_i) + w_i \sum_{j \neq i} w_j \end{aligned}$$

The result then follows from $\sum_{j \neq i} w_j = 1 - w_i$ and Theorem 1. $\square$

For the case $n = 2$, the possibility $p_i = 1/2$ must also be allowed. The result carries over to general growth functions f with $f(0) = 0$, again for the simple reason that repetition is excluded. If, however, the "drop-down" value $f(0)$ (the value to which a selected element is reset prior to the next selection) is greater than zero, repetition becomes possible. In some of Tenney's music (such as the piece about which he first published a description of this algorithm, *Changes*), he specifies a "very small" drop-down value (see Section 3).

Finally, we note that a wide variety of complex behaviors can result from combining the convex (large-α) power law with unequal weights. This seems to result in a competition between the tendency for locked-in permutations of all n elements due to the large α, and the strong bias for heavily-weighted elements. For example, Fig. 8 illustrates a selection process among $n = 12$ elements, with weights strongly biased towards "high" elements. The

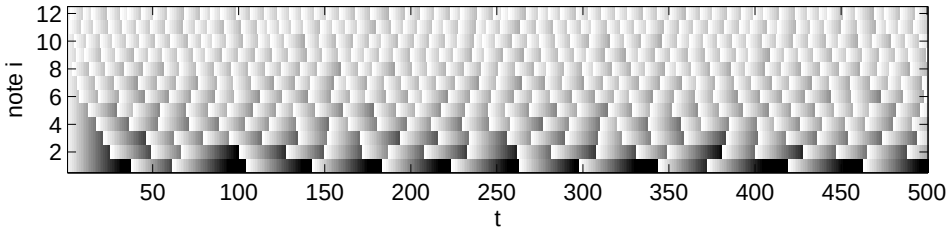


Figure 8: Grayscale image of counting function produced for the power-law with $\alpha = 6$ and $n = 12$ pitch-classes, with highly unequal weights $w_i = i^5$. $T = 500$ time steps are shown. This is discussed in Sec. 2.4.

resulting behavior consists of disordered clusters of arpeggiated sequences. It is striking that even with such extremely unequal weights (element 12 is $12^5 = 248832$ times more preferred than element 1), element 12 only occurs a few times more often than element 1.

3 Examples from Tenney's Work

Tenney's interest in the ideas of dissonant counterpoint dates back to the 1950s, as evidenced by pieces like *Seeds* (for ensemble) (1956; revised 1961) and *Monody* (for solo clarinet) (1959) [18]. These pieces, while through-composed without the use of a computer, show his nascent fascination with achieving what he later refers to—with respect to the early electronic works—as "variety":

> If I had to name a single attribute of music that has been more essential to my esthetic than any other, it would be *variety*. It was to achieve greater variety that I began to use random selection procedures in the Noise Study (more than from any philosophical interest in indeterminacy for its own sake), and the very frequent use of random number generation in all my composing programs has been to this same end. [30] (p. 40)

Tenney began using the computer for his compositions in 1961. These works, produced at Bell Laboratories, are among the first examples of computer music. Most dealt primarily with both the new possibilities of computer synthesis and his ideas of hierarchical temporal gestalt formation [38, 29]. Yet he recognized that randomly generated events without memory

of prior events would not produce the variety (Cowell: non-tautology) that he wanted.

While the early computer pieces predate a formalization of the dissonant counterpoint algorithm, the "seeds" of this idea are clear in his description of an approach to pitch selection:

> Another problem arose with this [Stochastic String] quartet which has led to changes in my thinking and my ways of working, and may be of interest here. Since my earliest instrumental music ("Seeds," in 1956), I have tended to avoid repetitions of the same pitch or any of its octaves before most of the other pitches in the scale of 12 had been sounded. This practice derives not only from Schoenberg and Webern, and 12-tone or later serial methods, but may be seen in much of the important music of the century (Varèse, Ruggles, etc).
>
> In the programs for both the Quartet and the Dialogue, steps were taken to avoid such pitch-repetitions, even though this took time, and was not always effective (involving a process of recalculation with a new random number, when such a repetition did occur, and this process could not continue indefinitely). In the quartet, a certain amount of editing was done, during transcription, to satisfy this objective when the computer had failed. [30]

Tenney continued to explore this process throughout his life, and began using the algorithm described in this paper as early as the 1980s. The first published description of it occurs in a sentence in his article on *Changes* (1985):

> Just after a pitch is chosen for an element, [the probability of] that pitch is reduced to a very small value, and then increased step by step, with the generation of each succeeding element (at any other pitch), until it is again equal to 1.0. The result of this procedure is that the immediate recurrence of a given pitch is made highly unlikely (although not impossible). [34] (p. 82).

Note that the above description seems to describe a linear model with a small positive "drop-down" value, and truncation, i.e.

$$f(c) = \min[\epsilon + ac, 1] \tag{8}$$

for $a > 0$ and some small $\epsilon > 0$. The fact that ϵ is positive allows pitches to be repeated.

From the composition of *Changes* in 1985 until 1995, Tenney wrote a number of computer-generated pieces, including *Road to Ubud* (1986; revised 2001), *Rune* (1988), *Pika-Don ("flash-boom")* (1991) and *Stream* (1991), each of which warrant further investigation with respect to the use of the dissonant counterpoint algorithm. However, many of Tenney's works after 1995 implement the dissonant counterpoint algorithm explicitly including: *Spectrum 1 – 8* (1995 – 2001); *Diaphonic Study* (1997); *Diaphonic Toccata* (1997); *Diaphonic Trio* (1997); *Seegersong #1* and *#2* (1999); *Prelude and Toccata* (2001); *To Weave (a meditation)* (2003); *Panacousticon* (2005); and *Arbor Vitae* (2006).[12]

At a certain point, the dissonant counterpoint algorithm simply became Tenney's de facto pseudo-random element chooser. He used it to determine pitches (*Seegersongs* and others), timbre/instrumentation (*Spectrum* pieces, *Panacousticon*), and register (*To Weave*), and even movement through harmonic space (*Arbor Vitae*). Early drafts of computer programs written to generate *Spectrum 6 – 8* are labeled with the word *diaphonic*. Tenney used that term to refer to most of his computer code after about 1995. Specific titles notwithstanding, he may have considered many or all of these works as "diaphonic" studies after Ruth Crawford Seeger's four studies from the early 1930s (taking their name from Charles Seeger's earlier use of the term to mean, roughly: "sounding apart" [27]). Over time, the algorithm's role seems to change from that of a principal formal determinant (as in the *Seegersongs* and *To Weave*) to an embedded, deep-level selection technique which was combined with and modulated by larger formal processes.

3.1 *Seegersongs*

Seegersong #1 and *#2* are perhaps the clearest examples of Tenney's use of the algorithm. These pieces exemplify Tenney's integration of the algorithm with larger formal concerns. Both *Seegersongs* used the convex growth function that Tenney most commonly employed in these later works, in this case by repeated doubling, thus

$$f(c) = 2^c \tag{9}$$

Seegersong #1 and *#2* explicitly model Ruth Crawford Seeger's approach to dissonant counterpoint in the avoidance of pitch-class repetition. However, they also suggest other aspects of her work, such as the technique of "phrase structure" discussed by Charles Seeger in the "Manual..." [26]

[12] For these latter pieces (unlike those between 1985–1995), the computer code is available. Other pieces may use the algorithm in some way that is not yet known.

and exemplified by Ruth Crawford Seeger in *Piano Study in Mixed Accents* (1930/31) as well as by Johanna Beyer in her two solo clarinet suites (1932) [10, 9, 22]. As with Seeger's *Piano Study in Mixed Accents*, both Tenney's *Seegersongs* consist of delimited phrases (or, in Tenney's terminology, *gestalt sequences*). Each phrase has an associated ascending or descending pitch-trajectory over some duration. This is achieved using a generalization of the dissonant counterpoint algorithm of Section 2.2, in which the weights w_i are allowed to change with time in a prescribed fashion, and therefore are labeled $w_{i,t}$. That is, the probabilities in the above algorithm are computed according to

$$p_i = \frac{w_{i,t} f(c_i)}{\sum_{k=1}^{n} w_{k,t} f(c_k)}, \quad i = 1, \ldots, n \tag{10}$$

Tenney used weights $w_{i,t}$ which decrease linearly with pitch distance from a pitch center, giving a triangularly-shaped weight vector which reaches zero a pitch distance of a tritone from the center. The center itself moves in time in a piecewise linear fashion, with each linear trajectory being a phrase. The resulting weights are shown as grayscale density in Fig. 9. The moving weights are used by the dissonant counterpoint algorithm to follow the desired registral trajectory. The interpolation points defining the linear trajectories are themselves chosen randomly within a slowly-changing pitch range illustrated by the gray region in Fig. 10b.

The large scale form of Tenney's *Seegersongs* resembles Ruth Crawford Seeger's *Piano Study in Mixed Accents*, in which the registral profile similarly ascends and then descends (Fig. 10a). However, in Tenney's re-imagining, the range's upper limit follows the positive part of a smoothly distorted cosine function peaking at the golden mean division of the piece's duration. The lower limit of the pitch range remains constant (see Fig. 10b).

3.2 The *Spectrum* pieces

The pitches in the *Spectrum* series are derived from a harmonic series with a fixed fundamental [15]. In these pieces, Tenney also used the algorithm to determine non-pitch parameters. In the *Spectrum* works that use percussion, the algorithm selects, for those instruments, from a set of pitched and unpitched sounds. When the algorithm selects a pitch that cannot be played accurately by a pitched percussion instrument, a number is returned indicating an unpitched percussion sound. In this case, "accurate" is defined as a pitch in equal-temperament that is more than 5 cents from its cognate harmonic. In the instructions, Tenney states that "numbers in place of note-heads, denote non-pitched sounds or instruments to be freely chosen by the

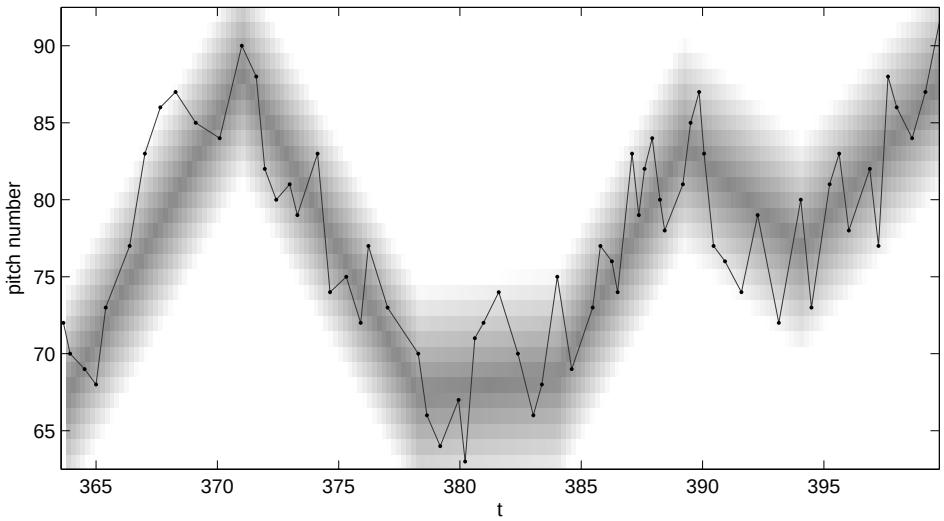


Figure 9: *Seegersong #2* excerpt showing pitch profile (solid lines) and dynamic weights $w_{i,t}$ (grey density: white is zero and darker larger positive values) used in the dissonant counterpoint algorithm (see text for how $w_{i,t}$ is generated).

player" [36].

In the *Spectrum* pieces with piano and/or harp, those fixed-pitch instruments are retuned and thus not subject to the process described above. However, because these instruments are polyphonic, the counts for more than one pitch (in the selection process) are reset to zero simultaneously. That is, all the notes for the chord are treated as having been selected. Tenney chooses the *number of chord tones* stochastically, based on a function of upper and lower density limits over time. All other parameters of the *Spectrum* pieces such as duration, loudness and pitch (which integrates the dissonant counterpoint algorithm) are determined in similar ways (for more on parametric profiles see [29]).

For the note selection algorithm, the *Spectrum* pieces use a growth function similar to that of the *Seegersongs*, but with a larger base for exponential growth:

$$f(c) = 4^c \tag{11}$$

Each part in each *Spectrum* piece is generated individually. Since the larger the base, the more convex the function (this being similar to the effect of a larger power explored in Section 2.3), note selection will tend to be

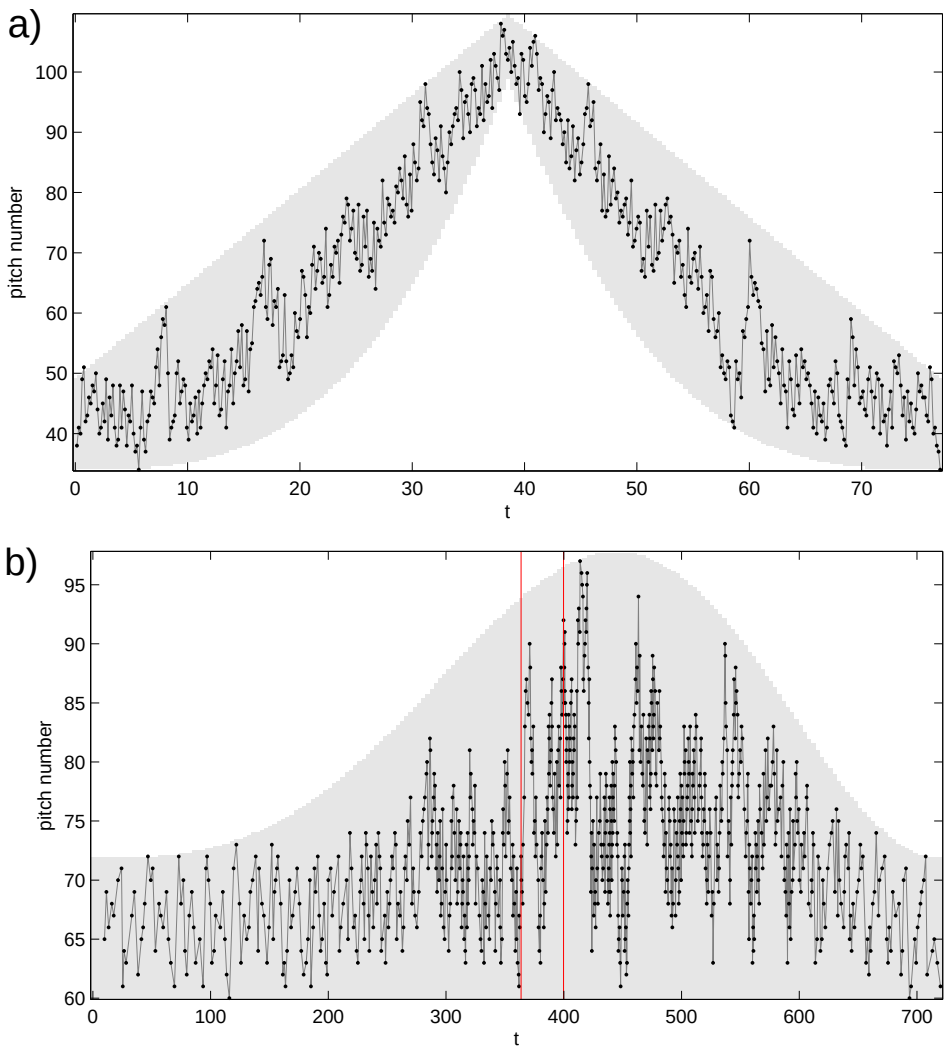


Figure 10: Pitch profiles of a) Ruth Crawford Seeger's *Piano Study in Mixed Accents* (less than a minute and a half long, about a 6 octave range), compared against b) Tenney's *Seegersong #2* (12 minutes long, about a 3 octave range). In each case time is horizontal (seconds) and pitch vertical (semitones, where 60 is middle C). The gray region shows the time-dependent pitch range used; see text for discussion of the algorithms. In b) the upper bound is proportional to $25 - 13\cos[2\pi(t/t_{\max})^{1.44}]$, and the vertical lines show the start and end of the excerpt in Fig. 9.

more correlated than in the *Seegersongs*. Perhaps, because there is a large variety of instruments in the *Spectrum* pieces, Tenney might have used the higher base in order to give individual voices a more coherent, even melodic character.

3.3 *To Weave (a meditation)*

In *To Weave (a meditation)*, for solo piano, the selection algorithm determines not only pitch-class but also register, or what can be called "voice." Three such voices are used in the piano part, with the max-1 algorithm of Section 2.4.

The algorithm determines voices note by note, where a voice is defined as one of three registers (low, middle or high). Each voice is thus a possible element for selection. For each note, the two voices not selected for the previous note become equally probable and the selected voice's probability is set to 0. In other words, if a pitch occurs in the low register, then the next pitch must occur within one of the two other registers (middle or high). The stochastic pitch sequence is *woven* "non-tautologically" into a three-voice virtual polyphony (thus the titular pun on the name of the pianist, Eve Egoyan, for whom it was written).

The growth function (2^c) is the same as the *Seegersongs*, but in *To Weave* pitch-class probabilities are incremented both globally and locally (for each individual voice). These two values are multiplied to determine the pitch-class probabilities ultimately used to select a pitch-class. Once selected, that pitch-class is placed within the range of the currently selected voice. As in the *Seegersongs*, the ranges of the three voices change over time, peaking at the golden mean point of the piece. According to Tenney:

> Waves for Eve, wave upon wave, little waves on bigger waves, et cetera, but precisely calibrated to peak at the phi-point of the golden ratio. To weave: a three-voice polyphonic texture in dissonant counterpoint, with a respectful nod in the direction of Carl Ruggles and Ruth Crawford Seeger. [37]

3.4 *Panacousticon*

In *Panacousticon* for orchestra, the algorithm selects both pitch-class and instrument. As in the *Spectrum* pieces, the pitches are derived from the harmonic series on one fundamental. Both implementations of the algorithm

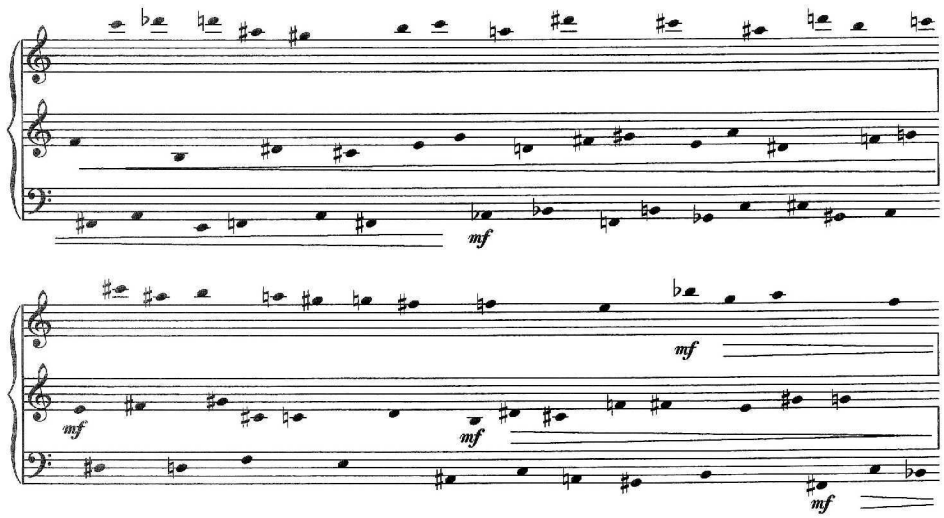


Figure 11: *To Weave (a meditation)* score excerpt.

(for pitch-class and instrumentation) use linear growth functions with an upper bound, of the form

$$f(c) = \min[c, 5] \tag{12}$$

Thus after an element is chosen, its probability reaches a maximum if it is not chosen within the next five selections.[13] For each note, the dissonant counterpoint algorithm is combined with another procedure that determines the register of the chosen pitch-class and which instruments are available to play the pitches (that is, the instruments that are not already sounding and whose range covers the determined pitch).

3.5 *Arbor Vitae*

In his last work, *Arbor Vitae* for string quartet, Tenney uses the algorithm to explore complex harmonic spaces using "harmonics of harmonics," in perhaps his most complex and unusual usage of this selection method. The pitch material, or the *harmonic space* (see, for example [33]) of *Arbor Vitae*

[13] Generally, the effect of such an upper bound on f is to equalize the probabilities of all elements not chosen within the upper bound number of time steps. This may be viewed as negative curvature as in Fig. 4, and it serves to reduce the already small autocorrelation of the resulting sequence at large times. An upper bound of 1 would be the same as a power law $\alpha = 0$, and is the max-1 rule of Section 2.4.

is complex. However, ratios are generated via reasonably simple procedures, in a manner which recalls Lou Harrison's idea of "free style" just intonation (such as in *A Phrase from Arion's Leap*, *Symfony in Free Style*, and *At the Tomb of Charles Ives*; see [16, 23, 19, 20, 17]).

The title is a metaphor for the work's harmonic structure [40]. The dissonant counterpoint algorithm first calculates *roots*, which are harmonics of a low B-flat. These are treated as temporary, *phantom* fundamentals. Next, the algorithm calculates *branches* which terminate in the pitch-classes for sounding tones. Lower harmonics are biased for root calculation by assigning *initial* probabilities (p_i proportional to $\frac{1}{\sqrt{i}}$ where i is the harmonic number of the possible root).[14] The growth function depends on harmonic number i. The selection algorithm can be summarized as that of Section 2.2 but with the probabilities computed via

$$p_i = \frac{f_i(c_i)}{\sum_{k=1}^{n} f_k(c_k)} \tag{13}$$

where $f_i(c)$ now depends both on element i and counts c, as follows,

$$f_i(c) = \begin{cases} 0, & c = 0 \\ i^{-1/2}, & c = 1 \\ i^{-1/4}, & c > 1 \end{cases} \tag{14}$$

Initial counts c_i are all set to 1, and counts are only updated for elements that have been selected at least once.

For branch selection, the set of possible elements are the primes — 3, 5, 7, 11 — of a given root. Due to the use of negative powers, the growth function becomes a kind of harmonic distance measure (with higher primes less favored). This tendency towards "consonance" is reflected in the bias towards selected roots which are closer to fundamentals, as well as in selected branches which are closer to those roots. Tenney seems to be imposing a kind of natural evolutionary "drag" on the tendency of the harmonic material in the piece to become too strongly dissassociated with its fundamentals. This ensures a kind of tonality, albeit a sophisticated and ever-changing one.

In other ways, compositional procedures of *Arbor Vitae* resemble those of *Seegersongs*, *Panacousticon*, *To Weave (a meditation)*, and the *Spectrum* pieces. Final pitch determinations use time-variant profiles of pitch ranges. Instrument selection, as in some of the other pieces, is performed by the dissonant counterpoint algorithm. As in *Panacousticon*, the software first determines whether an instrument's range can accomodate the selected pitch.

[14] In fact, rather than handling counts as in Section 2.2, Tenney *directly* updated relative probabilities p_i, normalizing them to sum to 1 whenever random selection was needed.

In contrast to *To Weave (a meditation)*, the growth function for instrument selection in *Arbor Vitae* has $f(0) = f(1) = 0$: in general no instrument may be chosen until *two* other instruments have played. For a more detailed discussion of this piece, see [5].

4 Conclusion

The dissonant counterpoint algorithm is, in some respects, just a simple method for choosing from a set of elements to give a random sequence with certain behaviors. In other respects, it is an ingenious way of marrying both an important historical style (that of Ruggles/Seeger/Cowell "dissonant counterpoint," or "diaphony") with a more modern and sophisticated, but poorly understood set of ideas from computer-aided music composition (Ames' *statistical feedback*). The algorithm elegantly embeds the latter idea in a manner characteristic of Tenney's interest in the idea of a *model* (a method that reflects how humans do or hear something).

We have analyzed the algorithm mathematically, explaining how a convex/concave choice of the growth function controls the correlation in time (rate of memory loss) of the sequence. This can lead to surprising musical effects, such as quasi-rhythmic permutations with long-range order. We included a formula explaining how weights determine the statistical selection frequencies (in the case of the max-1 growth function). We illustrated a few of the algorithm's wide variety of musical possibilities with two example compositions. Combined with our discussion of the algorithm's role in Tenney's work, this work suggests ways in which it might be used further, in experimental and musical contexts.

For example, none of the versions of the algorithm presented above consider an "ordered" set of elements in which the *proximity* of one to the other is significant. Simple examples of this are pitch-sets, registral values, durations, and so on. One might want to select from meaningful *regions* of the set of elements (e.g. "shorter durations," "higher pitches"). In a simple extension to the algorithm, one of the authors (Polansky) has implemented what he calls "gravity": not only is the chosen element's probability affected post-selection, but so are the probabilities of surrounding elements. By defining the shape of the "gravitation" (the width of the effect, the slope of the effect curves, whether the "gravity well" is negative or positive) one can increase or decrease the probabilities of neighborhood selection in a variety of ways. Polansky has used this extensively in his piece *22 Sounds* (for percussion quartet) [24].

The algorithm, in its simplest form, is meant to ensure a kind of maximal variety with a minimum amount of computation. However, by varying explicit parameters, it can produce, in both predictable and novel ways, a continuum of behaviors from completely non-deterministic to completely deterministic.

Acknowledgements

Thanks to Dan Rockmore for his important help in describing the relationship of the mathematics to the aesthetic ideas here. Kimo Johnson made many valuable suggestions, and helped work on some of the software. Thanks to Lauren Pratt and the Estate of James Tenney for making valuable material available to us. AHB is funded by NSF grant DMS-0811005.

A Matlab/Octave code examples

Here we give code that we use to simulate the dissonant counterpoint algorithm and collect statistics. It may be run with MATLAB [39] or its free alternative Octave [14]. To measure autocorrelation accurately we need many (e.g. 10^5) samples. We generate N realizations of length T simultaneously, since this is equivalent to, but (due to vectorization) much more efficient than, generating a single long realization of length NT. This also allows us to see 'vertical' correlations between different runs launched with the same initial conditions (as in Fig. 6).

```
n = 12;                        % # notes or elements
N = 500;                       % # simultaneous realizations
T = 500;                       % # timesteps
f = @(c) c.^4;                 % fourth power law growth function (convex)
w = ones(n,1);                 % selection bias weights (column vector)
a = zeros(N,T);                % histories of which notes chosen
c = ones(n,N,T+1);             % histories of counts for all notes
for t=1:T
  fc = repmat(w, [1 N]) .* f(c(:,:,t));        % feed c thru func & bias
  p = fc ./ repmat(sum(fc,1), [n 1]);          % selection probabilities
  cum = cumsum(p,1);                           % cumulative probs
  x = rand(1,N);                               % random iid in [0,1]
  a(:,t) = sum(repmat(x, [n 1]) > cum, 1) + 1; % chosen notes
  c(:,:,t+1) = c(:,:,t) + 1;                   % increment counts
```

```
  c(sub2ind(size(c), a(:,t)', 1:N, (t+1)*ones(1,N))) = 0; % reset chosen
end
figure; imagesc(a); colorbar; xlabel('t'); ylabel('run'); title('notes');
```

To compute autocorrelation of the element sequences `a` we use,

```
M = 100;                        % max correlation time to explore
ma = mean(a(:)); zma = a - ma; ca = zeros(1,M+1); l = 1:(T-M);
for t=0:M, ca(t+1) = mean(mean(zma(:,l).*zma(:,l+t))); end
figure; plot(0:M, ca./ca(1), '+-'); xlabel('\tau'), ylabel('C_a(\tau)');
```

Finally, to generate an audio file output of realization number `r` we use,

```
dt = 0.125; fs = 44100; Tsong = 60;          % Tsong: song length (sec)
fnot = 440*2.^((0:n-1)/n);                   % list of note freqs (Hz)
t = single(1:floor(fs*Tsong)-1)/fs;          % list of time ordinates
wavwrite(0.9*sin(2*pi*fnot(a(r,1+floor(t/dt))).*t)',fs,16,'out.wav');
```

References

[1] Charles Ames, *Two pieces for amplified guitar*, Interface: Journal of New Music Research **15** (1986), no. 1, 35–58.

[2] ———, *Automated composition in retrospect*, Leonardo Music Journal **20** (1987), no. 2, 169–185.

[3] ———, *Tutorial on automated composition*, Proceedings of the ICMC, International Computer Music Association. Urbana, Illinois, 1987, pp. 1–8.

[4] ———, *A catalog of statistical distributions: Techniques for transforming random, determinate and chaotic sequences*, Leonardo Music Journal **2** (1991), 55–70.

[5] ———, *Statistics and compositional balance*, Perspectives of New Music **28** (1991), no. 1, 80–111.

[6] ———, *A catalog of sequence generators*, Leonardo Music Journal **2** (1992), 55–72.

[7] ———, *Thresholds of confidence: An analysis of statistical methods for composition, part 1: Theory*, Leonardo Music Journal **5** (1995), 33–38.

[8] _____, *Thresholds of confidence: An analysis of statistical methods for composition, part 2: Applications*, Leonardo Music Journal **6** (1996), 21–26.

[9] Marguerite Boland, *Experimentation and process in the music of Johanna Beyer*, VivaVoce, Journal of the Internationaler Arbeitskreis Frau and Musik (in German) **76** (2007), `http://www.archiv-frau-musik.de`.

[10] _____, *Johanna Beyer: Suite for clarinet Ib*, Frog Peak Music (A Composers' Collective), Hanover, NH, 2007, Annotated performance edition.

[11] L. A. Bunimovich, *Existence of transport coefficients*, Encyclopaedia Math. Sci., vol. 101, pp. 145–178, Springer, 2000.

[12] Michael Casey, *HS: A symbolic programming language for computer assisted composition*, Master's thesis, Dartmouth College, 1992.

[13] Henry Cowell, *New musical resources*, Something Else Press Edition (reprinted 1969), 1930.

[14] John W. Eaton, *GNU Octave manual*, Network Theory Limited, 2002, `http://www.octave.org`.

[15] Robert Gilmore, Liner notes to *James Tenney:* Spectrum Pieces*)*, CD, New World Records 0692, 2009.

[16] Lou Harrison, *A Phrase for Arion's Leap*, Recording on *Tellus #14: Just Intonation*, (re–issued recording, 1986), 1974.

[17] Leta Miller and Fred Lieberman, *Lou Harrison: Composing a world*, Oxford University Press, 1998.

[18] Larry Polansky, *The Early Works of James Tenney*, Soundings (Peter Garland, ed.), vol. 13, Santa Fe: Soundings Press, 1983.

[19] _____, *Item: Lou Harrison as a speculative theorist*, A Lou Harrison Reader (Peter Garland, ed.), Soundings Press, 1987.

[20] _____, *Paratactical Tuning: An agenda for the future use of computers in experimental intonation*, Computer Music Journal **11** (1987), no. 1, 61–68.

[21] _____, *More on morphological mutations*, Proceedings of the ICMC, Interational Computer Music Association, San Jose, 1992, pp. 57–60.

[22] _____, Liner notes to *"Sticky Melodies": The Choral and Chamber Music of Johanna Magdalena Beyer*, CD, New World Records 80678, 2008.

[23] _____, *Three Excellent Ideas*, 2009, Talk given at the Drums Along the Pacific Festival, Cornish Institute for the Arts, Seattle, `http://eamusic.dartmouth.edu/~larry/misc_writings/talks/three_excellent_ideas/three_excellent_ideas.front.html`.

[24] _____, *22 Sounds (for percussion quartet)*, 2010.

[25] _____, *No Replacement (85 Verses for Kenneth Gaburo)*, Perspectives of New Music **33** (Winter–Summer, 1995), no. 1–2, 78–97.

[26] Charles Seeger, *Manual of dissonant counterpoint*, Studies in Musicology II: 1929-1979 (Ann Pescatello, ed.), Berkeley: University of California Press (republished 1994), 1930.

[27] _____, *On Dissonant Counterpoint*, Modern Music **7** (June–July, 1930), no. 4, 25–31.

[28] D. Stroock, *An introduction to Markov processes*, Springer, 2005.

[29] James Tenney, *Meta + Hodos*, Oakland, California: Frog Peak Music, 1964, (republished 1986).

[30] _____, *Computer Music Experiences, 1961–1964*, Electronic Music Reports **1** (1969).

[31] _____, *The chronological development of Carl Ruggles' melodic style*, Perspectives of New Music **16** (1977), no. 1, 36–69.

[32] _____, *Conlon Nancarrow's Studies for player piano*, Conlon Nancarrow: Selected Studies for Player Piano, Berkeley, California: Soundings Press, 1977.

[33] _____, *John Cage and the Theory of Harmony*, Soundings (Peter Garland, ed.), vol. 13, Santa Fe: Soundings Press, 1983.

[34] _____, *About 'Changes': Sixty-four studies for six harps*, Perspectives of New Music **25** (1987), no. 1–2, 64–87.

[35] _____, Liner notes to *James Tenney: Bridge and Flocking*, CD, hat ART CD 6193, 1996.

[36] _____, *Spectrum 8 (for solo viola and six instruments)*, Hanover, NH: Frog Peak Music (A Composers' Collective), 2001.

[37] _____, Liner notes to *Weave: Eve Egoyan*, CD, EVE0106, 2005.

[38] James Tenney and Larry Polansky, *Hierarchical temporal gestalt perception in music: a metric space model*, Journal of Music Theory **24** (1980), no. 2, 205–241.

[39] The MathWorks, Inc., *MATLAB software*, Copyright (c) 1984–2009, `http://www.mathworks.com/matlab`.

[40] Michael Winter, *On James Tenney's 'Arbor Vitae' for string quartet*, Contemporary Music Review **27** (2008), no. 1, 131–150.

Selected Discography of Cited Tenney Compositions

[41] Quatuor Bozzini, *Arbor Vitae*, CD, CQB 0806, 2008.

[42] Eve Egoyan, *Weave: Eve Egoyan*, CD, EVE0106, 2005.

[43] Maelstrom Percussion Ensemble, *Pika-don*, CD, hat(now)ART 151, 2004.

[44] Margaret Lancaster, *io*, CD, New World Records 80665, 2009.

[45] Erika Radermacher, Gertrud Schneider, Manfred Werder, and Tomas Bächli, *James Tenney: Bridge and Flocking*, hat ART CD 6193, 1996.

[46] Marc Sabat and Stephen Clark, *Music for Violin & Piano*, CD, hat(now)ART 120, 1996.

[47] James Tenney, *James Tenney: Selected Works 1961-1969*, CD, New World Records 80570, 2003.

[48] The Barton Workshop, *James Tenney: Melody, Ergodicity and Indeterminacy*, CD, Mode 185, 2007.

[49] _____, *James Tenney: Spectrum Pieces*, CD, New World Records 80692, 2009.

Journal of Computational Physics 229 (2010) 6898–6914

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

A new integral representation for quasi-periodic fields and its application to two-dimensional band structure calculations

Alex Barnett [a,*], Leslie Greengard [b]

[a] Department of Mathematics, Dartmouth College, Hanover, NH 03755, USA
[b] Courant Institute, New York University, 251 Mercer St., NY 10012, USA

ARTICLE INFO

Article history:
Received 29 January 2010
Accepted 22 May 2010
Available online 1 June 2010

Keywords:
Photonic crystal
Band structure
Eigenvalue
Periodic
Helmholtz
Maxwell
Integral equation
Bloch
Lattice

ABSTRACT

In this paper, we consider band structure calculations governed by the Helmholtz or Maxwell equations in piecewise homogeneous periodic materials. Methods based on boundary integral equations are natural in this context, since they discretize the interface alone and can achieve high order accuracy in complicated geometries. In order to handle the *quasi-periodic* conditions which are imposed on the unit cell, the free-space Green's function is typically replaced by its quasi-periodic cousin. Unfortunately, the quasi-periodic Green's function diverges for families of parameter values that correspond to resonances of the empty unit cell. Here, we bypass this problem by means of a new integral representation that relies on the free-space Green's function alone, adding auxiliary layer potentials on the boundary of the unit cell itself. An important aspect of our method is that by carefully including a few neighboring images, the densities may be kept smooth and convergence rapid. This framework results in an integral equation of the second kind, avoids spurious resonances, and achieves spectral accuracy. Because of our image structure, inclusions which intersect the unit cell walls may be handled easily and automatically. Our approach is compatible with fast-multipole acceleration, generalizes easily to three dimensions, and avoids the complication of divergent lattice sums.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

A number of problems in wave propagation require the calculation of *quasi-periodic* solutions to the governing partial differential equation in the frequency domain. For concreteness, let us consider the two-dimensional (locally isotropic) Maxwell equations in what is called TM-polarization [27,28]. In this case, the Maxwell equations reduce to a scalar Helmholtz equation

$$\Delta u(x,y) + \omega^2 \epsilon \mu u(x,y) = 0, \tag{1}$$

where ϵ and μ are the permittivity and permeability of the medium, respectively, and we have assumed a time dependence of $e^{-i\omega t}$ at frequency $\omega > 0$. Given a solution u to (1), it is straightforward to verify that the corresponding electric and magnetic fields **E**,**H** of the form

* Corresponding author. Tel.: +1 603 646 3178; fax: +1 603 646 1312.
E-mail addresses: ahb@math.dartmouth.edu (A. Barnett), greengard@cims.nyu.edu (L. Greengard).
URLs: http://www.math.dartmouth.edu/~ahb (A. Barnett), http://math.nyu.edu/faculty/greengar (L. Greengard).

0021-9991/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2010.05.029

$$\mathbf{E}(x,y,z) = \mathbf{E}(x,y) = (0,0,u(x,y))$$
$$\mathbf{H}(x,y,z) = \mathbf{H}(x,y) = \frac{1}{i\omega\mu}(u_y(x,y), -u_x(x,y), 0)$$

satisfy the full system

$$\nabla \times \mathbf{E} = i\omega\mu\mathbf{H}$$
$$\nabla \times \mathbf{H} = -i\omega\epsilon\mathbf{E}.$$

We are particularly concerned with doubly periodic materials whose refractive index $n = \sqrt{\epsilon\mu}$ is piecewise constant (Fig. 1). Such structures are typical in solid state physics, and are of particular interest at present because of the potential utility of photonic crystals, where the obstacles are dielectric inclusions with a periodicity on the scale of the wavelength of light [28]. Photonic crystals allow for the control of optical wave propagation in ways impossible in homogeneous media, and are finding a growing range of exciting applications to optical devices, filters [21], sensors, negative-index and meta-materials [36], and solar cells [7].

We assume that the crystal consists of a periodic array of obstacles (Ω_Λ) with refractive index $n \neq 1$, embedded in a background material with refractive index $n = 1$ (denoted by $\mathbb{R}^2 \setminus \overline{\Omega_\Lambda}$). We then rewrite (1) as a system of Helmholtz equations

$$(\Delta + n^2\omega^2)u = 0 \quad \text{in } \Omega_\Lambda \tag{2}$$
$$(\Delta + \omega^2)u = 0 \quad \text{in } \mathbb{R}^2 \setminus \overline{\Omega_\Lambda} \tag{3}$$

The expression $\overline{\Omega_\Lambda}$, above, is used to denote the closure of the domain Ω_Λ (the union of the domain and its boundary $\partial\Omega_\Lambda$). In this formulation, we must also specify conditions at the material interfaces. These are derived from the required continuity of the tangential components of the electric and magnetic fields across $\partial\Omega_\Lambda$ [27,28], yielding

$$u, u_n \text{ continuous across } \partial\Omega_\Lambda \tag{4}$$

where $u_n = \partial u/\partial n$ is the outward-pointing normal derivative.

The essential feature of doubly periodic microstructures in 2D (or triply periodic microstructures in 3D) is that, at each frequency, there may exist traveling wave solutions (Bloch waves) propagating in some direction defined by a vector $\mathbf{k}$.

Definition 1. *Bloch waves* are nontrivial solutions to (2)–(4), that are quasiperiodic, in the sense that

$$u(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}}\tilde{u}(\mathbf{x}), \tag{5}$$

where $\tilde{u}$ is periodic with the lattice period and $\mathbf{k} = (k_x, k_y)$ is real-valued. $\mathbf{k}$ is referred to as the *Bloch wavevector*.

Bloch waves characterize the bulk optical properties at frequency ω; they are analogous to plane waves for free space. If such waves are absent for all directions $\mathbf{k}$ for a given ω, then the material is said to have a *band-gap* [48]. The *size* of a band-gap is the length of the frequency interval $[\omega_1, \omega_2]$ in which Bloch waves are absent. Crystal structures with a large band-gap are 'optical insulators' in which defects may be used as guides [28], with the potential for enabling high-speed integrated optical computing and signal processing.

Definition 2. The *band structure* of a given crystal geometry is the set of parameter pairs (ω, $\mathbf{k}$) for which nontrivial Bloch waves exist.

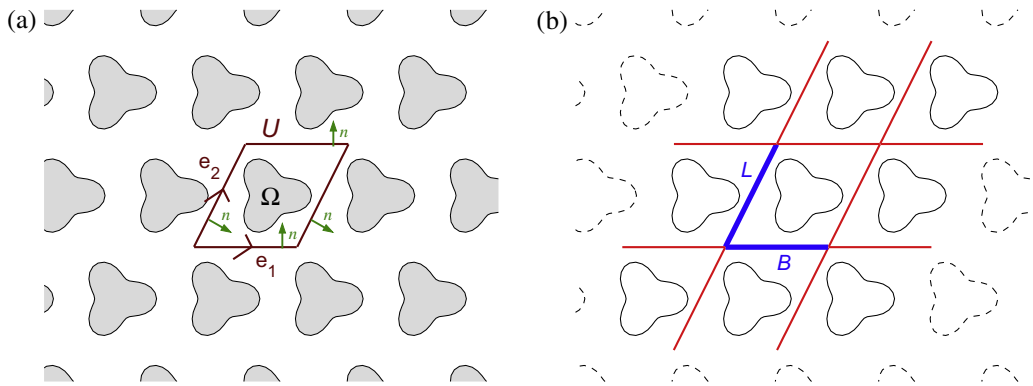


Fig. 1. (a) Problem geometry: an infinite dielectric crystal, in the case where the inclusion Ω lies within a parallelogram unit cell U. The (shaded) set of all inclusions in the lattice, denoted by Ω_Λ in the text, has refractive index n, while the white region has index 1. (b) Sketch of our quasi-periodizing scheme: we make use of layer potentials on the left (L) and bottom (B) walls, extended to the additional segments shown, which form a skewed 'tic-tac-toe' board, as well as the near neighbor images of Ω, outlined in solid lines.

The numerical prediction of band structure is a computationally challenging task, yet essential to the design and optimization of practical devices. It requires characterizing the nontrivial solutions to a homogeneous system of partial differential Eqs. (2) and (3) subject to homogeneous interface and periodicity conditions (4) and (5) in complicated geometry. Solving this eigenvalue problem is the focus of our paper.

In the next section, we briefly review existing approaches, and in Section 3, we present and test a method that relies on the quasi-periodic Green's function. We introduce our new mathematical formulation in Section 4. Numerical results are presented in Section 5, and we conclude in Section 6 with some remarks about the potential for wider application of this approach.

2. Existing approaches

In order to pose the band structure problem as an eigenvalue problem on the unit cell U (see Fig. 1), we will require some additional notation. The nonparallel vectors $\mathbf{e}_1, \mathbf{e}_2 \in \mathbb{R}^2$ define a *Bravais lattice* $\Lambda := \{m\mathbf{e}_1 + n\mathbf{e}_2 : m, n \in \mathbb{Z}\}$. Given a smooth, simply connected inclusion $\Omega \subset \mathbb{R}^2$, we may formally define the corresponding dielectric crystal by $\Omega_\Lambda := \{\Omega + \mathbf{d} : \mathbf{d} \in \Lambda\}$. As indicated above, we assume that Ω_Λ has refractive index $n \neq 1$, and that the background $\mathbb{R}^2 \setminus \overline{\Omega_\Lambda}$ has refractive index 1. For the moment, we assume that $\overline{\Omega} \subset U$ as illustrated in Fig. 1. We will discuss the case of Ω crossing ∂U in Section 5.1.

The quasi-periodicity condition (5) can be rewritten as a set of boundary conditions on the unit cell U, coupling the solution on the left (L) and right ($L + \mathbf{e}_1$) walls, as well as on the bottom (B) and top ($B + \mathbf{e}_2$) walls. More precisely, if we define

$$a := \mathbf{k} \cdot \mathbf{e}_1, \quad \alpha := e^{ia}, \quad b := \mathbf{k} \cdot \mathbf{e}_2, \quad \beta := e^{ib},$$

then quasi-periodicity is written

$$u|_{L+\mathbf{e}_1} = \alpha u|_L \tag{6}$$
$$u_n|_{L+\mathbf{e}_1} = \alpha u_n|_L \tag{7}$$
$$u|_{B+\mathbf{e}_2} = \beta u|_B \tag{8}$$
$$u_n|_{B+\mathbf{e}_2} = \beta u_n|_B, \tag{9}$$

where the normals have the senses shown in Fig. 1.

The homogeneous Eqs. (2)–(4), (6)–(9) define a partial differential equation (PDE) eigenvalue problem on the torus U. By convention, the band structure or Bloch eigenvalues are generally defined as the subset of the parameter space $\{(\omega, a, b): \omega > 0, -\pi \leqslant a < \pi, -\pi \leqslant b < \pi\}$ for which nontrivial solutions $u : U \to \mathbb{C}$ exist. The earlier definition of band structure, based on (5), allows for arbitrary values of $\mathbf{k}$. It is clear, however, that one only needs to consider a single period of $\mathbf{k}$'s projection onto $\mathbf{e}_1$, $\mathbf{e}_2$, which we have denoted by a, b, to characterize the entire set $(\omega, \mathbf{k})$ of nontrivial Bloch waves. This domain $\{(a, b): -\pi \leqslant a < \pi, -\pi \leqslant b < \pi\}$ is (essentially) what is referred to as the *Brillouin zone*.

Because the PDE is elliptic and U is compact, for each $\mathbf{k}$ there is a discrete set of eigenvalues $\{\omega_j(\mathbf{k})\}_{j=1}^{\infty}$, counting multiplicity, accumulating only at infinity. Each $\omega_j(\mathbf{k})$ is continuous in $\mathbf{k}$, so that the bands form *sheets*.

Popular numerical methods for band structure calculations are reviewed in [28]. Broadly speaking, they may be classified as either time-domain or frequency domain schemes. In the first case, an initial pulse is evolved via the full wave equation (typically using a finite difference or finite element approximation). If the simulation is sufficiently long, Fourier transformation in the time variable then reveals the full band structure. In the second case, the eigenvalue problem (2)–(4), (6)–(9) is discretized directly. Such frequency domain schemes can be further categorized as:

1. PDE-based methods, which involve discretizing the unit cell using finite difference or finite element methods [3,19,20],
2. plane-wave methods which expand the function $\tilde{u}$ in (5) as a Fourier series, and apply the partial differential operator in Fourier space [28,29],
3. semi-analytic multipole expansion methods which apply largely to cylindrical or spherical inclusions [10,43],
4. methods which use a basis of particular solutions to the PDE at a given frequency ω and enforce both interface and boundary conditions as a linear system, such as the "multiple multipole" or "transfer-matrix" method [23,46], and
5. boundary integral (boundary element) methods [49], which includes the method described here.

For a fixed $\mathbf{k}$, methods of type (1) and (2) result in large, sparse generalized eigenvalue problems whose lowest few eigenvalues approximate the first few bands $\omega_j(\mathbf{k})$. They have the advantage that they couple easily to existing robust linear algebraic techniques. PDE-based methods, however, require discretization of the entire cell in a manner that accurately resolves the geometry of the inclusion Ω. Plane-wave methods, which perform extremely well when the index of refraction n is smooth, have low order convergence when n is piecewise constant, as in the present setting. Both require a large number of degrees of freedom.

Methods of type (3), (4) or (5), on the other hand, represent the solution using specialized functions (solutions of the PDE) whose dependence on ω is nonlinear. As a result, they can be much more efficient and high order accurate, dramatically reducing the number of degrees of freedom required. Unfortunately, however, they result in a nonlinear eigenvalue problem

involving all the parameters ω, a and b, and somewhat non-standard techniques are required to find values of the parameters for which the system of equations is singular [47].

We are particularly interested in using boundary integral methods (BIEs), since they easily handle jumps in the index in complicated geometry, have a well understood mathematical foundation, and can achieve rapid convergence, limited only by the order of accuracy of the quadrature rules used. High order accuracy is important, not only because of the reduction in the size of the discretized problem, but in carrying out subsequent tasks, such as sensitivity analyses [17] through the numerical approximation of derivatives, and the computation of band slopes (group velocity), and band curvatures (group dispersion).

There is surprisingly little historical literature on using BIE for band structure calculations, although the last few years have begun to see some activity in this direction (see, for example, [49]). There is, however, an extensive literature on integral equations for *scattering* from periodic structures, which we do not seek to review here. For some recent work and additional references, see [14,42].

3. Integral equations based on the quasi-periodic Green's function

An elegant approach to designing integral representations for quasi-periodic fields involves the construction of the Green's function that imposes the desired conditions (6)–(9) exactly. We first need some definitions [16,41]. At wavenumber $\omega > 0$, the free-space Green's function for the Helmholtz equation, G is defined by $-(\Delta + \omega^2)G = \delta_{\mathbf{0}}$ where $\delta_{\mathbf{0}}$ is the Dirac delta function centered at the origin. In 2D, this yields

$$G(\mathbf{x}) = G^{(\omega)}(\mathbf{x}) = \frac{i}{4} H_0^{(1)}(\omega|\mathbf{x}|), \quad \mathbf{x} \in \mathbb{R}^2 \setminus \{\mathbf{0}\}, \tag{10}$$

where $H_0^{(1)}$ is the outgoing Hankel function of order zero. By formally summing over images of the Green's function placed on the lattice Λ, with correctly assigned phases, we get an explicit expression for the quasi-periodic Greens function

$$G_{\mathrm{QP}}(\mathbf{x}) = \sum_{\mathbf{d}\in\Lambda} e^{i\mathbf{k}\cdot\mathbf{d}} G(\mathbf{x} - \mathbf{d}) = \sum_{m,n\in\mathbb{Z}} \alpha^m \beta^n G(\mathbf{x} - m\mathbf{e}_1 - n\mathbf{e}_2). \tag{11}$$

We leave it to the reader to verify that G_{QP} does, indeed, satisfy (6)–(9). One small caveat: the series in (11) is conditionally convergent for real ω. The physically meaningful limit is taken by assuming some dissipation $\omega = \omega + i\varepsilon$ in the limit $\varepsilon \to 0^+$ (see [18] for a more detailed discussion). It will be useful to distinguish between the copy of the Green's function sitting in the unit cell U and the set of all other images. For this, we define the "regular" part of the quasi-periodic Green's function by

$$G^r_{\mathrm{QP}}(\mathbf{x}) = \sum_{\substack{m,n\in\mathbb{Z} \\ (m,n)\neq(0,0)}} \alpha^m \beta^n G(\mathbf{x} - m\mathbf{e}_1 - n\mathbf{e}_2). \tag{12}$$

This function is a smooth solution to the Helmholtz equation within U and clearly satisfies

$$G_{\mathrm{QP}}(\mathbf{x}) = G(\mathbf{x}) + G^r_{\mathrm{QP}}(\mathbf{x}). \tag{13}$$

A spectral representation also exists [9,18], built from the plane-wave eigenfunctions of the quasi-periodic torus U:

$$G_{\mathrm{QP}}(\mathbf{x}) = \frac{1}{\mathrm{Vol}(U)} \sum_{\mathbf{q}\in\Lambda^*} \frac{e^{i(\mathbf{k}+\mathbf{q})\cdot\mathbf{x}}}{|\mathbf{k}+\mathbf{q}|^2 - \omega^2}. \tag{14}$$

Here, $\Lambda^* := \{m\mathbf{r}_1 + n\mathbf{r}_2 : m, n \in \mathbb{Z}\}$ is the *reciprocal lattice* with vectors $\mathbf{r}_j$ defined by $\mathbf{e}_i \cdot \mathbf{r}_j = 2\pi\delta_{ij}$ for $i,j = 1,2$. From the denominators in (14) it is clear that G_{QP} may blow up for specific combinations of ω and $\mathbf{k}$. The quasi-periodic Green's function is, in fact, well-defined if and only if those parameters satisfy the following non-resonance condition.

Definition 3 (*Empty resonance*). A parameter set (ω, $\mathbf{k}$), equivalently (ω, a, b), is *empty resonant* if $\omega = |\mathbf{k} + \mathbf{q}|$ for some $\mathbf{q} \in \Lambda^*$, otherwise it is *empty non-resonant*.

Our terminology comes from the fact that the blow-up in G_{QP} is physically the resonance of the 'empty' unit cell U, with refractive index 1 everywhere and quasi-periodic boundary conditions. That is, G_{QP} is undefined if and only if (ω, a, b) lies on the band structure of the empty unit cell. The blow-up of the Green's function is less apparent from (11), but is manifested in the divergence of the series, even in the limit $\omega = \omega + i\varepsilon$ with $\varepsilon \to 0^+$.

It will be convenient sometimes to refer to a Green's function as a function of two variables, with $G(\mathbf{x},\mathbf{y}) := G(\mathbf{x} - \mathbf{y})$, and $G_{\mathrm{QP}}(\mathbf{x},\mathbf{y}) := G_{\mathrm{QP}}(\mathbf{x} - \mathbf{y})$. Then, for each $\mathbf{y} \in \mathbb{R}^2$, the function $G_{\mathrm{QP}}(\cdot,\mathbf{y})$ is quasi-periodic.

We now represent solutions to the PDE eigenvalue problem (2)–(4), (6)–(9) by the layer potentials,

$$u = \begin{cases} \mathcal{S}^{(n\omega)}\sigma + \mathcal{D}^{(n\omega)}\tau & \text{in } \Omega \\ \mathcal{S}^{(\omega)}_{\mathrm{QP}}\sigma + \mathcal{D}^{(\omega)}_{\mathrm{QP}}\tau & \text{in } U \setminus \overline{\Omega} \end{cases} \tag{15}$$

where the usual single and double layer densities [16] at any wavenumber $\omega > 0$ are defined by

$$(\mathcal{S}^{(\omega)}\sigma)(\mathbf{x}) = \int_{\partial\Omega} G^{(\omega)}(\mathbf{x},\mathbf{y})\sigma(\mathbf{y})ds_{\mathbf{y}} \tag{16}$$

$$(\mathcal{D}^{(\omega)}\tau)(\mathbf{x}) = \int_{\partial\Omega} \frac{\partial G^{(\omega)}}{\partial n_{\mathbf{y}}}(\mathbf{x},\mathbf{y})\tau(\mathbf{y})ds_{\mathbf{y}} \tag{17}$$

and their quasi-periodized versions are likewise

$$\left(\mathcal{S}^{(\omega)}_{\mathrm{QP}}\sigma\right)(\mathbf{x}) = \int_{\partial\Omega} G^{(\omega)}_{\mathrm{QP}}(\mathbf{x},\mathbf{y})\sigma(\mathbf{y})ds_{\mathbf{y}} \tag{18}$$

$$\left(\mathcal{D}^{(\omega)}_{\mathrm{QP}}\tau\right)(\mathbf{x}) = \int_{\partial\Omega} \frac{\partial G^{(\omega)}_{\mathrm{QP}}}{\partial n_{\mathbf{y}}}(\mathbf{x},\mathbf{y})\tau(\mathbf{y})ds_{\mathbf{y}}. \tag{19}$$

Here ds is the usual arc length measure on $\partial\Omega$, and the derivatives are with respect to the second variable in the outward surface normal direction at $\mathbf{y}$. It is clear [16] that the above four fields satisfy the Helmholtz equation at wavenumber ω in both Ω and $U \setminus \overline{\Omega}$. Note that we have chosen a *non-periodized* representation within the inclusion Ω in (15), which has some analytic advantages (see Theorem 4 and the last paragraph in the Appendix).

Since u in (15) satisfies (2), (3), (6)–(9), all that remains is to solve for densities σ, τ such that the matching conditions (4) are satisfied, which we now address.

Using superscripts + and − to denote limiting values on $\partial\Omega$, approaching from the positive and negative normal side, respectively, we use the field (15) and the standard jump relations for single and double layer potentials [16,22] to write

$$\begin{bmatrix} u^+ - u^- \\ u_n^+ - u_n^- \end{bmatrix} = \left(\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} D^{(\omega)}_{\mathrm{QP}} - D^{(n\omega)} & S^{(n\omega)} - S^{(\omega)}_{\mathrm{QP}} \\ T^{(\omega)}_{\mathrm{QP}} - T^{(n\omega)} & D^{(n\omega)*} - D^{(\omega)*}_{\mathrm{QP}} \end{bmatrix} \right) \begin{bmatrix} \tau \\ -\sigma \end{bmatrix} =: A_{\mathrm{QP}}\eta \tag{20}$$

Here I is the identity operator, while S and D are defined to be the limiting boundary integral operators (maps from $C(\partial\Omega) \to C(\partial\Omega)$) with the kernels $\mathcal{S}$ and $\mathcal{D}$ interpreted in the principal value sense. (S is actually weakly singular so the limit is already well-defined. A standard calculation [16,22] shows that D is weakly singular as well). The hypersingular operator T has the kernel $\frac{\partial^2 G(\mathbf{x},\mathbf{y})}{\partial n_{\mathbf{x}}\partial n_{\mathbf{y}}}$ and is unbounded as a map from $C(\partial\Omega) \to C(\partial\Omega)$. In these definitions, as in (16)–(19), it is implied that G inherits the appropriate superscripts and subscripts from S, D and T. Finally, * indicates the adjoint. The amounts by which the material matching conditions fail to be satisfied,

$$m := \begin{bmatrix} u^+ - u^- \\ u_n^+ - u_n^- \end{bmatrix}, \tag{21}$$

is a column vector of functions which we call the *mismatch*. We summarize the linear system (20) by $m = A_{\mathrm{QP}}\eta$ where $\eta := [\tau; -\sigma]$. It is important to note that the *difference* of hypersingular kernels, $T^{(\omega)}_{\mathrm{QP}} - T^{(n\omega)}$, in (20) is only weakly singular [16, Sec. 3.8]. This cancellation, achieved here by using the same pair of densities inside as outside the inclusion, is well known [44]. The result is that A_{QP} is a compact perturbation of the identity and (20) is a Fredholm system of integral equations of the second kind.

In the above scheme, we might hope that if it is possible to find nontrivial densities η whose field u gives zero mismatch m for a set of parameters (ω, a, b), then that set is a Bloch eigenvalue. Indeed (as with the case of simpler domain eigenvalue problems [39, Sec. 8]) we have a stronger result.

Theorem 4. *Let (ω, a, b) be empty non-resonant. Then (ω, a, b) is a Bloch eigenvalue if and only if Null $A_{QP} \neq \{0\}$.*

The proof occupies Appendix A. This suggests the core of a numerical scheme: at each of a sampling (e.g. a grid) of parameters (ω, a, b), find the lowest singular value $\sigma_{\min}(\widetilde{A}_{\mathrm{QP}})$ of a matrix discretization $\widetilde{A}_{\mathrm{QP}}$ of A_{QP}. The band structure will then be found where $\sigma_{\min}(\widetilde{A}_{\mathrm{QP}})$ is close to zero.

3.1. Discretization of the integral operators

Since the goal of this work is to explore periodization, we limit ourselves to the simplest case of $\partial\Omega$ being smooth. The methods of this paper extend without much effort to other shapes, but the quadrature issues become more involved. Recalling (13), note that the kernels in (20) are the sum of a component due to G which is weakly singular, plus the remainder due to G^r_{QP} which is smooth (analytic). We will make use of a Nyström discretization using the spectral quadrature scheme of Kress [31] for G and the trapezoidal rule for G^r_{QP}.

We first remind the reader of the periodic trapezoidal Nyström scheme [33], in the context of a general second kind boundary integral equation

$$\mu(\mathbf{x}) + \int_{\partial\Omega} k(\mathbf{x},\mathbf{y})\mu(\mathbf{y})ds_{\mathbf{y}} = f(\mathbf{x}), \quad \mathbf{x} \in \partial\Omega,$$

where $\partial\Omega$ is parametrized by the 2π-periodic analytic function $\mathbf{z} : [0, 2\pi) \to \mathbb{R}^2$. Changing variable gives

$$\mu(s) + \int_0^{2\pi} K(s,t)\mu(t)dt = f(s), \quad s \in [0, 2\pi),$$

where $K(s,t) := k(\mathbf{z}(s), \mathbf{z}(t))|\mathbf{z}'(t)|$ and $\mathbf{z}' = d\mathbf{z}/dt$. Choosing N quadrature points $t_j = 2\pi j/N$ with equal weights $2\pi/N$ gives the N-by-N linear system for the unknowns $\mu_j^{(N)}$, which approximate the exact values $\mu(t_j)$, as

$$\mu_k^{(N)} + \frac{2\pi}{N}\sum_{j=1}^{N} K(t_k, t_j)\mu_j^{(N)} = f(t_k), \quad k = 1, \ldots, N. \tag{22}$$

By Anselone's theory of collectively compact operators [33], the convergence of errors $\left|\mu_j^{(N)} - \mu(t_j)\right|$ inherits the order of the quadrature scheme applied to the exact integrand $K(s,\cdot)\mu$, which is analytic when k and f are.

Remark 5. For analytic integrands, the periodic trapezoidal rule has exponential convergence with error $O(e^{-2\gamma N})$ where γ is the smallest distance from the real axis of any singularity in the analytic continuation of the integrand. [33, Thm. 12.6].

The above discretization is used to populate the matrix entries in (20) that are due to the smooth compoment G_{QP}^r. (We explain how to compute this kernel itself in Section 3.2.)

For non-smooth kernels, such as G, the rule (22) must be replaced by a quadrature that correctly accounts for the singularity in order to retain high order accuracy. There are a variety of such schemes, such as those of [2,24,30]. By fixing the order of accuracy, they allow for straightforward coupling to fast multipole acceleration [12–14,42] by making local modifications of a simple underlying quadrature rule (such as the trapezoidal rule or a composite Gaussian rule). In the present context, we ignore such considerations and use a global rule due to Kress [31] that achieves spectral accuracy in the logarithmically singular case.

The essential idea of Kress' scheme (after transformation of variables to the interval $[0,2\pi]$) is to split a logarithmically singular kernel $K(s,t)$ in the form

$$K(s,t) = \log\left(4\sin^2\frac{s-t}{2}\right)K_1(s,t) + K_2(s,t) \tag{23}$$

with K_1 and K_2 periodic and analytic. K_2 is (again) handled with the trapezoidal rule. For K_1, the Kussmaul–Martensen quadrature rule is spectrally accurate:

$$\int_0^{2\pi} \log\left(4\sin^2\frac{s-t}{2}\right)g(t)dt \approx \sum_{j=1}^{N} R_j^{(N)}(s)g(t_j) \tag{24}$$

with quadrature weights (deriving from the Fourier series of the log factor) given by

$$R_j^{(N)}(s) = -\sum_{m=1}^{N/2-1}\frac{2}{m}\cos m(s-t_j) - \frac{2}{N}\cos\frac{N}{2}(s-t_j). \tag{25}$$

Thus, the matrix elements in discretizing (23) are $K(t_k,t_j) = R_{|j-k|}^{(N)}(0)K_1(t_k,t_j) + K_2(t_k,t_j)$. Finally, it is always the difference of two hypersingular operators T that appears in the integral Eq. (20). This difference is only logarithmically singular, so that Kress' rule can be used for every block of (20). We refer the reader to [31] for further details.

In summary, a matrix discretization $\widehat{A}_{\mathrm{QP}}$ of A_{QP} is formed by using the above quadrature rules for each of the 2-by-2 integral operator blocks in (20). This matrix maps density values to field values. However, in order to create a matrix whose singular values approximate those of A_{QP} we must instead normalize such that $2N$-dimensional Euclidean 2-norms correctly approximate $L^2(\partial\Omega)$-norms. This is done by symmetrizing using quadrature weights to give our final matrix

$$\widetilde{A}_{\mathrm{QP}} = W^{1/2}\widehat{A}_{\mathrm{QP}}W^{-1/2} \tag{26}$$

where W is diagonal with diagonal elements $w_j = w_{j+N} = (2\pi/N)|\mathbf{z}'(t_j)|$, for $j = 1,\ldots,N$.

The net result of the preceding discussion is that with the use of specialized quadratures on smooth boundaries, the singular values of $\widetilde{A}_{\mathrm{QP}}$ are spectrally accurate approximations to those of A_{QP}. We demonstrate this convergence for a small trefoil-shaped inclusion in Fig. 2(a); the convergence is spectral, until the error is approximately machine precision times the matrix 2-norm. The rate appears to be faster at a Bloch eigenvalue (in this case on the fourth band) than far from one. Fig. 2(b) shows that the minimum locates the parameter b to 14 digit accuracy for $N \geqslant 70$.

3.2. New method for evaluation of the quasi-periodic Greens function

In order to compute the elements of $\widetilde{A}_{\mathrm{QP}}$, one must evaluate G_{QP}^r defined by (12); in this section, we present a surprisingly simple (and apparently new) method for this. Since the sums (11) and (14) converge too slowly to be numerically useful, many sophisticated schemes have been devised. Some of these are based on the Fourier representation (such as [9]), but most are based on the observation that

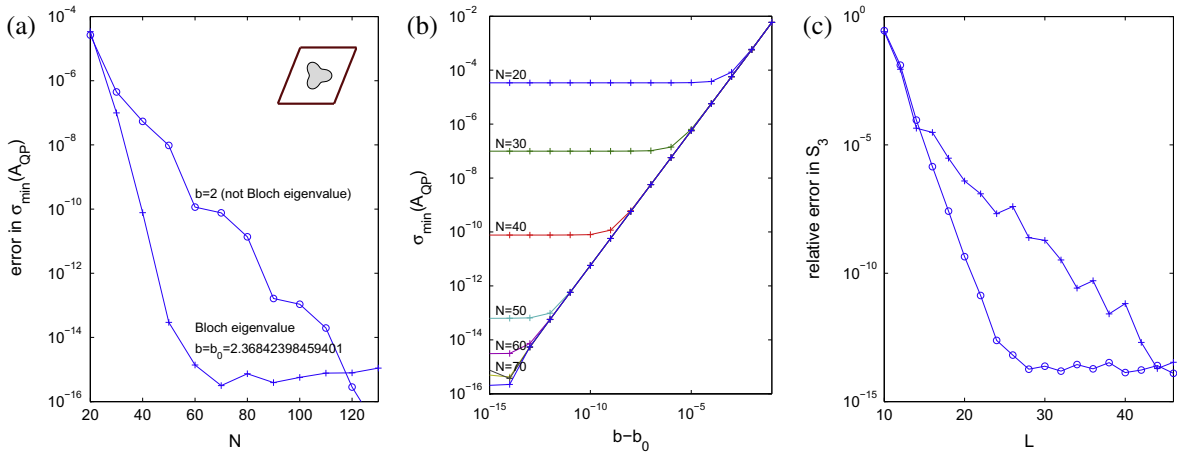


Fig. 2. Convergence for quasi-periodic Greens function scheme of Section 3. (a) Absolute error in $\sigma_{\min}(\widetilde{A}_{QP})$ vs. N the number of quadrature nodes on $\partial\Omega$, for Bloch parameters $a = \pi/2$ and the two different b values labeled. The unit cell with $\mathbf{e}_1 = (1,0)$, $\mathbf{e}_2 = (0.4,1)$, and inclusion, described by the radial function $r(\theta) = 0.2(1 + 0.3\cos 3\theta)$, are shown in the inset. Index is $n = 3$ and frequency $\omega = 4.5$. For $b = 2$, error is taken relative to the converged value 0.01879908530381247; for $b = b_0$, relative to 0. The matrix $\widetilde{A}_{QP}$ has 2-norm of about 25. (b) $\sigma_{\min}(\widetilde{A}_{QP})$ vs. difference in parameter b from the Bloch eigenvalue b_0, for several different numbers of quadrature points N. Note the horizontal log scale. This shows that it is the convergence rate at the Bloch eigenvalue that controls the accuracy with which the minimum can be found. (c) Relative error (+ symbols) in evaluation of lattice sum S_3 by the method of Section 3.2 vs. the maximum order L in (27). Parameters are as in Table 2 of [38], whose claim $S_3 = 2.13097899279352 + 5.66537068305984i$ is taken as the true value. Also, relative error (○ symbols) for $\widetilde{S}_3$ which excludes the 3×3 block of neighbors (parameters are the same; true value is taken as the converged value at $L = 50$).

$$G^r_{QP}(r,\theta) = \sum_{l=-L}^{L} S_l J_l(\omega r) e^{il\theta}, \tag{27}$$

where (r, θ) are the usual polar coordinates, and J_l the regular Bessel function of order l. As $L \to \infty$, this expression is uniformly convergent in the unit cell U, as long as there exists a circle about the origin which contains $\overline{U}$ but encloses no points in $\Lambda\backslash\{\mathbf{0}\}$. The coefficients S_l in this expansion are know as *lattice sums*, given by

$$S_l = \sum_{\substack{m,n\in\mathbb{Z}\\(m,n)\neq(0,0)}} \alpha^m \beta^n H^{(1)}_l(\omega r_{mn}) e^{-il\theta_{mn}},$$

where (r_{mn}, θ_{mn}) are the polar coordinates of $m\mathbf{e}_1 + n\mathbf{e}_2$, and $H^{(1)}_l$ is the outgoing Hankel function of order l. Thus, the issue of evaluating G^r_{QP} has been reduced to that of tabulating the lattice sums. This problem itself has a substantial literature (see, for example, [15,18,34,38,40]). Nevertheless, very few papers discuss the problem of empty resonances, at which point the lattice sums S_l blow up. One notable exception is the work of Linton and Thompson [35], who analyze this blowup for periodic one-dimensional arrays in two-dimensional scattering. They also propose a regularization method to overcome it.

We present here the construction of a small linear system whose solution yields the lattice sums rather easily (away from empty resonances). In physical terms, we compute the field induced by the free-space Green's function G, determine how it fails to satisfy quasi-periodicity, and use the representation (27) to enforce quasi-periodicity numerically. More precisely, given a field u, we define the *discrepancy* by

$$d = \begin{bmatrix} f \\ f' \\ g \\ g' \end{bmatrix} := \begin{bmatrix} u|_L - \alpha^{-1} u|_{L+\mathbf{e}_1} \\ u_n|_L - \alpha^{-1} u_n|_{L+\mathbf{e}_1} \\ u|_B - \beta^{-1} u|_{B+\mathbf{e}_2} \\ u_n|_B - \beta^{-1} u_n|_{B+\mathbf{e}_2} \end{bmatrix}. \tag{28}$$

We can interpret f, f' as functions on wall L and g, g' as functions on wall B. We construct a $4M$-component column vector $\boldsymbol{d}$ by sampling these four functions at Gaussian quadrature points $\left\{\mathbf{y}^{(L)}_m\right\}^M_{m=1}$ on L, and $\left\{\mathbf{y}^{(B)}_m\right\}^M_{m=1}$ on B. If we let the field $u(\mathbf{x}) = G(\mathbf{x})$, then for $m = 1,\ldots,M$, the mth element of $\boldsymbol{d}$ is $G\left(\mathbf{y}^{(L)}_m\right) - \alpha^{-1} G\left(\mathbf{y}^{(L)}_m + \mathbf{e}_1\right)$. The remaining $3M$ entries in $\boldsymbol{d}$ are computed in the analogous fashion.

Now let H be a (complex) matrix of size $4M \times (2L + 1)$, defined as follows. For $l = -L,\ldots,L$, fill the $(l + L + 1)$ th column in the same manner as $\boldsymbol{d}$, but using the field $u(\mathbf{x}) = J_l(\omega r)e^{il\theta}$. Letting $\boldsymbol{s} := \{S_l\}^L_{l=-L}$, it is straightforward to verify that the linear system

$$H\boldsymbol{s} = -\boldsymbol{d} \tag{29}$$

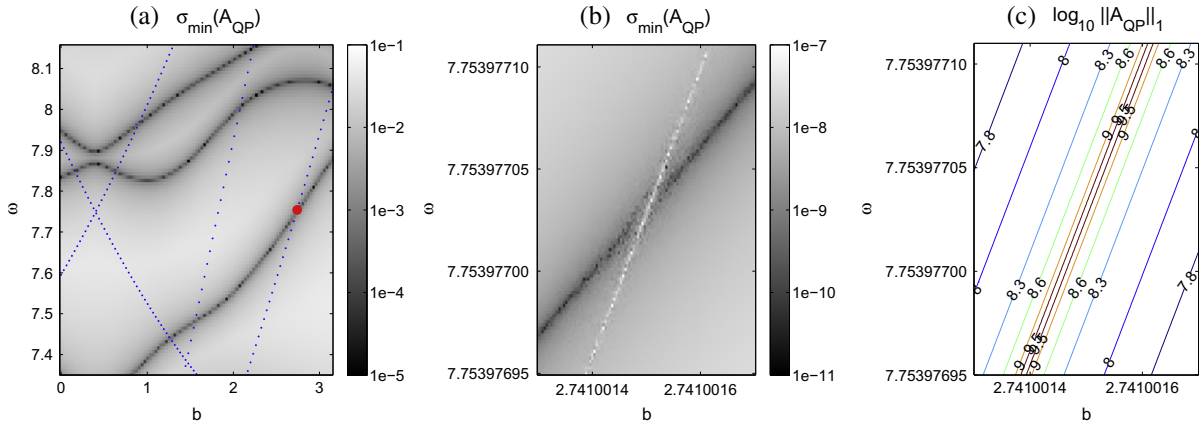


Fig. 3. Breakdown of quasi-periodic Greens function scheme, for the system of Fig. 2a except with $\mathbf{e}_2 = (0.5,1)$. (a) Minimum singular value of $\widetilde{A}_{QP}$ vs. $b = \mathbf{k}\cdot\mathbf{e}_2$ and ω, as a log density plot over a slice with fixed $a = 0.8$. Dark curves indicate the band structure, and superimposed dotted lines the 'empty' band structure where G_{QP} blows up. (b) Zooming in by a factor of 10^7 to the region shown by the dot in (a), showing failure to resolve band structure at the intersection. (c) log 1-norm of the matrix $\widetilde{A}_{QP}$ plotted over the same region as (b); it is of order the inverse of the distance to the empty band structure.

yields values for the lattice sums that annihilate the discrepancy induced by the source G. We solve the linear system in the least squares sense. This has to be done with some care, since the Bessel functions J_l become exponentially small for large l. A simple fix is to right-precondition the system by scaling the $(l + L + 1)$ th column of H by the factor $\rho_l := 1/J_l(\min[\omega R, l])$, where $R := \max_{\mathbf{x}\in U}|\mathbf{x}|$ is the unit cell radius. The entire procedure may be interpreted as finding the representation (27) which minimizes the L^2-norm of the discrepancy of the resulting G_{QP}.

Fig. 2(b) shows that the error in evaluating S_l, for $l = 3$, has exponential convergence in L. We fixed $M = 24$ (large enough that further increase had no effect). Fourteen digits of relative accuracy are achieved for $L \geqslant 46$, comparable in accuracy to [38]. Although the maximum achievable accuracy for S_l deteriorates exponentially as $|l|$ increases, the resulting accuracy of G^r_{QP} computed via (27) is close to 14 digits everywhere in U.[1] We do not claim that our method is optimal in terms of speed (although at 0.05 s to solve for all S_l values, it is adequate), merely that it is accurate, convenient and robust. To our knowledge it has not been proposed in the literature.

The convergence rate in the boundary L^2-norm of expansions such as (27) depends on the (conformal) distance from the domain to the nearest field singularity (a result of Vekua's theory and approximation in the complex plane [8, Ch. 6]). Thus, the rate may be improved by increasing this distance by removing the rest of the 3×3 block of nearest neighbors from the lattice sum, and representing

$$\widetilde{G}^r_{QP}(\mathbf{x}) := \sum_{(j,k)\in\mathbb{Z}^2\setminus\{-1,0,1\}^2} \alpha^j\beta^k G(\mathbf{x} - j\mathbf{e}_1 - k\mathbf{e}_2) = \sum_{l=-L}^{L} \widetilde{S}_l J_l(\omega r)e^{il\theta}. \tag{30}$$

To solve for $\{\widetilde{S}_l\}$, the right-hand side of the linear system is now chosen to be the direct summation of these neighbors, $u(\mathbf{x}) = \widetilde{G}(\mathbf{x}) := \sum_{j,k\in\{-1,0,1\}} \alpha^j\beta^k G(\mathbf{x} - j\mathbf{e}_1 - k\mathbf{e}_2)$. We may then evaluate $G_{QP} = \widetilde{G} + \widetilde{G}^r_{QP}$. As Fig. 2(c) shows, the convergence rate for $\widetilde{S}_l$, and hence for G_{QP}, is now a factor 2–3 better. Hence we use this method below, fixing $L = 30$.

3.3. The empty resonance problem

Given a photonic crystal (inclusion Ω with index n), using the methods of Sections 3.1 and 3.2 we are able to construct the matrix $\widetilde{A}_{QP}$ for any given frequency and Bloch parameters (ω, a, b). Fig. 3(a) shows the minumum singular value of this matrix as a function over the (b, ω) plane, for constant a: the band structure is visible as the zeros of this function. We have also superimposed the band structure of the empty unit cell (dotted lines). Theorem 4 guarantees that, away from the empty unit cell band structure, no spurious modes will be found, and that no modes are missed.

However, zooming into one of the many intersections of the two sets of curves (Fig. 3(b)), we see that in the neighborhood of the empty band structure, the desired singular values take on arbitrary fluctuating values that obscure the theoretical behavior near their intersection. This prevents any attempt to locate the desired zero set to an accuracy better than $O(\sqrt{\varepsilon_{mach}})$, where ε_{mach} is the machine precision. As Fig. 3(c) shows, this is explained by the blowup of the entries of the matrix $\widetilde{A}_{QP}$ as one approaches the empty band structure. This, in turn, causes unbounded roundoff error when computing small singular values in finite-precision arithmetic.

[1] This is to be expected from arguments similar to [4, Eq. (5)]: the residual of the linear system, around 10^{-14}, approximates the boundary error norm, which in turn controls the interior error norm when using a basis of particular solutions to the Helmholtz equation.

Remark 6. The above demonstrates a fundamental flaw inherent in the use of the quasi-periodic Greens function in band structure problems; there are empty-resonant parameter sets (sheets in the space (ω, a, b)) where the desired band structure cannot be computed. Furthermore, loss of accuracy is inevitable near these parameter sets.

This motivates the development of a more robust scheme.

4. Periodizing using auxiliary densities on the unit cell walls

4.1. Inclusion images and a new linear system

Section 3.2 illustrated the fact, well known in the fast multipole literature [6,14,12,13,18], that summing the nearest neighbors directly (i.e. excluding them from the quasi-periodic field representation) results in much improved convergence rates. This motivates defining generalizations of (16) and (17) that include summation over the appropriately phased 3×3 nearest neighbor images, as shown in Fig. 1(b),

$$(\widetilde{\mathcal{S}}^{(\omega)}\sigma)(\mathbf{x}) = \int_{\partial\Omega} \sum_{j,k\in\{-1,0,1\}} \alpha^j \beta^k G^{(\omega)}(\mathbf{x}, \mathbf{y} + j\mathbf{e}_1 + k\mathbf{e}_2)\sigma(\mathbf{y}) ds_{\mathbf{y}} \tag{31}$$

$$(\widetilde{\mathcal{D}}^{(\omega)}\tau)(\mathbf{x}) = \int_{\partial\Omega} \sum_{j,k\in\{-1,0,1\}} \alpha^j \beta^k \frac{\partial G^{(\omega)}}{\partial n_{\mathbf{y}}}(\mathbf{x}, \mathbf{y} + j\mathbf{e}_1 + k\mathbf{e}_2)\tau(\mathbf{y}) ds_{\mathbf{y}} \tag{32}$$

We now choose a layer potential representation for u that involves only free space kernels:

$$u = \begin{cases} \mathcal{S}^{(n\omega)}\sigma + \mathcal{D}^{(n\omega)}\tau & \text{in}\Omega \\ \widetilde{\mathcal{S}}^{(\omega)}\sigma + \widetilde{\mathcal{D}}^{(\omega)}\tau + u_{\mathrm{QP}}[\xi] & \text{in}U \setminus \overline{\Omega} \end{cases} \tag{33}$$

The auxiliary field u_{QP} will be represented by a new set of layer potentials that lie on the "tic-tac-toe" stencil of Fig. 1(b), consisting of the boundary of U and its closest extensions, none lying in the interior of U. We will return to this in Section 4.2. For the moment, let us denote the unknown densities that determine u_{QP} by ξ. By construction, the representation (33) satisfies (2) and (3) in U, so that it remains only to impose both the matching/continuity conditions (4) and quasi-periodicity (6)–(9). Imposing the mismatch m defined by (21) and the discrepancy d defined by (28) on u, the unknowns in (33) must satisfy a linear system of the form:

$$E\begin{bmatrix}\eta \\ \xi\end{bmatrix} := \begin{bmatrix} A & B \\ C & Q \end{bmatrix}\begin{bmatrix}\eta \\ \xi\end{bmatrix} = \begin{bmatrix} m \\ d \end{bmatrix}, \tag{34}$$

where, as before, $\eta := [\tau; -\sigma]$, We will describe the operators A, B, C, and Q in more detail shortly. For the moment, note that if there exists a density $[\eta; \xi]$ which generates a nontrivial field with vanishing mismatch and discrepancy, then it is a solution to (2), (3), (4), (7)–(9) and the corresponding parameters (ω, a, b) must be a Bloch eigenvalue. Numerical evidence supports the following stronger claim, the analog of Theorem 4.

Conjecture 7. (ω, a, b) is a Bloch eigenvalue if and only if Null $E \neq \{0\}$.

This suggests, as in Section 3, computing the band structure by locating the parameter families where (a discretization of) E is singular. The point of the new scheme is that it should be robust; if the conjecture holds, then (in contrast to the quasi-periodic Green's function approach), there will be no spurious parameter values where the method breaks down.

To discuss the operators in E, we need some additional notation. We assume that the wavenumber ω and quasiperiodicity parameters (a, b) are given. Let W be a curve in $\mathbb{R}^2$ on which single and double layer densities are defined, with the corresponding potentials written as

$$(\mathcal{S}_W\sigma)(\mathbf{x}) = \int_W G(\mathbf{x}, \mathbf{y})\sigma(\mathbf{y}) ds_{\mathbf{y}} \tag{35}$$

$$(\mathcal{D}_W\tau)(\mathbf{x}) = \int_W \frac{\partial G}{\partial n_{\mathbf{y}}}(\mathbf{x}, \mathbf{y})\tau(\mathbf{y}) ds_{\mathbf{y}}. \tag{36}$$

Letting V be a (possibly distinct) target curve in $\mathbb{R}^2$, we define the operators

$$(S_{V,W}\sigma)(\mathbf{x}) = \int_W G(\mathbf{x}, \mathbf{y})\sigma(\mathbf{y}) ds_{\mathbf{y}} \quad \mathbf{x} \in V \tag{37}$$

$$(D_{V,W}\tau)(\mathbf{x}) = \int_W \frac{\partial G}{\partial n_{\mathbf{y}}}(\mathbf{x}, \mathbf{y})\tau(\mathbf{y}) ds_{\mathbf{y}} \quad \mathbf{x} \in V \tag{38}$$

$$(D^*_{V,W}\sigma)(\mathbf{x}) = \int_W \frac{\partial G}{\partial n_{\mathbf{x}}}(\mathbf{x}, \mathbf{y})\sigma(\mathbf{y}) ds_{\mathbf{y}} \quad \mathbf{x} \in V \tag{39}$$

$$(T_{V,W}\tau)(\mathbf{x}) = \int_W \frac{\partial^2 G}{\partial n_{\mathbf{x}} \partial n_{\mathbf{y}}}(\mathbf{x}, \mathbf{y})\tau(\mathbf{y}) ds_{\mathbf{y}} \quad \mathbf{x} \in V. \tag{40}$$

When $V = W$, these operators are to be understood in the principal value sense. By analogy with (31) and (32), versions of these operators whose kernels include the phased sum over 3×3 images of the source are indicated with a tilde ($\sim$): that is, $\widetilde{S}_{V,W}, \widetilde{D}_{V,W}, \widetilde{D}^*_{V,W}$, and $\widetilde{T}_{V,W}$.

We are now in a position to provide explicit expressions for the operators A, B, C, Q in (34). Comparing (33) to (15), it is clear that the operator A is the same as A_{QP} in (20) but with the replacement of $S^{(\omega)}_{\mathrm{QP}}$, $D^{(\omega)}_{\mathrm{QP}}$ and $T^{(\omega)}_{\mathrm{QP}}$, by $\widetilde{S}_{\partial\Omega,\partial\Omega}$, $\widetilde{D}_{\partial\Omega,\partial\Omega}$ and $\widetilde{T}_{\partial\Omega,\partial\Omega}$, respectively. It is straightforward to verify that A is a compact perturbation of the identity.

The operator C describes the effect of the inclusion densities on the discrepancy d. Its eight sub-blocks are found by inserting (31) and (32) into (33) then evaluating (28), giving

$$C = \begin{bmatrix} \widetilde{D}_{L,\partial\Omega} - \alpha^{-1}\widetilde{D}_{L+\mathbf{e}_1,\partial\Omega} & -\widetilde{S}_{L,\partial\Omega} + \alpha^{-1}\widetilde{S}_{L+\mathbf{e}_1,\partial\Omega} \\ \widetilde{T}_{L,\partial\Omega} - \alpha^{-1}\widetilde{T}_{L+\mathbf{e}_1,\partial\Omega} & -\widetilde{D}^*_{L,\partial\Omega} + \alpha^{-1}\widetilde{D}^*_{L+\mathbf{e}_1,\partial\Omega} \\ \widetilde{D}_{B,\partial\Omega} - \beta^{-1}\widetilde{D}_{B+\mathbf{e}_2,\partial\Omega} & -\widetilde{S}_{B,\partial\Omega} + \beta^{-1}\widetilde{S}_{B+\mathbf{e}_2,\partial\Omega} \\ \widetilde{T}_{B,\partial\Omega} - \beta^{-1}\widetilde{T}_{B+\mathbf{e}_2,\partial\Omega} & -\widetilde{D}^*_{B,\partial\Omega} + \beta^{-1}\widetilde{D}^*_{B+\mathbf{e}_2,\partial\Omega} \end{bmatrix}$$

Consider now the any of the four upper sub-blocks of C. There are nine phased copies of $\partial\Omega$ which contribute to the field on the left (L) and right ($L + \mathbf{e}_1$) wall. From symmetry and translation invariance considerations, however, it is easy to check that the contributions from the six left-most images on L (dotted curves in Fig. 4(a)) are equal to the contributions of the six right-most images on $L + \mathbf{e}_1$ (dotted curves in Fig. 4(b)). In the (1,1) sub-block, for example, we have:

$$\widetilde{D}_{L,\partial\Omega} - \alpha^{-1}\widetilde{D}_{L+\mathbf{e}_1,\partial\Omega} = \sum_{k\in\{-1,0,1\}} \beta^k\left(\alpha D_{L,\partial\Omega+\mathbf{e}_1+k\mathbf{e}_2} - \alpha^{-2} D_{L,\partial\Omega-2\mathbf{e}_1+k\mathbf{e}_2}\right)$$

A rotated version of the analysis applies to the lower four sub-blocks in C. The result is that the entries in C involve only source-target interactions at distances *greater* than the size of the unit cell, ensuring the rapid convergence of a representation in terms of smooth functions.

We next discuss the representation of ξ and $u_{\mathrm{QP}}[\xi]$ in more detail, which will determine the form of blocks Q and B of the full system matrix E.

4.2. Choice of auxiliary densities and their images

The auxiliary field u_{QP} is determined by the choice of layer potentials on the boundary of (and outside of) U. We will use double and single layer densities on both the left (L) and bottom (B) boundaries of U, as well as on the other segments of the "tic-tac-toe" board in Fig. 1(b). More precisely, we define the vector of unknowns ξ by $\xi := [\tau_L;\ -\sigma_L;\ \tau_B;\ -\sigma_B]$, and set

$$u_{\mathrm{QP}} = \sum_{\substack{j\in\{0,1\}\\ k\in\{-1,0,1\}}} \alpha^j\beta^k\left(\mathcal{S}_{L+j\mathbf{e}_1+k\mathbf{e}_2}\sigma_L + \mathcal{D}_{L+j\mathbf{e}_1+k\mathbf{e}_2}\tau_L\right) + \sum_{\substack{j\in\{-1,0,1\}\\ k\in\{0,1\}}} \alpha^j\beta^k\left(\mathcal{S}_{B+j\mathbf{e}_1+k\mathbf{e}_2}\sigma_B + \mathcal{D}_{B+j\mathbf{e}_1+k\mathbf{e}_2}\tau_B\right) \tag{41}$$

The inclusion of the image segments leads to cancellations that are numerically advantageous in the operator Q, just as we found that images helped with the operator C in the preceding section.

We should first clarify the definition (28) of the discrepancy functions: field values should be interpreted as their limiting values on the wall approaching from *inside* U, since it is the field in U that (33) and (41) represent. For example, $f := u^+|_L - \alpha^{-1}u^-|_{L+\mathbf{e}_1}$, where, as before $u^\pm(\mathbf{x}) := \lim_{\varepsilon\to 0^+} u(\mathbf{x} \pm \varepsilon\boldsymbol{n})$, and $\boldsymbol{n}$ is the normal at $\mathbf{x}$.

Recall now that the operator Q expresses the effect of the four densities in ξ on the four discrepancy functions f, f', g, g'. If (41) contained only the terms $j = k = 0$, this would correspond to densities σ_L and τ_L placed on L, and σ_B and τ_B placed on B. While this is mathematically acceptable, it results in various complicated self-interactions and interactions between segments that share a common corner. This would lead to singularities in densities requiring more complicated discretization and quadrature. Although there has been significant progress in this direction (see, for example, [11,25]), in the present context we have the luxury of including the ten additional image segments in (41), which cancel both the self and near-field

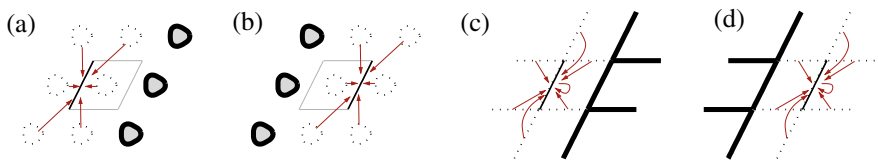


Fig. 4. Discrepancy cancellation due to neighbor image sums. Each arrow represents the influence of a source density on a target segment. (a) For the four upper sub-blocks of C, the six nearest source images (dotted) contribute to the discrepancy on the left wall L. (b) The six nearest source images (dotted) contribute exactly the same field (suitable phased) to the right wall ($L + \mathbf{e}_1$). The net result is that only the distant sources, shown in bold, have a non-zero effect. The same holds for all four upper sub-blocks of C. A rotated version applies to the lower four sub-blocks of C. (c) and (d) Contributions to the sub-blocks Q_{LL} and Q_{LB} of Q. The seven indicated terms (dotted source segments) cancel in the two diagrams, leaving only the contributions from distant wall segments shown in bold. A rotated version applies to the sub-blocks Q_{BL} and Q_{BB}.

corner interactions. As a result, our implementation is simpler and involves fewer degrees of freedom. The cancellation mechanism is shown in Fig. 4. The effect on $u^+|_L$ of the seven segments touching L, for example, cancels the effect on $u^-|_{L+\mathbf{e}_1}$ of the seven segments touching $L+\mathbf{e}_1$, leaving only ten far field contributions.

It is important to note that the local terms due to the jump relations do not cancel: e.g. a density function τ_L placed on L contributes a term $\frac{1}{2}\tau_L$ to $u^+|_L$, while $\alpha\tau_L$ placed on $L+\mathbf{e}_1$ contributes $-\frac{1}{2}\alpha\tau_L$ to $u^-|_{L+\mathbf{e}_1}$. These two terms add to contribute τ_L to f. One may check in this fashion that the jump relations contribute an identity to the diagonal sub-blocks of Q. This yields the crucial result that Q is the identity plus a compact operator, with the compact part generated by interactions at a distance greater than the size of the unit cell. After the above cancellations and simplification, we have,

$$Q = I + \begin{bmatrix} Q_{LL} & Q_{LB} \\ Q_{BL} & Q_{BB} \end{bmatrix}$$

where

$$Q_{LL} = \begin{bmatrix} \sum_{j\in\{-1,1\},k\in\{-1,0,1\}} j\alpha^j\beta^k D_{L,L+j\mathbf{e}_1+k\mathbf{e}_2} & -\sum_{j\in\{-1,1\},k\in\{-1,0,1\}} j\alpha^j\beta^k S_{L,L+j\mathbf{e}_1+k\mathbf{e}_2} \\ \sum_{j\in\{-1,1\},k\in\{-1,0,1\}} j\alpha^j\beta^k T_{L,L+j\mathbf{e}_1+k\mathbf{e}_2} & -\sum_{j\in\{-1,1\},k\in\{-1,0,1\}} j\alpha^j\beta^k D^*_{L,L+j\mathbf{e}_1+k\mathbf{e}_2} \end{bmatrix}$$

$$Q_{LB} = \begin{bmatrix} \sum_{k\in\{0,1\}} \beta^k\left(\alpha D_{L,B+\mathbf{e}_1+k\mathbf{e}_2} - \alpha^{-2} D_{L,B-2\mathbf{e}_1+k\mathbf{e}_2}\right) & \sum_{k\in\{0,1\}} \beta^k\left(-\alpha S_{L,B+\mathbf{e}_1+k\mathbf{e}_2} + \alpha^{-2} S_{L,B-2\mathbf{e}_1+k\mathbf{e}_2}\right) \\ \sum_{k\in\{0,1\}} \beta^k\left(\alpha T_{L,B+\mathbf{e}_1+k\mathbf{e}_2} - \alpha^{-2} T_{L,B-2\mathbf{e}_1+k\mathbf{e}_2}\right) & \sum_{k\in\{0,1\}} \beta^k\left(-\alpha D^*_{L,B+\mathbf{e}_1+k\mathbf{e}_2} + \alpha^{-2} D^*_{L,B-2\mathbf{e}_1+k\mathbf{e}_2}\right) \end{bmatrix}$$

$$Q_{BL} = \begin{bmatrix} \sum_{j\in\{0,1\}} \alpha^j\left(\beta D_{B,L+j\mathbf{e}_1+\mathbf{e}_2} - \beta^{-2} D_{B,L+j\mathbf{e}_1-2\mathbf{e}_2}\right) & \sum_{j\in\{0,1\}} \alpha^j\left(-\beta S_{B,L+j\mathbf{e}_1+\mathbf{e}_2} + \beta^{-2} S_{B,L+j\mathbf{e}_1-2\mathbf{e}_2}\right) \\ \sum_{j\in\{0,1\}} \alpha^j\left(\beta T_{B,L+j\mathbf{e}_1+\mathbf{e}_2} - \beta^{-2} T_{B,L+j\mathbf{e}_1-2\mathbf{e}_2}\right) & \sum_{j\in\{0,1\}} \alpha^j\left(-\beta D^*_{B,L+j\mathbf{e}_1+\mathbf{e}_2} + \beta^{-2} D^*_{B,L+j\mathbf{e}_1-2\mathbf{e}_2}\right) \end{bmatrix}$$

$$Q_{BB} = \begin{bmatrix} \sum_{j\in\{-1,0,1\},k\in\{-1,1\}} k\alpha^j\beta^k D_{B,B+j\mathbf{e}_1+k\mathbf{e}_2} & -\sum_{j\in\{-1,0,1\},k\in\{-1,1\}} k\alpha^j\beta^k S_{B,B+j\mathbf{e}_1+k\mathbf{e}_2} \\ \sum_{j\in\{-1,0,1\},k\in\{-1,1\}} k\alpha^j\beta^k T_{B,B+j\mathbf{e}_1+k\mathbf{e}_2} & -\sum_{j\in\{-1,0,1\},k\in\{-1,1\}} k\alpha^j\beta^k D^*_{B,B+j\mathbf{e}_1+k\mathbf{e}_2} \end{bmatrix}$$

Finally, we discuss the B operator from (34), which describes the effect of the auxiliary densities ξ on the mismatch. As with A, since the mismatch involves values on only a single curve $\partial\Omega$, there is no opportunity for cancellation. Inserting (41) into (21) we get

$$B = \sum_{j\in\{0,1\},k\in\{-1,0,1\}} \alpha^j\beta^k \begin{bmatrix} D_{\partial\Omega,L+j\mathbf{e}_1+k\mathbf{e}_2} & -S_{\partial\Omega,L+j\mathbf{e}_1+k\mathbf{e}_2} & 0 & 0 \\ T_{\partial\Omega,L+j\mathbf{e}_1+k\mathbf{e}_2} & -D^*_{\partial\Omega,L+j\mathbf{e}_1+k\mathbf{e}_2} & 0 & 0 \end{bmatrix} + \sum_{j\in\{-1,0,1\},k\in\{0,1\}} \alpha^j\beta^k \begin{bmatrix} 0 & 0 & D_{\partial\Omega,B+j\mathbf{e}_1+k\mathbf{e}_2} & -S_{\partial\Omega,B+j\mathbf{e}_1+k\mathbf{e}_2} \\ 0 & 0 & T_{\partial\Omega,B+j\mathbf{e}_1+k\mathbf{e}_2} & -D^*_{\partial\Omega,B+j\mathbf{e}_1+k\mathbf{e}_2} \end{bmatrix} \tag{42}$$

Summarizing the above, E is a compact perturbation of the identity. Its blocks C and Q involve interaction distances greater than the unit cell size. Its block A involves distances controlled by the shape of the inclusion and its nearest approach to its neighboring images. Its block B involves distances determined by the nearest approach of $\partial\Omega$ to ∂U.

4.3. Numerical implementation and discretization of B

We discretize the four blocks of the integral operator E in (34) to give the matrix $\widetilde{E} \in \mathbb{C}^{(2N+4M)\times(2N+4M)}$ as follows. We sample the densities on $\partial\Omega$ at equispaced points with respect to the given definition of the curve, as in Section 3.1. We sample the densities on the walls L and B at M standard Gaussian nodes, as in Section 3.2. A is then discretized in the same way as A_{QP} in Section 3.1 with a mix of the periodic trapezoidal rule and Kress' singular quadratures for the self-interaction of $\partial\Omega$. The (Nyström) method (22) may be used for the off-diagonal block C, and also for the wall's self-interaction Q. No special singular quadratures are needed in Q, due to the cancellations discussed above.

The B operator (42) involves computing the field due to source densities on walls L and B (and their images shown in Fig. 1(b)) at targets on $\partial\Omega$. When the distance from the inclusion to boundary dist $(\partial\Omega, \partial U)$ is large, the plain Nyström method may be used to construct the discretized matrix $\widehat{B}$. We will refer to this as discretization method B1. With nodes $\mathbf{y}_m$ and weights w_m on wall L, and nodes $\mathbf{x}_j$ on $\partial\Omega$, for example, the term $S_{\partial\Omega,L}$ in the (1,2)-block of (42) becomes the matrix $\widehat{S} \in \mathbb{C}^{N\times M}$ with elements $\widehat{S}_{jm} = \frac{i}{4}H_0^{(1)}(\omega|\mathbf{x}_j - \mathbf{y}_m|)w_m$.

When dist $(\partial\Omega, \partial U)$ becomes small, of course, the convergence rate of method B1 will become unacceptably poor. However, by construction, for a Bloch eigenfunction the field (41) generated by the wall densities in ξ has no singularities in the 3×3 neighboring block of unit cells. Hence these densities remain smooth, poor convergence being merely due to inaccurate evaluation of their field close to the walls. This leaves room for a large number of options:

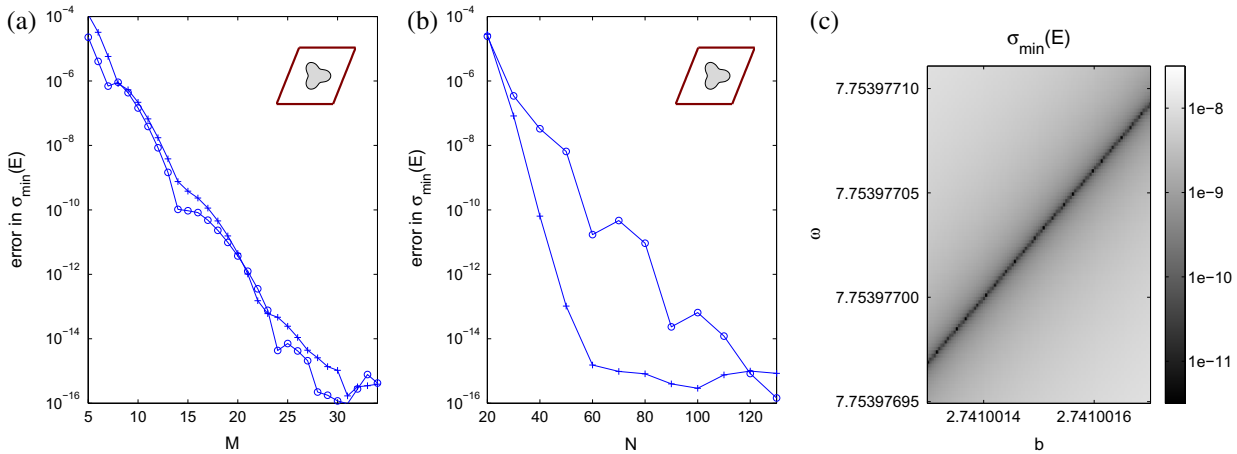


Fig. 5. Convergence of new periodizing scheme using auxilliary densities (described in Section 4), using method B1, for the same geometry and parameters as in Fig. 2(a). The meaning of the two curves is also the same as in the earlier figure. (a) Absolute error in $\sigma_{\min}(\widetilde{E})$ vs. M the number of nodes on each unit cell wall, for fixed $N = 70$ nodes on $\partial\Omega$. (b) Same as (a) except convergence vs. N, for fixed $M = 30$. (c) Same as Fig. 3(b) but using the new scheme: note the absence of pollution by the empty band structure.

(B2) For the rows of $\widehat{B}$ corresponding to target points on $\partial\Omega$ that are distance d_0 or closer to ∂U, use adaptive Gauss–Kronrod quadrature[2] with integrand given by the product of the kernel function and the Lagrange polynomial interpolant [32, Sec. 8.1] for the density at the M quadrature points. d_0 is some $O(1)$ constant. For the other rows, use method B1.

(B3) Project onto an order-L cylindrical J-expansion at the origin. This is done by computing a representation (27) for each of the point monopole or dipole sources in the quadrature approximation to the source densities on the walls, and then evaluating this at the target quadrature points on $\partial\Omega$ to fill the elements of $\widehat{B}$. The example term discussed for B1 gives $\widehat{S} = RP$, where the "source-to-local" matrix $P \in \mathbb{C}^{(2L+1)\times M}$ has elements

$$P_{lm} = \frac{i}{4} H_l^{(1)}(\omega|\mathbf{y}_m|) e^{-il\theta_m} w_m$$

and converts single layer density values to J-expansion coefficients. This follows from Graf's addition formula [1, Eq. 9.1.79]. The expansion matrix $R \in \mathbb{C}^{N\times(2L+1)}$ has elements $R_{jl} = J_l(\omega|\mathbf{x}_j|)e^{il\phi_j}$. In the above θ_m, ϕ_j are polar angles of points $\mathbf{y}_m$, $\mathbf{x}_j$, respectively. Similar formulae apply for double layers and evaluation of derivatives. To reduce dynamic range (hence roundoff error) we in fact scale the J-expansion by the factors ρ_l of Section 3.2 (this does not change the mathematical definition of $\widehat{S}$.)

(B4) Use a more sophisticated quadrature approach, such as those of [5,26,37].

Methods B2–B4 evaluate u_{QP} due to a spectral interpolant of the discretized wall densities, with an accuracy that persists up to the boundary of U. Note that this does not increase the number M of degrees of freedom associated with each such density. Since the underlying density is smooth (in fact analytic), the convergence rate is high and we are able to keep M very modest.

We have implemented methods B1, B2 and B3. We use the quadrature weights to scale the matrix $\widehat{E}$ to give $\widetilde{E}$ in an analogous fashion to (26), so that singular values of $\widetilde{E}$ approximate those of E.

Finally, there are many possible ways to locate parameter values (ω, a, b) where $\widetilde{E}$ is singular. In this paper, we will simply plot its smallest singular value $\sigma_{\min}(\widetilde{E})$ vs. the Bloch parameters, as in Section 3.

5. Results of proposed scheme

We first test the convergence of the new scheme for the same small inclusion used in Section 3, with the simplest discretization method for B, namely B1. As before, we test two Bloch parameter b values, one which is far from an eigenvalue, and one of which is guaranteed to be an eigenvalue according to Theorem 4. Fixing $N = 70$, which was found in Section 3.1 to be fully converged when at an eigenvalue, we first vary M, the number of nodes per unit cell wall. Fig. 5(a) shows the convergence of the minimum singular value of the discretized matrix $\widetilde{E}$ to its converged value (when far from an eigenvalue), or to zero (when at an eigenvalue). The convergence is spectral, and in both cases full machine accuracy is reached at $M = 30$. (For $N > 70$ the results are unchanged.) Thus for a matrix of order $2N + 4M = 260$, we are able to locate the desired band structure with relative error around 10^{-15} in the Bloch parameters (a, b). Filling such a matrix takes around 0.45 s and computing the complex SVD around 0.15 s.[3] Furthermore, by storing coefficient matrices in the expansion $\widetilde{E} = \sum_{-1\leqslant j,k\leqslant 2} \alpha^j \beta^k \widetilde{E}^{(j,k)}$ at fixed ω, we can fill $\widetilde{E}$ for new a, b values in 0.05 s.

[2] This was implemented with MATLAB's `quadgk`, which uses a pair of 15th and 7th order formulae, with relative tolerance set to 10^{-12}.

[3] All timings are reported for a laptop running MATLAB 2008a with a 2 GHz Intel Core Duo CPU.

Fig. 5(b) shows that, with M in the new quasi-periodizing scheme sufficient to yield machine precision, the error convergence rate with respect to N is the same as that of the old scheme. Fig. 5(c) demonstrates the robustness of the scheme, by plotting the smallest singular value over the same region of parameter space as Fig. 3(b). Notice that the location of the desired band structure (black line) is unchanged, but that the divergent behavior near the empty resonant band structure has entirely vanished.

5.1. Inclusions approaching and intersecting the unit cell wall

Given a crystal of inclusions, it may be impossible to choose a parallelogram unit cell U whose boundary does not come close to or even intersect $\partial\Omega$. Although this is not an issue for the scheme of Section 3, for the new scheme which relies on ∂U it is a potential problem.

We first show that, as expected, with method B1 the error performance deteriorates exponentially as $\partial\Omega$ approaches ∂U. In Fig. 6(a) we plot the minimum singular value at a Bloch eigenvalue, as a function of distance d that the inclusion has been translated in the x direction (translation does not affect the Bloch eigenvalue.) Numerical parameters N and M are held fixed. The logarithm of the error grows roughly linearly with d and reaches $O(1)$ for dist $(\partial\Omega, \partial U) = 0$, indicated by the dotted vertical line at around $d = 0.23$. Method B2, also shown in Fig. 6(a), uses adaptive quadrature for accurate evaluation of u_{QP} in all of U. For very small d, the inclusion is still centrally located (far from the wall) and B2 is identical to B1, with an error of 10^{-15}. The error is around 10^{-12} as one approaches the wall (more or less independent of d), limited by the accuracy of `quadgk`. This proves that the deterioration seen with B1 is associated with the B operator block, and can be remedied merely by careful discretization of B without increasing the matrix size. We did not bother continuing the computation with B1 or B2 after the inclusion crosses the wall; here they fail because (41), as constructed, represents u_{QP} only inside U (jump relations cause the values outside U to be different). We note that to use B1 or B2 correctly, one would have to wrap the boundary points outside U back into the cell, evaluate at the wrapped point, and correct for phase. Method B2 is not very useful in practice since the call to a black box adaptive quadrature routine causes the matrix fill time to increase to 55 s.

Finally we use method B3 with $L = 16$, $M = 30$ and with $L = 22$, $M = 40$. In the first case, errors grow slowly to around 10^{-12} as dist $(\partial\Omega, \partial U)$ reaches zero, and then continue to grow slowly to a plateau at around 10^{-9}, even though most of $\partial\Omega$ now falls *outside* of the unit cell. The cost of B3 is not much more than B1, taking 0.7 s to fill $\tilde{E}$. Note that the J-expansion used to represent u_{QP} has effectively carried out analytic continuation beyond U. This is stable because our image structure has pushed the singularities out beyond the nearest image cells. It is perhaps worth observing that some care must be taken in setting L. With $M = 30$, increasing L above 16 would worsen errors (not shown). The reason is that the coefficients $|l| > 16$ involve more oscillatory integrands which are not resolved by $M = 30$ points. Increasing M to 40 permits increased precision with $L = 22$, as seen in Fig. 6(a).

There is another potential pitfall with method B3 as implemented; if both L and d get larger, there may arise singular values of $\tilde{E}$ which become exponentially small, associated with highly oscillatory non-physical densities on the farthest part of $\partial\Omega$. For illustration, with $L = 16$ and $d = 0.6$, the second smallest singular value is 10^{-4}; with $L = 22$ the second smallest singular value shrinks to 10^{-6}. (When $d = 0$, the second smallest singular value is 10^{-1}.) This is troublesome for eigenvalue search methods that track $\sigma_{\min}(\tilde{E})$ vs. Bloch parameters, since the desired minima will be obscured by these spurious small

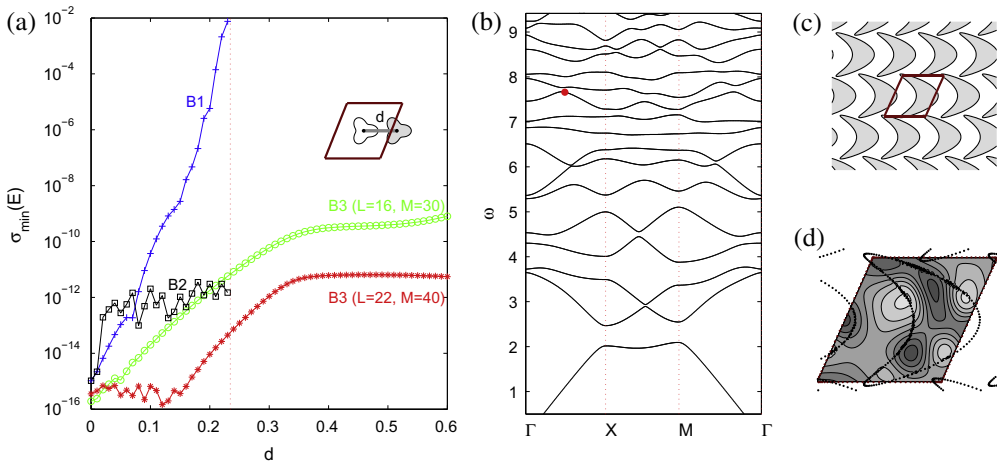


Fig. 6. (a) Dependence of $\sigma_{\min}(\tilde{E})$ on x-translation distance d of $\partial\Omega$ relative to the system of Fig. 2(a), for fixed $N = 70$, and $M = 30$. The vertical line shows where $\partial\Omega$ starts to touch ∂U. B is discretized as follows: method B1 (+ symbols), method B2 with $d_0 = 0.2$ (□ symbols), method B3 with $L = 16$ (∘ symbols), method B3 with $L = 22$ and $M = 40$ (* symbols). Inset shows unit cell and inclusion at $d = 0.6$. (b) Band structure for crescent-shaped photonic crystal shown in (c), index $n = 2$, shape $(0.265\cos 2\pi t + 0.318\cos 4\pi t, 0.53\sin 2\pi t)$, $0 \leqslant t < 1$, unit cell $\mathbf{e}_1 = (1,0)$, $\mathbf{e}_2 = (0.45,1)$. A tour $\Gamma X M \Gamma$ of the Brillouin zone is shown, where Γ is $(a,b) = (0,0)$, X is $(\pi,0)$, and M is (π,π). (d) Contours of the Bloch mode $\mathrm{Re}[u]$ with parameters shown by the dot on the band diagram.

singular values everywhere except in a small neighborhood of the desired band structure. We will discuss search methods less sensitive to this problem in a future paper. For now the lesson is that, when parts of an inclusion extend far beyond U, there is a price to pay for making use of analytic continuation.

5.2. Application to band structure

We compute the band structure of a more difficult crystal in Fig. 6(b). Ω is far from circular, hence simple multipole methods [23] would not be accurate. The closest approach to its neighbors is only 0.06, so that $N = 150$ points are needed in discretizing the inclusion boundary. Note that any parallelogram unit cell must intersect $\partial\Omega$, so the method of [49] cannot be used without modification. We use method B3 with $M = 35$ and $L = 18$. As illustrated before in Fig. 2(b), the minimum values of $\sigma_{\min}(\widetilde{E})$ on the band structure indicate the size of the errors in the Bloch parameters found. By this measure, sampling 100 random points on the first 15 bands, we find a median error of 3×10^{-10} and a maximum 1.6×10^{-9}. 1.7 s were required to fill the matrix $\widetilde{E}$ of order 440 once for a given ω, a, b (and 0.13 s for subsequent values of a, b). The SVD required 0.7 s for a matrix of this size. We located the band structure using 8000 such evaluations and a specialized search algorithm, which we will describe in a forthcoming paper. The search algorithm is also accelerated by computing the determinant of $\widetilde{E}$ rather than the SVD, at a cost of 0.1 s for each matrix. The total CPU time required was 35 min.

Fig. 6(d) shows a single Bloch mode on the 11th band for this crescent-shaped crystal. This took 16 s to evaluate on a 100×100 grid over U using (33), and the J-expansion for (41) (with no fast-multipole acceleration).

6. Conclusions

We have presented two algorithms for locating the band structure of a two-dimensional photonic crystal, in the z-invariant Maxwell setting. The first (Section 3) uses the quasi-periodic Green's function. Theorem 4 guarantees the success of this method (no spurious or missed modes) as long as the band structure for the empty unit cell is avoided, where we have shown that the method fails. The second method (Section 4) introduces a small number of additional degrees of freedom on the walls to represent the periodizing part of the field: numerical evidence suggests that it is immune to breakdown for any Bloch parameters (Conjecture 7). The two schemes are connected by the following observation.

Remark 8. Computing the Schur complement formula for the operator system (34) recovers the quasi-periodic Green's function approach described by (20). In particular,

$$A_{\mathrm{QP}} = A - BQ^{-1}C.$$

The quasi-periodic Green's function approach fails when Q becomes singular and A_{QP} blows up. The full system (34), on the other hand, remains well-behaved.

We have shown spectral convergence for both schemes, achieving close to machine precision accuracy on simple crystals using only a few hundred degrees of freedom, hence CPU times of less than 1 s for testing at a single parameter set (ω, a, b). In the new scheme we have shown (method B3) how to handle the passage of the inclusion through the unit cell boundary, without much sacrifice in accuracy, without much extra numerical effort, and with no bookkeeping needed to determine which points of $\partial\Omega$ lie in U. The latter is convenient for larger-scale or three-dimensional (3D) computations if existing scattering codes are to be used to fill the A operator block. Other ways to handle this intersection problem exist, such as a variant of B2 which wraps points on $\partial\Omega$ back into U, with which we have preliminary success.

We have not discussed the methods we use for the nonlinear eigenvalue problem, due to space constraints. The scheme of Yuan et al. [49] uses a quadratic eigenvalue problem, and factorizes the scattering matrix of the inclusion at each ω, hence may be faster than our scheme for small systems. However, moving to large-scale systems with more than 10^4 degrees of freedom, such a factorization would be impractical compared to an iterative version of our scheme.

Some generalizations of what we present are straightforward, such as multiple inclusions per unit cell, non-simply connected inclusions, or inclusions with corners (using quadrature rules such as [11,25]). There exist regimes, however, that would require some modification. These include two phase dielectrics one or more of which are connected through the bulk (sometimes called bicontinuous), and unit cells which are highly skew or have large aspect ratios.

Our new representation for quasi-periodic fields can also be used for scattering calculations from periodic one-dimensional arrays of inclusions in 2D and one or two-dimensional arrays in 3D. Because we rely entirely on the free-space Green's function, it should be straightforward to create quasi-periodic solvers from existing scattering codes. We will describe such solvers at a later date.

Acknowledgements

We thank Greg Beylkin, Zydrunas Gimbutas and Ivan Graham for insightful discussions. The work of AHB was supported by NSF grant DMS-0811005, and by the Class of 1962 Fellowship at Dartmouth College. The work of LG was supported by the Department of Energy under contract DEFG0288ER25053 and by AFOSR under MURI grant FA550-06-1-0337.

Appendix A. Proof of Theorem 4

Recall the Green's representation formulae [16, Sec. 3.2]. If u satisfies $(\Delta + \omega^2)u = 0$ in Ω, recalling that u^- and u_n^- signify limits on $\partial\Omega$ approaching from the inside, and the normal always points outwards from Ω, then

$$-\mathcal{S}^{(\omega)}u_n^- + \mathcal{D}^{(\omega)}u^- = \begin{cases} -u & \text{in } \Omega \\ 0 & \text{in } \mathbb{R}^2 \setminus \overline{\Omega} \end{cases} \tag{A.1}$$

The exterior representation has the opposite sign: let u satisfy $(\Delta + \omega^2)u = 0$ in $\mathbb{R}^2 \setminus \overline{\Omega}$ and the Sommerfeld radiation condition, that is,

$$\frac{\partial u}{\partial r} - i\omega u = o(r^{-1/2}), \quad r := |\mathbf{x}| \to \infty \tag{A.2}$$

holds uniformly with respect to direction $\mathbf{x}/r$. Then,

$$-\mathcal{S}^{(\omega)}u_n^+ + \mathcal{D}^{(\omega)}u^+ = \begin{cases} 0 & \text{in } \Omega \\ u & \text{in } \mathbb{R}^2 \setminus \overline{\Omega} \end{cases} \tag{A.3}$$

We will need the following quasi-periodic analogues.

Lemma 9. *Let u satisfy $(\Delta + \omega^2)u = 0$ in Ω, and $\overline{\Omega} \subset U$, Then for each Bloch phase (α,β),*

$$-\mathcal{S}_{\mathrm{QP}}^{(\omega)}u_n^- + \mathcal{D}_{\mathrm{QP}}^{(\omega)}u^- = \begin{cases} -u & \text{in } \Omega \\ 0 & \text{in } U \setminus \overline{\Omega} \end{cases} \tag{A.4}$$

Proof. Write G_{QP} using (11) and notice that each term other than $(m, n) = (0,0)$ contributes zero. This is because all points in U lie outside each closed curve $\partial\Omega - m\mathbf{e}_1 - n\mathbf{e}_2$, and we may apply the second (extinction) case of (A.1) to show that they have no effect in U. □

Lemma 10. *Let u satisfy $(\Delta + \omega^2)u = 0$ in $U \setminus \overline{\Omega}$ and quasi-periodicity (6)–(9), and $\overline{\Omega} \subset U$. Then*

$$-\mathcal{S}_{\mathrm{QP}}^{(\omega)}u_n^+ + \mathcal{D}_{\mathrm{QP}}^{(\omega)}u^+ = \begin{cases} 0 & \text{in } \Omega \\ u & \text{in } U \setminus \overline{\Omega} \end{cases} \tag{A.5}$$

Proof. We follow the usual method of proof [16, Thm. 3.3] but with the quasi-periodicity condition playing the role of the radiation condition. Apply Green's 2nd identity to the functions u and $G_{\mathrm{QP}}(\mathbf{x},\cdot)$ in the domain $U \setminus \overline{\Omega}$ if $\mathbf{x} \in \Omega$, or the domain $\{\mathbf{y} \in U \setminus \overline{\Omega} : |\mathbf{x} - \mathbf{y}| > \varepsilon\}$ if $\mathbf{x} \in U \setminus \overline{\Omega}$. In the latter case the limit $\varepsilon \to 0$ is taken, and (11) shows that only the $(m,n) = (0,0)$ term contributes to the limit of the integral over the sphere of radius ε. In both cases the boundary integrals contain the term

$$\int_{\partial U} \frac{\partial G_{\mathrm{QP}}}{\partial n_{\mathbf{y}}}(\mathbf{x},\mathbf{y})u(\mathbf{y}) - G_{\mathrm{QP}}(\mathbf{x},\mathbf{y})u_n(\mathbf{y})\, ds_{\mathbf{y}}, \tag{A.6}$$

which vanishes by cancellation on opposing walls, since u is quasi-periodic with phases (α, β), but $G_{\mathrm{QP}}(\mathbf{x},\cdot)$ is anti-quasiperiodic, i.e. quasi-periodic with phases $(\alpha^{-1}, \beta^{-1})$. □

Turning now to Theorem 4, to prove the *if* part, we show that whenever the operator has a nontrivial nullspace, a Bloch eigenfunction u may be constructed, i.e. a solution to (2)–(9), that we must take care to show is nontrivial. Let $\eta = [\tau; -\sigma] \neq 0$ be a nontrivial density such that $A_{\mathrm{QP}}\eta = 0$. Immediately we have that the resulting field u given by (15) satisfies (2)–(9). We now define a complementary field over the whole plane minus $\partial\Omega$,

$$v = \begin{cases} \mathcal{S}_{\mathrm{QP}}^{(\omega)}\sigma + \mathcal{D}_{\mathrm{QP}}^{(\omega)}\tau & \text{in } \Omega \\ -\mathcal{S}^{(n\omega)}\sigma - \mathcal{D}^{(n\omega)}\tau & \text{in } \mathbb{R}^2 \setminus \overline{\Omega} \end{cases} \tag{A.7}$$

Suppose $u \equiv 0$. Then $u^- = u_n^- = 0$ and by the jump relations for $\mathcal{S}^{(n\omega)}\sigma + \mathcal{D}^{(n\omega)}\tau$ we get $v^+ = -\tau$ and $v_n^+ = \sigma$. Similarly, since $u^+ = u_n^+ = 0$ by the jump relations for $\mathcal{S}_{\mathrm{QP}}^{(\omega)}\sigma + \mathcal{D}_{\mathrm{QP}}^{(\omega)}\tau$ we get $v^- = -\tau$ and $v_n^- = \sigma$. It is easy to check that v solves the (swapped-wavenumber) transmission problem,

$$(\Delta + \omega^2)v = 0 \quad \text{in } \Omega \tag{A.8}$$

$$(\Delta + n^2\omega^2)v = 0 \quad \text{in } \mathbb{R}^2 \setminus \overline{\Omega} \tag{A.9}$$

$$\frac{\partial v}{\partial r} - in\omega v = o(r^{-1/2}), \quad r \to \infty, \text{ uniformly in direction} \tag{A.10}$$

$$v^+ - v^- = h \tag{A.11}$$

$$v_n^+ - v_n^- = h' \tag{A.12}$$

with homogeneous boundary discontinuity data $h = h' = 0$. By uniqueness for this problem [16, Thm. 3.40] we get that $v \equiv 0$ in $\mathbb{R}^2$, from which the jump relations back to u imply $\sigma = \tau = 0$, which contradicts our assumption of nontrivial density. Thus u is a Bloch eigenfunction.

To prove the *only if* part we show that, given the existence of a Bloch eigenfunction, we may exhibit a (nontrivial) density η such that $A_{\mathrm{QP}}\eta = 0$. Let w be a Bloch eigenfunction with eigenvalue (ω, a, b). Then let v solve (A.8)–(A.12) with the inhomogeneous data $h = -2w|_{\partial\Omega}$ and $h' = -2w_n|_{\partial\Omega}$. (Note that w obeys continuity (4), hence $w|_{\partial\Omega} = w^+ = w^-$ and $w_n|_{\partial\Omega} = w_n^+ = w_n^-$). By [16, Thm. 3.41] we know that a unique solution exists. We now claim that the densities

$$\sigma = w_n|_{\partial\Omega} + v_n^+ \tag{A.13}$$

$$\tau = -w|_{\partial\Omega} - v^+ \tag{A.14}$$

generate precisely the eigenfunction w, i.e. the representation u of (15) obeys $u \equiv w$ in U. We show this by substituting the densities into (15), then applying (A.1) and (A.3) in Ω, and Lemma 10 in $U \setminus \overline{\Omega}$:

$$u = \begin{cases} \mathcal{S}^{(n\omega)} w_n|_{\partial\Omega} - \mathcal{D}^{(n\omega)} w|_{\partial\Omega} + \mathcal{S}^{(n\omega)} v_n^+ - \mathcal{D}^{(n\omega)} v^+ & \text{in } \Omega \\ \mathcal{S}_{\mathrm{QP}}^{(\omega)} w_n|_{\partial\Omega} - \mathcal{D}_{\mathrm{QP}}^{(\omega)} w|_{\partial\Omega} + \mathcal{S}_{\mathrm{QP}}^{(\omega)} v_n^+ - \mathcal{D}_{\mathrm{QP}}^{(\omega)} v^+ & \text{in } U \setminus \overline{\Omega} \end{cases}$$
$$= \begin{cases} w & \text{in } \Omega \\ -w + \mathcal{S}_{\mathrm{QP}}^{(\omega)} v_n^+ - \mathcal{D}_{\mathrm{QP}}^{(\omega)} v^+ & \text{in } U \setminus \overline{\Omega} \end{cases}$$

On the remaining term, we use v's known jumps h and h' to get

$$\mathcal{S}_{\mathrm{QP}}^{(\omega)} v_n^+ - \mathcal{D}_{\mathrm{QP}}^{(\omega)} v^+ = \mathcal{S}_{\mathrm{QP}}^{(\omega)} v_n^- - \mathcal{D}_{\mathrm{QP}}^{(\omega)} v^- - 2\mathcal{S}_{\mathrm{QP}}^{(\omega)} w_n|_{\partial\Omega} + 2\mathcal{D}_{\mathrm{QP}}^{(\omega)} w|_{\partial\Omega} = -2w$$

where we applied Lemma 9 to the first pair, and Lemma 10 to the second as before. Substituting this above shows that $u \equiv w$ in U. Since w has zero mismatch, the density vector $\eta := [\tau; -\sigma]$ satisfies $A_{\mathrm{QP}}\eta = 0$. Finally, η must be nontrivial since $\eta = 0$ would imply $u \equiv 0$ by (15) which contradicts it being equal to the eigenfunction w.

We close with a couple of remarks about the proof. Barring a sign, v in (A.7) is the extension of u's representation (15) into its non-physical regions, a trick originating, in the homogeneous context, with the proof in [16, Thm. 3.41]. Because (15) uses G_{QP} outside, but G inside, the complementary problem is a *nonperiodic* transmission problem, which has known existence and uniqueness. The related analysis of [45] uses G_{QP} both inside and outside. This results in a periodic problem as the complementary problem, and it is not so clear that one can eliminate the possibility of spurious modes.

References

[1] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 10th ed., Dover, New York, 1964.
[2] B.K. Alpert, Hybrid Gauss-trapezoidal quadrature rules, SIAM J. Sci. Comput. 20 (1999) 1551–1584.
[3] W. Axmann, P. Kuchment, An efficient finite element method for computing spectra of photonic and acoustic band-gap materials, J. Comput. Phys. 150 (1999) 468–481.
[4] A.H. Barnett, T. Betcke, Stability and convergence of the method of fundamental solutions for Helmholtz problems on analytic domains, J. Comput. Phys. 227 (2008) 7003–7026.
[5] J. Beale, M.-C. Lai, A method for computing nearly singular integrals, SIAM J. Numer. Anal. 38 (2001) 1902–1925.
[6] C.L. Berman, L. Greengard, A renormalization method for the evaluation of lattice sums, J. Math. Phys. 35 (1994) 6036–6048.
[7] P. Bermel, C. Luo, L. Zeng, L.C. Kimerling, J.D. Joannopoulos, Improving thin-film crystalline silicon solar cell efficiencies with photonic crystals, Opt. Express 15 (25) (2007) 16986–17000.
[8] T. Betcke, Computations of eigenfunctions of planar regions, Ph.D. Thesis, Oxford University, UK, 2005.
[9] G. Beylkin, C. Kurcz, L. Monzón, Fast algorithms for Helmholtz Green's functions, Proc. R. Soc. A 464 (2008) 3301–3326.
[10] L.C. Botten, R.C. McPhedran, N.A. Nicorovici, A.A. Asatryan, C.M. de Sterke, P.A. Robinson, K. Busch, G.H. Smith, T.N. Langtry, Rayleigh multipole methods for photonic crystal calculations, Prog. Electromag. Res. 41 (2003) 21–60.
[11] J. Bremer, V. Rokhlin, I. Sammis, Universal quadratures for boundary integral equations on two-dimensional domains with corners, Yale University Department of Computer Science Technical Report 1420.
[12] H. Cheng, W.Y. Crutchfield, Z. Gimbutas, G.L.F. Ethridge, J. Huang, V. Rokhlin, N. Yarvin, J. Zhao, A wideband fast multipole method for the Helmholtz equation in three dimensions, J. Comput. Phys. 216 (2006) 300–325.
[13] H. Cheng, W.Y. Crutchfield, Z. Gimbutas, G.L.J. Huang, V. Rokhlin, N. Yarvin, J. Zhao, Remarks on the implementation of the wideband FMM for the Helmholtz equation in two dimensions, Contemp. Math. Am. Math. Soc. Providence, RI, 408 (2006) 99–110.
[14] W.C. Chew, J.M. Jin, E. Michielssen, J. Song, Fast and Efficient Algorithms in Computational Electromagnetics, Artech House, Boston, MA, 2001.
[15] S.K. Chin, N.A. Nicorovici, R.C. McPhedran, Green's function and lattice sums for electromagnetic scattering by a square array of cylinders, Phys. Rev. E 49 (5) (1994) 4590–4602.
[16] D. Colton, R. Kress, Integral Equation Methods in Scattering Theory, Wiley, 1983.
[17] W. Crutchfield, H. Cheng, L. Greengard, Sensitivity analysis of photonic crystal fiber, Opt. Express 12 (2004) 4220–4226.
[18] A. Dienstfrey, F. Hang, J. Huang, Lattice sums and the two-dimensional, periodic Green's function for the Helmholtz equation, Proc. R. Soc. Lond. A 457 (2001) 67–85.
[19] D.C. Dobson, An efficient method for band structure calculations in 2d photonic crystals, J. Comput. Phys. 149 (1999) 363–376.
[20] K. Dossou, M. Byrne, L.C. Botten, Finite element computation of grating scattering matrices and application to photonic crystal band calculations, J. Comput. Phys. 219 (2006) 120–143.
[21] S. Fan, P.R. Villeneuve, J.D. Joannopoulos, H.A. Haus, Channel drop filters in a photonic crystal, Opt. Express 3 (1998) 4–11.
[22] R.B. Guenther, J.W. Lee, Partial differential equations of mathematical physics and integral equations, Prentice-Hall, Englewood Cliffs, NJ, 1988.
[23] C. Hafner, The Generalized Multipole Technique for Computational Electromagnetics, Artech House Books, Boston, 1990.
[24] J. Helsing, Integral equation methods for elliptic problems with boundary conditions of mixed type, J. Comput. Phys. 228 (2009) 8892–8907.
[25] J. Helsing, R. Ojala, Corner singularities for elliptic problems: integral equations, graded meshes, quadrature, and compressed inverse preconditioning, J. Comput. Phys. 227 (2008) 8820–8840.

[26] J. Helsing, R. Ojala, On the evaluation of layer potentials close to their sources, J. Comput. Phys. 227 (2008) 2899–2921.
[27] J.D. Jackson, Classical Electrodynamics, third ed., Wiley, 1998.
[28] J.D. Joannopoulos, S.G. Johnson, R.D. Meade, J.N. Winn, Photonic Crystals: Molding the Flow of Light, second ed., Princeton Univ. Press, Princeton, NJ, 2008.
[29] S.G. Johnson, J.D. Joannopoulos, Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis, Opt. Express 8 (3) (2001) 173–190.
[30] S. Kapur, V. Rokhlin, High-order corrected trapezoidal quadrature rules for singular functions, SIAM J. Numer. Anal. 34 (1997) 1331–1356.
[31] R. Kress, Boundary integral equations in time-harmonic acoustic scattering, Math. Comput. Modell. 15 (1991) 229–243.
[32] R. Kress, Numerical analysis, Graduate Texts in Mathematics, vol. 181, Springer-Verlag, 1998.
[33] R. Kress, Linear Integral Equations, Applied Mathematical Sciences, second ed., vol. 82, Springer, 1999.
[34] K.M. Leung, Y. Qiu, Multiple-scattering calculation of the two-dimensional photonic band structure, Phys. Rev. B 48 (11) (1993) 7767–7771.
[35] C.M. Linton, I. Thompson, Resonant effects in scattering by periodic arrays, Wave Motion 44 (2007) 165–175.
[36] N.M. Litchinitser, V.M. Shalaev, Photonic metamaterials, Laser Phys. Lett. 5 (6) (2008) 411–420.
[37] A. Mayo, Fast high order accurate solution of laplace's equation on irregular regions, SIAM J. Sci. Stat. Comput. 6 (1985) 144–157.
[38] R.C. McPhedran, N.A. Nicorovici, L.C. Botten, K.A. Grubits, Lattice sums for gratings and arrays, J. Math. Phys. 41 (2000) 7808–7816.
[39] M. Mitrea, Boundary value problems and Hardy spaces associated to the Helmholtz equation in Lipschitz domains, J. Math. Anal. Appl. 202 (1996) 819–842.
[40] A. Moroz, Exponentially convergent lattice sums, Opt. Lett. 26 (2001) 1119–1121.
[41] P. Morse, H. Feshbach, Methods of Theoretical Physics, vol. 2, McGraw-Hill, 1953.
[42] Y. Otani, N. Nishimura, A periodic FMM for Maxwell's equations in 3d and its applications to problems related to photonic crystals, J. Comput. Phys. 227 (2008) 4630–4652.
[43] D. Pissoort, E. Michielssen, F. Olyslager, D.D. Zutter, Fast analysis of 2D electromagnetic crystal devices using a periodic green function approach, J. Lightwave Technol. 23 (7) (2005) 2294–2308.
[44] V. Rokhlin, Solution of acoustic scattering problems by means of second kind integral equations, Wave Motion 5 (1983) 257–272.
[45] S. Shipman, S. Venakides, Resonance and bound states in photonic crystal slabs, SIAM J. Appl. Math. 64 (2003) 322–342.
[46] J. Smajic, C. Hafner, D. Erni, Automatic calculation of band diagrams of photonic crystals using the multiple multipole method, Appl. Comput. Electromag. Soc. J. 18 (2003) 172–180.
[47] A. Spence, C. Poulton, Photonic band structure calculations using nonlinear eigenvalue techniques, J. Comput. Phys. 204 (2005) 65–81.
[48] E. Yablonovitch, Photonic band-gap structures, J. Opt. Soc. Am. B 10 (1993) 283–295.
[49] J. Yuan, Y.Y. Lu, X. Antoine, Modeling photonic crystals by boundary integral equations and Dirichlet-to-Neumann maps, J. Comput. Phys. 227 (2008) 4617–4629.

SIAM J. SCI. COMPUT.
Vol. 32, No. 3, pp. 1417–1441
© 2010 Society for Industrial and Applied Mathematics

AN EXPONENTIALLY CONVERGENT NONPOLYNOMIAL FINITE ELEMENT METHOD FOR TIME-HARMONIC SCATTERING FROM POLYGONS*

A. H. BARNETT† AND T. BETCKE‡

Abstract. In recent years nonpolynomial finite element methods have received increasing attention for the efficient solution of wave problems. As with their close cousin the method of particular solutions, high efficiency comes from using solutions to the Helmholtz equation as basis functions. We present and analyze such a method for the scattering of two-dimensional scalar waves from a polygonal domain that achieves exponential convergence purely by increasing the number of basis functions in each element. Key ingredients are the use of basis functions that capture the singularities at corners and the representation of the scattered field towards infinity by a combination of fundamental solutions. The solution is obtained by minimizing a least-squares functional, which we discretize in such a way that a matrix least-squares problem is obtained. We give computable exponential bounds on the rate of convergence of the least-squares functional that are in very good agreement with the observed numerical convergence. Challenging numerical examples, including a nonconvex polygon with several corner singularities, and a cavity domain, are solved to around 10 digits of accuracy with a few seconds of CPU time. The examples are implemented concisely with `MPSpack`, a MATLAB toolbox for wave computations with nonpolynomial basis functions, developed by the authors. A code example is included.

Key words. scattering, polygons, Helmholtz equation, fundamental solutions

AMS subject classifications. 65N30, 35J05

DOI. 10.1137/090768667

1. Introduction.

Nonpolynomial finite element methods for wave problems such as the PUFEM [3], the least-squares finite element method [24], or the ultraweak variational formulation [8] have recently received much attention in the mathematics and engineering communities. The idea of these methods is to use Trefftz-type basis functions, i.e., exact solutions to the PDE inside each element, then solve for the coefficients which best satisfy the interelement matching and/or boundary conditions. This better captures solution behavior than standard polynomial bases, and it is closely related to the method of particular solutions or collocation methods [23]. Typical choices of basis functions are plane waves [26], Fourier–Bessel expansions [14, 27], and fundamental solutions of the Helmholtz equation [6, 16, 4].

The main advantage of nonpolynomial finite element methods over standard Galerkin methods is that they allow one to choose element sizes which are many wavelengths in diameter, while standard Galerkin methods with linear elements typically need of order 10 elements per wavelength to resolve a wave to any reasonable accuracy. Thus at high frequency the total number of degrees of freedom is much reduced, and due to the high-order (often exponential) convergence of nonpolynomial finite element methods, this advantage is even more pronounced if high accuracy is required.

*Received by the editors August 19, 2009; accepted for publication (in revised form) February 23, 2010; published electronically May 21, 2010.
http://www.siam.org/journals/sisc/32-3/76866.html
†Department of Mathematics, Dartmouth College, Hanover, NH 03755 (ahb@math.dartmouth.edu). The work of this author was supported by National Science Foundation grant DMS-0811005.
‡Department of Mathematics, University of Reading, Reading, Berkshire RG6 6AH, UK (t.betcke@reading.ac.uk). The work of this author was supported by Engineering and Physical Sciences Research Council grant EP/F067798/1.

Great progress has also been made in recent years in the development of wavenumber independent boundary integral equation methods for the solution of scattering problems (see, for example, [7, 18, 10, 13]). An excellent overview of these new methods is given in [9].

In this paper we analyze a nonpolynomial finite element method for time-harmonic scattering from a sound-soft polygonally shaped bounded obstacle $\Omega \subset \mathbb{R}^2$. That is, we are looking for the solution u of the problem

$$\Delta u + k^2 u = 0 \quad \text{in } \mathbb{R}^2 \backslash \overline{\Omega}, \tag{1.1}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{1.2}$$

$$\frac{\partial u_s}{\partial r} - iku_s = o(r^{-1/2}), \tag{1.3}$$

where $u = u_{inc} + u_s$ is the total field in which u_{inc} is the incident wave and u_s the scattered field, and where r is the radial coordinate. The wavenumber is $k = 2\pi/\lambda$, where λ is the wavelength. The Sommerfeld radiation condition (1.3) is to be understood to hold uniformly in all directions.

The main idea of the presented method is to use fractional order Fourier–Bessel functions at the corners of the polygon to match the asymptotic behavior of u there, and to represent the scattered field u_s towards infinity by fundamental solutions (i.e., effective sources), which automatically satisfy (1.3). The matching conditions between different elements are implemented using a least-squares finite element formulation. A related approach was proposed without analysis by Stojek in [27]. The main differences in our formulation are the choice of basis functions towards infinity and the setup of the discretized problem.

Least-squares finite element methods were also analyzed by Monk and Wang in [24]. Their analysis focused on Helmholtz problems on smooth domains with local absorbing boundary conditions. For such problems they also proved interior error estimates based on the least-squares functional.

In this paper we choose each element as large as possible in such a way that each of the finite subdomains contains one corner of the polygon. The accuracy is then improved by increasing the number of basis functions in each subdomain. Hence, h-refinement is not necessary. This is similar to the domain decomposition method of Descloux and Tolley for the Laplace eigenvalue problem on polygonal domains [12].

We give exponential convergence bounds with a rate that can be numerically computed using techniques from conformal maps (see also [5]). The asymptotic convergence rate is wavenumber-independent. However, we observe that the minimum number of basis functions N required to enter into the regime of exponential convergence depends linearly on k, leading to a complexity of $O(k^3)$ for the fully discrete scheme due to the need for solving dense linear systems. But due to the fast exponential convergence, and the relatively small preprocessing effort, the overall computation time for this method is very attractive and is competitive for low to medium frequencies with the wavenumber-independent approach recently introduced by Chandler-Wilde and Langdon [10]. Furthermore, the method presented here works on convex and nonconvex polygons, while the approach in [10] is restricted to convex polygons. We will demonstrate that high accuracy (up to 10 digits) is easily achievable.

Another advantage of the proposed approach compared to boundary element (or boundary integral) methods is that no quadrature of singular integrals close to the corners is necessary. Furthermore, the proposed method can be easily adapted to the sound-hard scattering case by a small change in the basis functions close to the

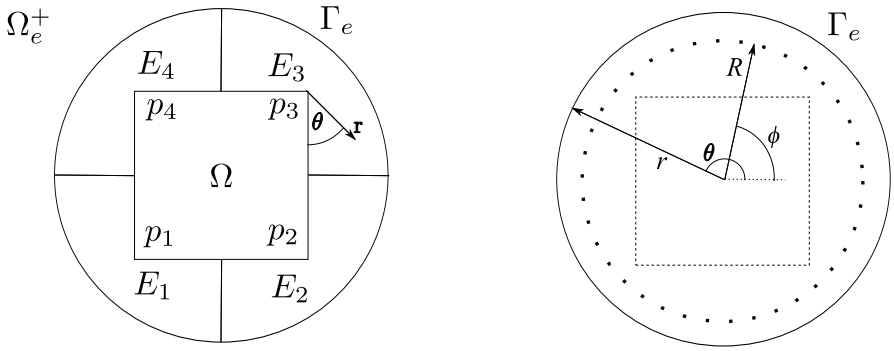


FIG. 2.1. *Geometry of subdomains, illustrated for the square scatterer. For the exterior angles $\pi/\alpha_1, \ldots, \pi/\alpha_4$ at the corners $p_1, \ldots, p_4$, we have in this example that $\alpha_1 = \cdots = \alpha_4 = 2/3$. Left plot: The local coordinate systems of the Fourier–Bessel basis functions are oriented along the boundary lines adjacent to the corners of the polygon. Right plot: Fundamental solutions basis coordinate system, showing the charge points $\boldsymbol{y}_j$ lying on a circle.*

corners. In contrast, combined integral formulations for sound-hard scattering involve the evaluation of hypersingular integral operators.

The paper is organized as follows. In section 2 we present the formulation of the method. Section 3 describes how we discretize this into a linear system. The approximation of the scattered field u_s towards infinity is done using fundamental solutions approximations, which are presented in more detail in section 4. In particular, a combined monopole-plus-dipole formulation is given that avoids problems with interior resonances. The convergence analysis in section 5 depends on approximation theoretic results for Bessel functions and an analysis of the convergence of the fundamental solutions approximations. A detailed analysis of the exponential convergence of fractional order Fourier–Bessel functions in cornered domains based on techniques from complex analysis was given by Betcke in [5]. Here, we state only the main convergence result of that paper adapted to our situation. Numerical examples are given in section 6. All computations are implemented quite simply in `MPSpack`, a MATLAB toolbox for wave computations developed by the authors; a short example code to compute scattering from the square and plot Figure 6.3 is given in Appendix A. Finally, since maintaining reasonably small coefficients is important for high accuracy, the numerical stability of the method is discussed in section 7. We conclude in section 8.

2. Formulation of the method. Let Ω be a polygonal domain with boundary Γ having q corners at the points $p_1, \ldots, p_q \in \mathbb{R}^2$ and associated exterior angles $0 < \pi/\alpha_j < 2\pi$, $j = 1, \ldots, q$. We define the artificial exterior boundary $\Gamma_e := \partial\Omega_e$ to be an analytic Jordan curve whose simply connected interior Ω_e completely contains $\overline{\Omega}$, i.e., $\operatorname{dist}(\Gamma, \Gamma_e) > 0$. We will also frequently need the exterior domain $\Omega_e^+ = \mathbb{R}^2 \backslash \overline{\Omega_e}$. Also, denote by $E := \Omega_e \backslash \overline{\Omega}$ the part of Ω_e exterior to Ω. The geometry and notation are shown in Figure 2.1.

We subdivide E into q simply connected subdomains E_i, $i = 1, \ldots, q$, with boundaries $\Gamma_i = \partial E_i$ in such a way that the following conditions are satisfied:

- $E_i \cap E_j = \emptyset$ for all $i \neq j$,
- $\bigcup_i \overline{E_i} = \overline{E}$,
- $\Gamma_i \cap \partial\Omega$ consists of two straight lines which meet at the corner p_i and whose continuation towards infinity does not intersect E_i.[1]

[1]The continuation condition is a technical condition needed for the convergence estimates on the finite subdomains (see [5, Fig. 6] for an example violating this condition).

Remark 2.1. The requirement that every subdomain E_i contain a corner of Ω is not strictly necessary and is used only to simplify the description. An extension to the case of "free" elements that have no intersection with Ω is straightforward. This is used in the cavity example in section 6.3.

In each subdomain E_i we approximate the total field u by a linear combination of Fourier–Bessel functions of the form

$$u(r,\theta) \approx \sum_{j=1}^{N_i} c_j^{(i)} J_{j\alpha_i}(kr) \sin j\alpha_i\theta. \tag{2.1}$$

The local polar coordinate systems of the Fourier–Bessel functions are shown in Figure 2.1. This guarantees that the basis functions automatically satisfy the zero boundary conditions on Γ. We denote by V_i the finite dimensional linear space spanned by all Fourier–Bessel functions of the form (2.1) in the subdomain E_i.

The approximation by Fourier–Bessel functions is motivated by the corner behavior of Helmholtz solutions. One can show that in a wedge with interior angle π/α, any Helmholtz solution u that satisfies zero Dirichlet boundary conditions on the sides of the wedge has close to the corner the absolutely convergent expansion

$$u(r,\theta) = \sum_{j=1}^{\infty} c_j J_{\alpha j}(kr) \sin \alpha j\theta$$

for certain coefficients $c_j \in \mathbb{C}$ [12]. If $\alpha \notin \mathbb{N}$, then any nontrivial u has a singularity at the corner of the wedge. Otherwise, if $\alpha \in \mathbb{N}$, then u can be analytically continued across the corner of the wedge [23].

In Ω_e^+ we approximate as follows the scattered field u_s by a linear combination of fundamental solutions whose origins lie in Ω_e. Let $\Gamma_F \subset E$ be a closed analytic Jordan curve, and choose points $\mathbf{y}_j \in \Gamma_F$. Then our approximation, which satisfies the Helmholtz equation in Ω_e^+, is

$$u_s(\mathbf{x}) \approx \sum_{j=1}^{N_e} c_j^{(e)} \left(\frac{i}{4} \frac{\partial}{\partial \nu(\mathbf{y}_j)} H_0^{(1)}(k|\mathbf{x}-\mathbf{y}_j|) - \frac{\eta}{4} H_0^{(1)}(k|\mathbf{x}-\mathbf{y}_j|) \right), \quad \mathbf{x} \in \Omega_e^+, \tag{2.2}$$

where $\nu(\mathbf{y}_j)$ is the outward-facing unit normal direction of the curve Γ_F at the point $\mathbf{y}_j$, $H_0^{(1)}$ is the outgoing Hankel function of zeroth order, and $\eta \neq 0$ is a real parameter. Typically, by a scaling argument we choose $\eta = k$. Formulation (2.2) can be interpreted as a discrete version of an integral equation of the first kind for approximating u (see section 4). The finite dimensional space of all linear combinations of the fundamental solutions in Ω_e^+ of form (2.2) is denoted by V_e.

Remark 2.2. One could alternatively approximate the scattered field u_s by a multipole expansion of the form

$$u_s(r,\theta) \approx \sum_{m=-N_e}^{N_e} \alpha_m H_m^{(1)}(kr) e^{im\theta}.$$

This was proposed by Stojek in [27]. The disadvantage of this expansion is that when Γ_e is anything other than a circle about the origin, severe numerical stability problems arise at large wavenumber due to the huge dynamic range of Hankel functions at large m. Fundamental solutions do not suffer from this problem, and hence allow more flexibility (e.g., see section 6.2).

Combining the above basis sets, the trial space V of the finite element method is the space of functions v such that $v_i := v|_{E_i} \in V_i$ for all i, and $v_e := v|_{\Omega_e^+} \in V_e$. It is useful to express the number of basis functions in each subdomain as a multiplier of a common factor N. Let $N_i = n_i N$, $i = 1, \ldots, q$, and $N_e = n_e N$. The total dimension of the trial space is then

$$N_V = \left[n_e + \sum_{i=1}^{q} n_i \right] N =: n_V N. \tag{2.3}$$

We now explain the least-squares finite element formulation as proposed by Stojek in [27] and Monk and Wang in [24]. Let Γ_{ij} be the internal boundary between two elements E_i and E_j. Let $\nu(\mathbf{x})$ be a unit normal to the curve Γ_{ij}, $i < j$, defined for almost every $\mathbf{x} \in \Gamma_{ij}$, pointing away from domain E_i. The jump of a function u defined on E_i and E_j is

$$[u](\mathbf{x}) := \lim_{\epsilon \to 0} u(\mathbf{x} + \epsilon\nu(\mathbf{x})) - u(\mathbf{x} - \epsilon\nu(\mathbf{x})), \quad \mathbf{x} \in \Gamma_{ij}.$$

Denoting by $\partial_\nu := \nabla \cdot \nu$ the normal derivative in the sense explained above, and by $[\partial_\nu u]$ its jump, define the interelement matching error functional

$$\begin{aligned} J(v) := &\sum_{i<j} \int_{\Gamma_{ij}} |[\partial_\nu v]|^2 \, ds + k^2 \, |[v]|^2 \, ds \\ &+ \sum_{i=1}^{q} \int_{\Gamma_i \cap \Gamma_e} |[\partial_\nu(\hat{u}_{inc} + v)]|^2 + k^2 \, |[\hat{u}_{inc} + v]|^2 \, ds, \end{aligned} \tag{2.4}$$

where

$$\hat{u}_{inc}(\mathbf{x}) := \begin{cases} u_{inc}(\mathbf{x}), & \mathbf{x} \in \Omega_e^+, \\ 0, & \mathbf{x} \in \Omega_e. \end{cases} \tag{2.5}$$

Here we need the restriction $\hat{u}_{inc}$ of the incoming wave u_{inc} to Ω_e^+ since inside Ω_e we directly approximate the full field u, while in Ω_e^+ we approximate the scattered field u_s.

The least-squares finite element approximation v_{LS} is now defined as

$$v_{LS} = \arg\min_{v \in V} J(v). \tag{2.6}$$

3. Implementation. In this section we give details of the numerical implementation of the least-squares formulation (2.6). Denote by $c = \left[c^{(1)}, \ldots, c^{(q)}, c^{(e)}\right]^T \in \mathbb{C}^{n_V N}$ the coefficient vector associated with all basis functions in the elements E_i, $i = 1, \ldots, q$, and Ω_e^+. Since the basis functions in each element are linearly independent, every $v \in V$ has a unique representation in terms of a vector $c \in \mathbb{C}^{n_V N}$. We can therefore reformulate (2.6) equivalently as

$$c_{LS} = \arg\min_{c \in \mathbb{C}^{N_V}} J(c).$$

For notational convenience we use the same name for the least-squares functional $J(v)$ defined in (2.4) and its parameterized version in terms of a coefficient vector c.

We now derive a matrix representation of $J(c)$. Consider first two elements E_i and E_j, within which basis functions are $g_1^{(i)}, \ldots, g_{N_i}^{(i)}$ and $g_1^{(j)}, \ldots, g_{N_j}^{(j)}$, respectively.

We assume that these functions and their derivatives are also defined on Γ_{ij}. Denote by ξ_τ, $\tau = 1, \ldots, m_{ij}$, quadrature points on Γ_{ij} with corresponding weights $\omega_\tau > 0$ appropriate for integration with respect to arc length (e.g., in section 6 we use the Clenshaw–Curtis rule). Then from Γ_{ij} the contribution to $J(c)$ is

$$\int_{\Gamma_{ij}} k^2 \left| \sum_{p=1}^{N_i} c_p^{(i)} g_p^{(i)}(s) - \sum_{p=1}^{N_j} c_p^{(j)} g_p^{(j)}(s) \right|^2 + \left| \sum_{p=1}^{N_i} c_p^{(i)} \partial_\nu g_p^{(i)}(s) - \sum_{p=1}^{N_j} c_p^{(j)} \partial_\nu g_p^{(j)}(s) \right|^2 ds$$
$$\approx \left\| \begin{bmatrix} W_{ij} & \\ & W_{ij} \end{bmatrix} \begin{bmatrix} kA_i & -kA_j \\ \partial_\nu A_i & -\partial_\nu A_j \end{bmatrix} \begin{bmatrix} c^{(i)} \\ c^{(j)} \end{bmatrix} \right\|_2^2,$$

where for each $i = 1, \ldots, q$, A_i is the matrix with elements $(A_i)_{\tau p} = g_p^{(i)}(\xi_\tau)$, and $(\partial_\nu A_i)_{\tau p} = \partial_\nu g_p^{(i)}(\xi_\tau)$ is the matrix of normal derivatives. Quadrature weights now reside in the diagonal matrix W_{ij} with elements $(W_{ij})_{\tau\tau} = \omega_\tau^{1/2}$, $\tau = 1, \ldots, m_{ij}$. On each outer boundary the contribution to $J(c)$ is

$$\int_{\Gamma_i \cap \Gamma_e} k^2 \left| \sum_{p=1}^{N_i} c_p^{(i)} g_p^{(i)}(s) - \sum_{p=1}^{N_e} c_p^{(e)} g_p(s) - u_{inc}(s) \right|$$
$$+ \left| \sum_{p=1}^{N_i} c_p^{(i)} \partial_\nu g_p^{(i)}(s) - \sum_{p=1}^{N_e} c_p^{(e)} \partial_\nu g_p(s) - \partial_\nu u_{inc}(s) \right|^2 ds$$
$$\approx \left\| \begin{bmatrix} W_{ie} & \\ & W_{ie} \end{bmatrix} \left(\begin{bmatrix} kA_i & -kA_e \\ \partial_\nu A_i & -\partial_\nu A_e \end{bmatrix} \begin{bmatrix} c^{(i)} \\ c^{(j)} \end{bmatrix} - \begin{bmatrix} b_i \\ \tilde{b}_i \end{bmatrix} \right) \right\|_2^2,$$

where g_ℓ is the ℓth fundamental solutions basis function with source at $\mathbf{y}_\ell$, and, analogous to the above, $(A_e)_{\tau p} = g_p(\xi_\tau)$, $(\partial_\nu A_e)_{\tau p} = \partial_\nu g_p(\xi_\tau)$, and quadrature points now live on $\Gamma_i \cap \Gamma_e$. The inhomogeneous vector is filled with the known values $(b_i)_\tau = u_{inc}(\xi_\tau)$ and $(\tilde{b}_i)_\tau = \partial_\nu u_{inc}(\xi_\tau)$ for $i = 1, \ldots, m_{ie}$.

Note that, by construction, all of the above quadratures involve smooth (in fact analytic) functions on the respective segments; no quadrature near the corner singularities is needed. However, the integrand is oscillatory on the wavelength scale, and this will inform the choice of the number of quadrature nodes.

Stacking up all contributions from the different interfaces, we obtain

$$J(c) \approx \| W\,(Ac - b)\,\|_2^2.$$

The matrix A is built up from the subblocks A_i and $\partial_\nu A_i$ as described above, and the right-hand side vector b is zero apart from the contributions of the incident wave at the quadrature points on Γ_e. In [27, 24, 17] the authors discretize the variational derivative of $J(v)$ to obtain a linear system of equations that they solve. This approach is equivalent to solving the normal equation

$$A^H W^2 A c = A^H W^2 b. \tag{3.1}$$

However, as long as the problem size is not too large we can, and do, equivalently solve directly the least squares problem

$$\min_{c \in \mathbb{C}^{N_V}} \| W\,(Ac - b)\,\|_2 \tag{3.2}$$

using a dense least-squares solver, such as the backslash (`mldivide`) in MATLAB. This is numerically more stable than solving (3.1) if WA is ill-conditioned.

4. The method of fundamental solutions for scattering problems. In this section we give a brief introduction to the method of fundamental solutions (MFS) for Helmholtz problems in exterior domains and motivate the combined MFS formulation (2.2). Overviews of the MFS, its theory, and applications are given in [6, 15]. The results of this section form the basis for the convergence analysis in section 5.1.

Let D be a simply connected domain with analytic boundary ∂D. Let u_s be the unique solution [11] to the Helmholtz equation (1.1) in $\mathbb{R}^2\backslash\overline{D}$ and the Sommerfeld condition (1.3) with boundary value data $u_s = f$ on ∂D. The general idea of the MFS is to choose charge points in D, with strengths (coefficients), whose combined field can well approximate u_s in $\mathbb{R}^2\backslash\overline{D}$ as the number of charges grows. We focus on the special case where these charges lie on a closed curve $\Gamma_F \subset D$ some distance inside D. Then the MFS may be viewed as a quadrature approximation to a single-layer potential of the form

$$u_s(\mathbf{x}) \approx \int_{\Gamma_F} \frac{i}{4} H_0^{(1)}(k|\mathbf{x}-\mathbf{y}|) g(\mathbf{y}) ds_{\mathbf{y}}. \tag{4.1}$$

However, the MFS is more general than merely a discretization of a standard boundary integral representation: depending on the smoothness of u_s (or singularities of its continuation into D), there might not exist any layer density g that can represent u_s exactly. Nevertheless there still may exist a sequence of densities in suitable spaces whose field converges to u_s. The discrete version of this statement is that, with the MFS, there exists a sequence of charge points and strengths whose field converges to u_s. If Γ_F is suitably chosen, one can show that there are such sequences that give rapid exponential convergence in certain norms (see [4] and our section 5.1).

Unlike boundary integral methods which enable a second kind formulation, (4.1) is necessarily first kind. However, it has the advantages of easily provable convergence rate, of simplicity of evaluation arbitrarily close to ∂D, and of no singular quadrature schemes being required. Disadvantages of the MFS are that the right choice of Γ_F and the quadrature points $\mathbf{y}_j$ on Γ_F can be difficult for complicated domains, and that the MFS typically leads to ill-conditioned linear systems. In our case the exterior boundary Γ_e is typically a very simple shape (circles or simple analytic curves), making it easy to automate the choice of Γ_F, and the effects of the ill-conditioning can be controlled in such a way that computing u to many digits of accuracy is still possible (see section 7).

We now focus on the case that D is a disk with radius r and Γ_F is a circle with radius $R < r$ (see Figure 2.1, which shows this geometry for the case $D = \Omega_e$ that we will use in section 5.1). We identify the plane $\mathbb{R}^2$ with $\mathbb{C}$. Define the single-layer operator $S : L^2[0, 2\pi] \to L^2[0, 2\pi]$ by

$$(Sg)(\theta) = \int_0^{2\pi} \frac{i}{4} H_0^{(1)}(k|re^{i\theta} - Re^{i\phi}|) g(\phi) d\phi.$$

The task is to find a density g such that the boundary data is matched, that is, $Sg \approx f$. For the interior Helmholtz problem on the disk, this was analyzed in [4]. Here, we focus on the eigenvalue expansion of S in order to reveal some stability properties for exterior problems and to lay the groundwork for section 5.1.

Using Graf's addition formula [1, eq. 9.1.79] we obtain

$$\frac{i}{4} H_0^{(1)}(k|re^{i\theta} - Re^{i\phi}|) = \sum_{m\in\mathbb{Z}} \frac{i}{4} H_m^{(1)}(kr) e^{-im\phi} J_m(kR) e^{im\theta}.$$

Defining the Fourier coefficient $\hat{g}(m)$ of a function $g \in L^2[0,2\pi]$ by $\hat{g}(m) = \frac{1}{2\pi}\int_0^{2\pi} g(\phi)e^{-im\phi}d\phi$, it follows that

$$(Sg)(\theta) = \sum_{m\in\mathbb{Z}} \hat{s}(m)\hat{g}(m)e^{im\theta},$$

where the eigenfunctions of S are the complex exponentials, and the eigenvalues of S are

$$\hat{s}(m) = \frac{i\pi}{2}H_m^{(1)}(kr)J_m(kR).$$

If kR hits a zero of the Bessel function of order m, then $\hat{s}(m) = 0$. This occurs if and only if k^2 is a Dirichlet eigenvalue of the disk of radius R. But the exterior Helmholtz problem has a unique solution for all $k > 0$, and therefore we need a formulation that is not affected by interior resonances. We achieve this by a combined formulation analogous to those for boundary integral equations [11].

We define the modified operator $S_\eta : L^2[0,2\pi] \to L^2[0,2\pi]$ by

(4.2)
$$(S_\eta g)(\theta) = \frac{i}{4}\int_0^{2\pi}\left(\frac{\partial}{\partial R}H_0^{(1)}(k|re^{i\theta} - Re^{i\phi}|) + i\eta H_0^{(1)}(k|re^{i\theta} - Re^{i\phi}|)\right)g(\phi)d\phi.$$

This combines single- and double-layer densities. Its eigenvalues $\hat{s}_\eta(m)$ are easily computed as

$$\begin{aligned}\hat{s}_\eta(m) &= \frac{\pi}{2}H_m^{(1)}(kr)\left[i\frac{\partial}{\partial R}J_m(kR) - \eta J_m(kR)\right] \\ &= \frac{\pi}{2}H_m^{(1)}(kr)\left[\frac{ik}{2}\left(J_{m-1}(kR) - J_{m+1}(kR)\right) - \eta J_m(kR)\right].\end{aligned} \tag{4.3}$$

LEMMA 4.1. *Let $\eta \in \mathbb{R}\backslash\{0\}$. Then $\hat{s}_\eta(m) \neq 0$ for all $m \in \mathbb{Z}$.*

Proof. For Hankel functions of the first kind, it holds that $|H_m^{(1)}(z)| \neq 0$ for all $z \in \mathbb{C}$ [25, Chap. 7]. Then from the first expression in (4.3) we have $\hat{s}_\eta(m) = 0$ if and only if $\frac{\partial}{\partial R}J_m(kR) = 0$ and $J_m(kR) = 0$. But this is not possible, since otherwise by Green's representation theorem the function $v(r,\theta) := J_m(kr)e^{im\theta}$ would be identically zero in the disk with radius R. ∎

We now claim that the eigenvalues $\hat{s}_\eta(m)$ and their derivatives with respect to the outer disc radius r decay exponentially with a rate that depends on the radius ratio of R to r.

LEMMA 4.2. *For $\eta \neq 0$ and $\epsilon > 0$ arbitrarily small there exist constants $c_s > 0$ and $C_s > 0$ such that for $m \in \mathbb{Z}$, both the following hold:*

$$c_s\left(\frac{r}{R}\right)^{-|m|} \leq |\hat{s}_\eta(m)| \leq C_s\left(\frac{r}{R}\right)^{-|m|}, \tag{4.4}$$

$$\left|\frac{\partial}{\partial r}\hat{s}_\eta(m)\right| \leq C_s\left[\left(\frac{r}{R}\right) - \epsilon\right]^{-|m|}, \tag{4.5}$$

where c_s and C_s depend on k, R, and r, and C_s additionally depends on ϵ.

Proof. Consider the three terms in (4.3). Large-order asymptotics for Bessel functions [1, eq. 9.3.1] yield

$$J_m(kR) \sim \frac{1}{\sqrt{2\pi m}}\left(\frac{ekR}{2m}\right)^m, \quad Y_m(kr) \sim -\sqrt{\frac{2}{\pi m}}\left(\frac{ekr}{2m}\right)^{-m}$$

for fixed z and $m \to \infty$. Since $H_m(z) = J_m(z) + iY_m(z)$ it follows that $H_m(z) \sim iY_m(z)$, and therefore

$$H_m^{(1)}(kr)J_m(kR) \sim -\frac{i}{\pi m}\left(\frac{R}{r}\right)^m,$$
$$H_m^{(1)}(kr)J_{m-1}(kR) \sim -\frac{2i}{ekR\pi}\left(\frac{R}{r}\right)^m,$$
$$H_m^{(1)}(kr)J_{m+1}(kR) \sim -\frac{ekRi}{2\pi}\frac{1}{m^2}\left(\frac{R}{r}\right)^m.$$

Inserting these into (4.3), we get

$$\hat{s}_\eta(m) \sim \frac{1}{2eR}\left(\frac{R}{r}\right)^m - \frac{Rek^2}{8m^2}\left(\frac{R}{r}\right)^m + \frac{i\eta}{2m}\left(\frac{R}{r}\right)^m \sim \frac{1}{2eR}\left(\frac{R}{r}\right)^m.$$

Together with the reflection laws $J_{-m}(z) = (-1)^m J_m(z)$ and $H_{-m}^{(1)}(z) = (-1)^m H_m^{(1)}(z)$, the upper bound in (4.4) follows. Furthermore, by Lemma 4.1 there exists $c_s > 0$ such that also the lower bound in (4.4) is valid. We apply similar estimates to the expression

(4.6)
$$\frac{\partial}{\partial r}\hat{s}_\eta(m) = \frac{k\pi}{4}\left[H_{m-1}^{(1)}(kr) - H_{m+1}^{(1)}(kr)\right]\left[\frac{ik}{2}\left(J_{m-1}(kR) - J_{m+1}(kR)\right) - \eta J_m(kR)\right]$$

to obtain

$$\frac{\partial}{\partial r}\hat{s}_\eta(m) \sim Cm\left(\frac{r}{R}\right)^{-|m|}$$

for some constant $C \in \mathbb{C}$. The dominant contribution is $H_{m+1}^{(1)}(kr)J_{m-1}(kR)$. Absorbing the algebraic factor m into the exponential bound and choosing C_s sufficiently large yields (4.5). ∎

The MFS is now obtained by applying to the above layer potential formulation the discrete sources ansatz $g(\phi) = \sum_{j=1}^N c_j\delta(\phi - \phi_j)$, where $\phi_j = \frac{2\pi j}{N}$, $j = 1, \ldots, N$. It follows that

$$(S_\eta g)(\theta) = \sum_{j=1}^N c_j\left[\frac{i}{4}\frac{\partial}{\partial R}H_0^{(1)}(k|re^{i\theta} - Re^{i\phi_j}|) - \frac{\eta}{4}H_0^{(1)}(k|re^{i\theta} - Re^{i\phi_j}|)\right]. \tag{4.7}$$

Note $g \notin L^2[0, 2\pi]$, but S_η has a smooth bounded kernel so is well defined with g merely integrable.

We will also need the Fourier series for this choice of g, which we note converges only distributionally. Denote by $\hat{c}$ the discrete Fourier transform of the coefficient vector $c = \left[c_1, \ldots, c_N\right]$ defined by

$$\hat{c}_s = \frac{1}{N}\sum_{j=1}^N c_j e^{-is\phi_j}, \quad -\frac{N}{2} < s \le \frac{N}{2}.$$

Then applying our ansatz to the definition of the Fourier coefficients,

$$\hat{g}(m) = \frac{1}{2\pi}\sum_{j=1}^N c_j e^{im\phi_j} = \frac{N}{2\pi}\hat{c}_{(m \bmod N)}, \tag{4.8}$$

where $m \bmod N$ denotes the unique integer lying in the range $-N/2+1, \ldots, N/2$ which differs from m by an integer multiple of N. This "wrapping" of the vector $\hat{c}$ is due to the delta-comb in g.

Equation (4.7) is a special case of the general monopole-plus-dipole formulation (2.2), with Γ_F circular and the evaluation restricted to the concentric circle ∂D. The basis functions in (2.2), viewed as functions on $\mathbb{R}^2 \backslash D$, automatically satisfy the Sommerfeld radiation conditions. Thus the only remaining way to solve the boundary value problem laid out at the start of this section would be to choose the coefficients c_j to match the boundary data f, e.g., by collocation or a least-squares method on ∂D. We remind the reader that in this work we instead couple the MFS (with the choices $D = \Omega_e$ and $N = N_e$) to other basis representations in E.

5. Convergence analysis. Let u be the exact solution of (1.1)–(1.3) with given u_{inc}. Since u and its derivative are continuous across element boundaries, we may study the convergence of approximations to u *independently* in each subdomain. In particular, we will make use of the rate of approximation by Fourier–Bessel functions in each subdomain E_i, and the rate of approximation on the boundary Γ_e by fundamental solutions. Convergence to u then automatically leads to a minimization of the jump on the interfaces, as we now show. Using the triangle inequality and estimating the boundary L^2 norm by the L^∞ norm over the whole domain, we estimate (2.4) by

$$
\begin{aligned}
J(v) &= \sum_{i<j} \int_{\Gamma_{ij}} |[\partial_\nu v - \partial_\nu u]|^2 ds + k^2 |[v-u]|^2 ds \\
&\quad + \sum_{i=1}^{q} \int_{\Gamma_i \cap \Gamma_e} |[\partial_\nu(\hat{u}_{inc} + v - u)]|^2 + k^2 |[\hat{u}_{inc} + v - u]|^2 ds \\
&\le C_1 \sum_i \left\{ k^2 \|u_s - v_e\|^2_{L^2(\Gamma_i \cap \Gamma_e)} + \|\partial_\nu u_s - \partial_\nu v_e\|^2_{L^2(\Gamma_i \cap \Gamma_e)} \right\} \\
&\quad + C_2 \left\{ k^2 \sum_i \|u - v_i\|^2_{L^\infty(E_i)} + \sum_{i<j} \|\partial_\nu u - \partial_\nu v_i\|^2_{L^\infty(\Gamma_{ij})} \right\},
\end{aligned}
\tag{5.1}
$$

where $v \in V$ (recalling the definitions of v_i, v_e from section 2), and C_1, C_2 are mesh-dependent constants.

The analysis of the rate of convergence of approximating u by $v_i \in V_i$ in each subdomain E_i is based on transforming the approximation problem involving Bessel functions into a polynomial approximation problem in the complex plane. For this it is useful to work in the L^∞ norm. On Γ_e the rate of convergence of the fundamental solutions will be analyzed using Fourier series estimates, for which the natural space is L^2.

The convergence analysis for approximating solutions of $\Delta u + \lambda u = 0$ with fractional Bessel functions on domains with one corner singularity and zero Dirichlet boundary conditions at the wedges forming the singular corner was given by Betcke in [5]. The only notable new feature here is that now complex rather than real functions are approximated. This can easily be achieved by splitting the approximation problem on the subdomains E_i into separate problems for the approximation of the real and imaginary parts of u.

The following theorem summarizes the results of [5, Thm. 5.2 and Lem. 5.5] adapted to our notation in this paper. We refer to [5, sect. 5] for the proof.

THEOREM 5.1. *For every element E_i, $i = 1, \ldots, q$, the following statement holds.*

There exists $\rho_i > 1$ such that, for every $1 < \tau < \rho_i$ and each interface Γ_{ij} between E_i and a neighboring subdomain E_j, it holds that

$$\min_{v \in V_i} \left(\|\nabla u - \nabla v\|^2_{L^\infty(\Gamma_{ij})} + k^2 \|u - v\|^2_{L^\infty(E_i)} \right) = O(\tau^{-2N_i})$$

as $N_i \to \infty$.

The rates ρ_i may be computed; they are the conformal distance of the nearest singularity in u to (a conformal map of) the domain E_i [5].

To estimate the convergence on Γ_e of the fundamental solutions approximation to the scattered field u_s, we consider (as in section 4) only the case of concentric circles.[2] The source points for the fundamental solutions are given by $\mathbf{y}_j = Re^{i\phi_j}$, $j = 1, \ldots, N_e$, with $\phi_j = \frac{2\pi j}{N_e}$, and the exterior circle Γ_e has radius r. We impose $\max_i |p_i| < R < r$. The proof of the following theorem is given in section 5.1.

THEOREM 5.2. *Let $\epsilon > 0$ and $\tilde{\epsilon} > 0$ be arbitrarily small. Define $\rho := \min_i \frac{r}{|p_i|}$ and $\rho_\epsilon := \rho - \epsilon$. Let ν be the outward normal direction at the disk with radius r and assume that N_e is even. Then*

(5.2)

$$\min_{v \in V_e} \left(\|\partial_\nu u_s - \partial_\nu v\|^2_{L^2(\Gamma_e)} + k^2 \|u_s - v\|^2_{L^2(\Gamma_e)} \right) = \begin{cases} O\left(\left(\frac{r}{R} - \tilde{\epsilon} \right)^{-2N_e} \right), & \frac{r}{R} < \rho_\epsilon^{\frac{1}{2}}, \\ O\left(\rho_\epsilon^{-N_e} \right), & \frac{r}{R} > \rho_\epsilon^{\frac{1}{2}} \end{cases}$$

as $N_e \to \infty$.

Remark 5.3. If we consider only the rate of convergence for approximating the boundary value of u_s by $v \in V_e$ and not the normal derivatives, then $\tilde{\epsilon} = 0$ is possible.

Combining (5.1), Theorem 5.1, and Theorem 5.2, we immediately obtain the following.

THEOREM 5.4. *Let $\rho_1, \ldots, \rho_q$ be the measures of the exponential convergence rates given in Theorem 5.1, let $\rho_e := \min\left\{ \left(\frac{r}{R}\right)^2, \min_i \frac{r}{|p_i|} \right\}$, and let $n_1, \ldots, n_q$ and n_e be the multipliers for the number of basis functions as defined in (2.3). Assume that n_e is even, and let $\omega > 1$ be defined by*

$$\omega^2 := \min\{\rho_1^{2n_1}, \ldots, \rho_q^{2n_q}, \rho_e^{n_e}\}.$$

Then

$$\min_{v \in V} J(v) = O\left((\omega - \epsilon)^{-2N} \right)$$

for any $\epsilon > 0$ arbitrarily small.

Remark 5.5. An ideal choice of the multipliers $n_1, \ldots, n_q$ and n_e is such that

$$\rho_1^{2n_1} \approx \cdots \approx \rho_q^{2n_q} \approx \rho_e^{n_e}.$$

This choice keeps to a minimum the number of basis functions required for a certain accuracy.

[2] The convergence analysis of the MFS for integral equations with periodic kernels on more general domains has recently been considered by Kangro [22]. To understand the method presented in this paper, the case of the MFS in the exterior of a disk is sufficient for many applications and can be treated completely using arguments from Fourier analysis.

Remark 5.6. It is an open question whether there exists $C > 0$ such that

$$\|u - v\|^2_{L^2(E)} \leq CJ(v). \tag{5.3}$$

In [24, Thm. 3.1] such an estimate was proved for the case that Ω is a sufficiently smooth domain and Γ_e has the impedance condition $\partial_\nu u - iku = 0$. The crucial difference here is that we use fundamental solutions approximations instead of local absorbing boundary conditions. Similar estimates have also been investigated by Hiptmair, Moiola, and Perugia in the context of plane wave discontinuous Galerkin methods [21]. In section 6 we present numerical examples which support the conjecture (5.3).

5.1. Convergence of the MFS approximations. In this section we give a proof of Theorem 5.2. The proof proceeds along similar lines as the proof of [4, Thm. 3]. We proceed by finding a specific function $v \in V_e$, for which we show that (5.2) holds. In the following, $C > 0$ will always denote an unspecified constant that may change in the course of the derivation and that depends on k, R, r, and ϵ but not on N_e.

Let $t[v] := \|u_s - v\|_{L^2(\Gamma_e)}$ for $v \in V_e$. The multipole (Laurent-type) representation of u_s is

$$u_s(r, \theta) = \sum_{m \in \mathbb{Z}} \alpha_m H^{(1)}_m(kr) e^{im\theta}$$

for $r > \max_i |p_i|$ and $\theta \in [0, 2\pi]$. It follows immediately that the Fourier coefficients $\hat{u}_s(m)$ of the angular function $u_s(r, \cdot)$ are

$$\hat{u}_s(m) = \alpha_m H^{(1)}_m(kr).$$

The following lemma gives a bound on the decay of $\hat{u}_s(m)$ as $m \to \infty$.

LEMMA 5.7. *Let* $\rho := \min_i \frac{r}{|p_i|}$ *and* $\rho_\epsilon := \rho - \epsilon$. *Then*

$$|\hat{u}_s(m)| \leq C\rho_\epsilon^{-|m|} \tag{5.4}$$

for any arbitrarily small $\epsilon > 0$, *where* C *depends on* k, R, r, *and* ϵ.

Proof. Let $f(t) := u_s(r, 2\pi t)$ for $t \in [0, 1]$. We need to show that f can be analytically continued as a holomorphic function to the strip $t + i\tau$ for $|\tau| < \frac{1}{2\pi} \log \rho$. Denote by $\tilde{p} := \arg\max_i |p_i|$ the corner of $\partial\Omega$ with largest absolute value. Reparameterize the scattered field $u_s(x, y)$ in the Cartesian (x, y) coordinates as $U_s(z, \overline{z}) := u_s(x, y)$, where $z = x + iy$. Let $\Omega_{\tilde{p}} := \{z : |\tilde{p}| < |z| < \frac{r^2}{|\tilde{p}|},\ \mathrm{Im}\{z\} > 0\}$ be a half annulus, and define $\Omega^*_{\tilde{p}} := \{z : \ \overline{z} \in \Omega_{\tilde{p}}\}$. By a result from Vekua [19, 28], U_s can be analytically continued as a holomorphic function of two independent complex variables z and z^* into the whole of $\Omega_{\tilde{p}} \times \Omega^*_{\tilde{p}}$, such that $U_s(z, z^*) = u_s(x, y)$ if and only if $z^* = \overline{z}$. From the definition of f it follows that $f(t) = U_s(re^{i2\pi t}, re^{-i2\pi t})$. Hence, by the analyticity of U_s we have that $f(t + i\tau) = U_s(re^{-2\pi\tau + i2\pi t}, re^{2\pi\tau - i2\pi t})$ is analytic for $0 < t < 1/2$ and $-\frac{1}{2\pi} \log \rho < \tau < \frac{1}{2\pi} \log \rho$. By rotating the half annulus and repeating the argument, the analyticity in this strip can be extended to the whole interval $t \in [0, 1]$. The estimate (5.4) now follows from the decay of Fourier coefficients of analytic functions (see, for example, [20, sect. 13.2]). ∎

The Fourier series representation of $v \in V_e$, associated with density g, restricted to the circle of radius r is given by (see section 4)

$$\hat{v}(m) = \widehat{S_\eta g}(m) = \hat{s}_\eta(m)\hat{g}(m),$$

where $g(\theta) = \sum_{j=1}^{N_e} c_j \delta(\phi - \phi_j)$ and $\hat{g}(m) = \frac{N_e}{2\pi} \hat{c}_{(m \bmod N_e)}$.

By reparameterizing the L^2 integral over Γ_e and using Parseval's identity, it follows that

$$t[v]^2 = 2\pi r \sum_{m \in \mathbb{Z}} |\hat{u}_s(m) - \hat{v}(m)|^2. \tag{5.5}$$

We now take a specific $v \in V_e$ by choosing the MFS coefficients of v such that for all $-N_e/2 < m \leq N_e/2$, we have

$$\hat{g}(m) = \frac{\hat{u}_s(m)}{\hat{s}_\eta(m)} = \frac{\alpha_m H_m^{(1)}(kr)}{\frac{\pi}{2} H_m^{(1)}(kr) \left[\frac{ik}{2}\left(J_{m-1}(kR) - J_{m+1}(kR)\right) - \eta J_m(kR)\right]}. \tag{5.6}$$

Then the terms involving $-N_e/2 < m \leq N_e/2$ cancel out in (5.5) and we can estimate

$$t[v]^2 \leq C \left(\sum_{m \notin [-\frac{N_e}{2}+1, \ldots, \frac{N_e}{2}]} |\hat{u}_s(m)|^2 + \sum_{m \notin [-\frac{N_e}{2}+1, \ldots, \frac{N_e}{2}]} |\hat{s}_\eta(m)\hat{g}(m)|^2 \right) =: C(E_u^2 + E_s^2).$$

Using (5.4) we can now estimate E_u^2 as

$$E_u^2 \leq C \sum_{m \notin [-\frac{N_e}{2}+1, \ldots, \frac{N_e}{2}]} \rho_\epsilon^{-2|m|} \leq C \rho_\epsilon^{-N_e}. \tag{5.7}$$

In order to estimate E_s^2 we rewrite it as

$$E_s^2 = \sum_{n=-\frac{N_e}{2}+1}^{\frac{N_e}{2}} \hat{g}(n)^2 \sum_{b \in \mathbb{Z} \setminus \{0\}} \hat{s}_\eta(bN_e + n)^2.$$

Bounding the inner sum and substituting $|\hat{g}(n)| = \left|\frac{\hat{u}_s(n)}{\hat{s}_\eta(n)}\right| \leq C\left(\frac{r}{R\rho_\epsilon}\right)^{|n|}$ gives

$$E_s^2 \leq C \sum_{n=-\frac{N_e}{2}+1}^{\frac{N_e}{2}} |\hat{g}(n)|^2 \left(\frac{r}{R}\right)^{-2N_e+2|n|} \leq C \left(\frac{r}{R}\right)^{-2N_e} \sum_{n=-\frac{N_e}{2}+1}^{\frac{N_e}{2}} \left(\frac{r^2}{R^2\rho_\epsilon}\right)^{2|n|}. \tag{5.8}$$

If $\rho_\epsilon > \frac{r^2}{R^2}$, it follows from (5.8) that $E_s^2 \leq C\left(\frac{r}{R}\right)^{-2N_e}$. This decays slower than the estimate in (5.7). Hence, we have $t[v]^2 \leq C\left(\frac{r}{R}\right)^{-2N_e}$. If $\rho_\epsilon < \frac{r^2}{R^2}$, it follows from (5.8) that

$$E_s^2 \leq C \left(\frac{r}{R}\right)^{-2N_e} \left(\frac{r^2}{R^2\rho_\epsilon}\right)^{N_e} = C\rho_\epsilon^{-N_e},$$

which has the same exponential rate of decay as (5.7), and therefore $t[v]^2 \leq C\rho_\epsilon^{-N_e}$ in this case.

We now show that up to an arbitrarily small correction in rate $\tilde{\epsilon} > 0$ the error $t'[v] := \|\partial_\nu u_s - \partial_\nu v\|_{L^2(\Gamma_e)}$ has the same exponential bounds as $t[v]$. Similarly to (5.5) we have

$$t'[v] = 2\pi r \sum_{m \in \mathbb{Z}} |\widehat{\partial_\nu u_s}(m) - \widehat{\partial_\nu v}(m)|^2, \tag{5.9}$$

where

$$\widehat{\partial_\nu u_s}(m) = \frac{\partial}{\partial r}\hat{u}_s(m)$$

and

$$\widehat{\partial_\nu v}(m) = \frac{\partial}{\partial r}\hat{v}(m) = \left(\frac{\partial}{\partial r}\hat{s}_\mu(m)\right)\hat{g}(m).$$

The second equality follows from the fact that $\hat{g}(m)$ is independent of r (see (5.6)). Furthermore, from (5.6) it follows that

$$\frac{\frac{\partial}{\partial r}\hat{u}_s(m)}{\frac{\partial}{\partial r}\hat{s}_\mu(m)} = \frac{\hat{u}_s(m)}{\hat{s}_\eta(m)} = \hat{g}(m)$$

for $-N/2 < m \leq N/2$. Hence, the Fourier coefficients involving $-N/2 < m \leq N/2$ also cancel out for $t'[v]$. By Lemma 4.2, $\frac{\partial}{\partial r}\hat{s}_\eta(m)$ has, up to an arbitrarily small rate correction $\tilde{\epsilon}$, the same exponential upper bound as $\hat{s}_\eta(m)$. Also, since $\frac{\partial}{\partial r}u_s$ has its only possible singularities at the corners p_i, its Fourier coefficients $\frac{\partial}{\partial r}\hat{u}_s(m)$ have the same exponential bound (5.4) as those of u_s. We can therefore apply the same arguments to estimate $t'[v]$ as we did for $t[v]$ and arrive (up to $\tilde{\epsilon}$ coming from the bound on $\frac{\partial}{\partial r}\hat{s}_\eta(m)$) at the same exponential bounds as for $t[v]$, which concludes the proof of Theorem 5.2.

6. Numerical examples. In this section we demonstrate the numerical properties of the method and present computational results for scattering on a square, a snowflake domain, and a cavity. All computations were performed using MATLAB 2009a on a dual processor workstation with two 2.8 GHz Intel Xeon quad-core processors and 64GB RAM. However, the MATLAB multithreading rarely utilized more than one CPU core. Hence, the timings will be comparable on a standard desktop machine.

6.1. Scattering from a square. Let Ω be the unit square given by $\Omega = (-0.5, 0.5)^2$. We use a domain decomposition as shown in Figure 2.1. Γ_e is a circle with radius r, and the source points for the fundamental solutions are given by $\mathbf{y}_j = Re^{i\phi_j}$, $\phi_j = \frac{2\pi j}{N_e}$, $j = 1, \ldots, N_e$. We found that the fundamental solutions basis size $N_e := 2N$ was sufficient, where N is the number of Fourier–Bessel sine functions in each element E_i. For the fundamental solutions we use the combined formulation (2.2) with $\eta = k$. For quadrature we used the Clenshaw–Curtis rule with a number of nodes chosen high enough that further increases had negligible effect; typically this number is a small multiple of N per interface Γ_{ij}. (Code for this setup is given in Appendix A.)

Rate of convergence. In Figure 6.1 we show the rate of convergence of $t[v_{LS}] := J(v_{LS})^{1/2}$, which measures the norm of the jumps in function values and normal derivatives on the interfaces. We fix $R = 0.8$ and the wavenumber $k = 1$. The left plot shows the case $r = 0.9$ and the right plot $r = 1.3$. The measured convergence (solid-dotted line) for the latter is much faster than for the former. In both cases we have also plotted the asymptotic bound ω^{-N} from Theorem 5.4. The asymptotic estimates for the convergence on the finite elements $E_1, \ldots, E_4$ can be computed using numerical conformal mapping techniques (see [5] for more details). The convergence of the MFS approximations is directly given by Theorem 5.2.

What is the optimal value of r? If r is small, then the singular corners of the square have a large relative distance to the neighboring elements, suggesting a fast convergence on these elements. However, the rate of convergence for the MFS approximations will be slow since the radius r is close to the radius of the corners of the polygon. If r is large, then the MFS approximations converge fast but the relative distance of a singular corner to a neighboring element E_i is very small, leading to slow convergence on the finite elements. This suggests that there is an optimal radius $r = r_{opt}$, which is achieved when the asymptotic rate of convergence of the MFS is identical to the asymptotic rate of convergence on the finite elements.

In the left plot of Figure 6.2 we compute (using conformal mapping) the asymptotic exponential convergence factor ω from Theorem 5.4 as a function of the radius r of the outer circle Γ_e. The optimum asymptotic rate is achieved for $r \approx 1.036$. In that case the rates of convergence on the finite elements E_j and on Ω_e^+ are almost identical, with a rate of approximately 1.465^{-N}. The right plot of Figure 6.2 shows the convergence of $t[v_{LS}]$ in this case. For $N = 80$ the value of $t[v_{LS}]$ is close to machine precision.

Sound-hard boundary conditions. In this paper we have analyzed the case of sound-soft boundary conditions for u, that is, $u = 0$ on $\partial\Omega$. However, the scheme presented in this paper can be trivially extended to sound-hard boundary conditions, that is, $\partial_\nu u = 0$, where ∂_ν denotes the outward pointing normal derivative on $\partial\Omega$. For this we switch from Fourier–Bessel sine function approximations to Fourier–Bessel cosine function approximations in each E_i, replacing (2.1) by

$$u(r,\theta) \approx \sum_{j=0}^{N_i} c_j^{(i)} J_{j\alpha_i}(kr) \cos j\alpha_i\theta.$$

It is simple to check that these functions satisfy the homogeneous boundary conditions for the normal derivative. The fundamental solutions basis functions in Ω_e^+ do not change. This is a great advantage compared to combined integral equation formulations, which usually involve the approximation of a hypersingular integral operator for the sound-hard scattering case.

Timing results. In Figure 6.3 we show the real part of the full field for the sound-soft scattering case (left plot) and the sound-hard scattering case (right plot) for $k = 50$, i.e., about 8 wavelengths on a side. The incident field is a plane wave with direction $-\frac{\pi}{6}$.

The solution time was approximately 11 seconds in both cases. This includes the setup of the problem and the time for solving the least-squares problem $\min_c \|W(Ac-b)\|_2$, where A is of dimension 4816×600 (sound-soft case), or 4816×604 (sound-hard case), corresponding to $N = 100$ and $m_{ij} = 200$ Clenshaw–Curtis quadrature points on each interface Γ_{ij}. In the sound-soft case we have $t[v_{LS}] \approx 1.3 \times 10^{-10}$ and in the sound-hard case $t[v_{LS}] \approx 9.2 \times 10^{-11}$. Creating the solution plots takes longer, since the solution needs to be evaluated on a large number of grid points.

6.2. Scattering from a snowflake domain. We now show results for a snowflake domain, which is nonconvex but star-shaped. The domain and the corresponding decomposition are shown in the top two plots of Figure 6.4. Note that the external boundary Γ_E is not a circle but an equipotential line of the exterior (Laplace) Green's function of the domain. Although we have given the analysis only for Γ_E being a circle, we demonstrate in this example that domain-adapted analytic curves also work well.

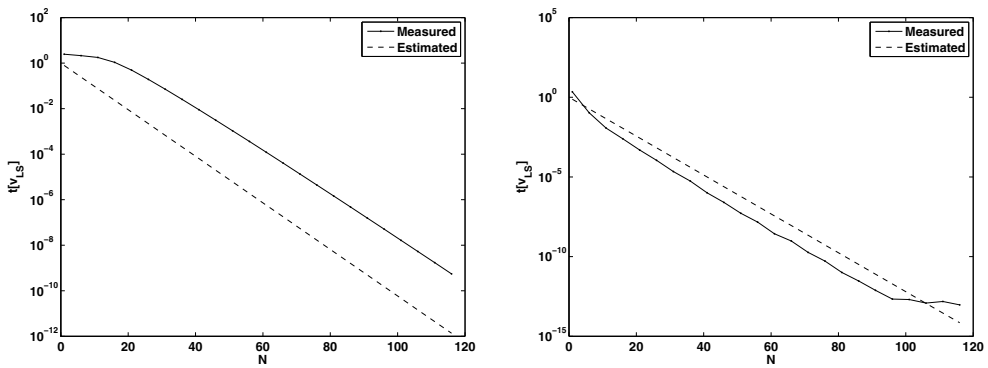


FIG. 6.1. *Left plot: Convergence of $t[v_{LS}]$ (solid-dotted line) for growing N for the case $r = 0.9$ compared to the estimated rate of convergence from Theorem 5.4 (dashed line). Right plot: Same as left plot but now with $r = 1.3$. The convergence is much faster for this choice of r.*

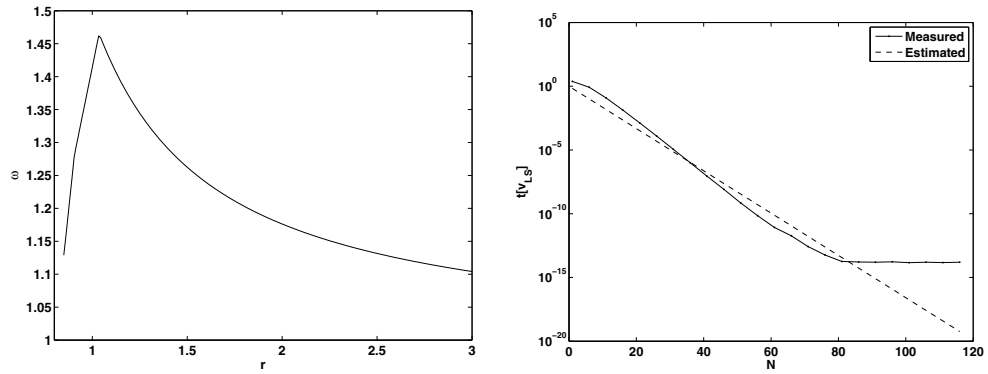


FIG. 6.2. *Left plot: Asymptotic convergence factor ω in dependence of r. Large values of ω indicate faster convergence of $t[v_{LS}]$. Right plot: Measured (solid-dotted line) and estimated (dashed line) convergence for the optimal radius $r = 1.036$.*

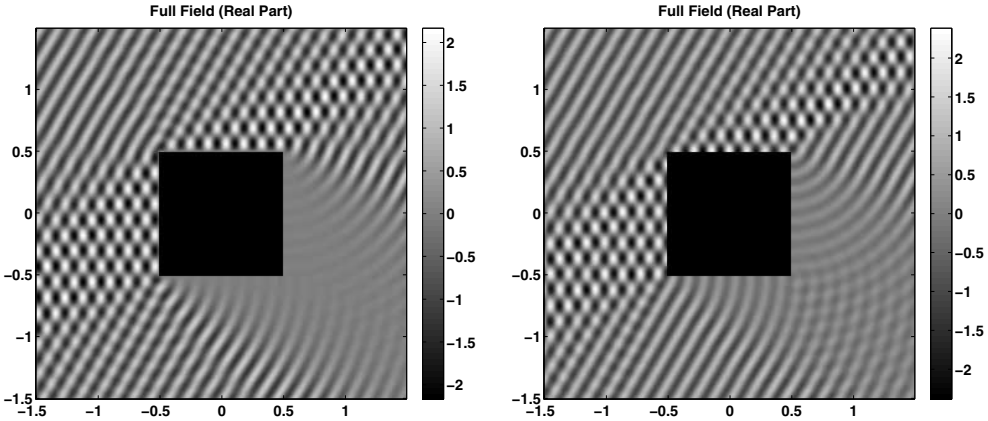


FIG. 6.3. *Real part of the full field for sound-soft scattering on the square (left plot) and sound-hard scattering on the square (right plot). In both cases $k = 50$ and the incident wave is a plane wave with angle $-\frac{\pi}{6}$.*

TABLE 6.1
Results for different wavenumbers for sound-soft scattering on the snowflake domain.

k	N	m	N_V	$t[v_{LS}]$	Time
50	70	1980	1071	$3 \cdot 10^{-8}$	8 s
100	90	2460	1377	$4 \cdot 10^{-9}$	15 s
200	130	3660	1989	$5 \cdot 10^{-9}$	44 s
500	260	8700	3978	$2 \cdot 10^{-7}$	7 m

In the bottom two plots of Figure 6.4 we show the scattered field (left) and the full field (right) for sound-soft scattering by an incident plane wave with angle $-\frac{\pi}{3}$ and wavenumber $k = 100$. Setting up and solving the dense least-squares problem takes 15 seconds. The error $t[v_{LS}]$ is approximately $4 \cdot 10^{-9}$.

In Table 6.1 we show results for different wavenumbers. N was chosen such that the error $t[v_{LS}]$ is in the approximate range 10^{-7} to 10^{-9}. The number of fundamental solutions used was $N_e = 3.3N$. m and N_V denote the dimensions of the matrix $A \in \mathbb{C}^{m \times N_V}$ in the least-squares problem $\min_c \|W(Ac - b)\|_2$. The table lists the total times for setting up the matrix and solving the least-squares problem. Up to around $k = 200$ the solution time is excellent. For higher k the $O(k^3)$ dependence of the solution time becomes a problem. Also at this high wavenumber, we have noticed that the best achievable $t[v_{LS}]$ seems to be around 10^{-7}. By contrast, Figure 6.5 shows that, at lower k, accuracies approaching machine precision are achievable. This figure also shows that a small increase in the number of basis functions leads to a large decrease of $t[v_{LS}]$, once the number of basis functions is high enough to get into the regime of exponential convergence. From our experiments the number of basis functions required to enter this regime grows approximately linearly with k, as supported by Figure 6.5. Once this regime is entered, the convergence rate appears to be independent of k, as suggested by the analysis in section 5.

In Remark 5.6 we stated the conjecture that $\|u - v\|^2_{L^2(E)} \leq CJ(v)$ for some $C > 0$. To support this conjecture we also plotted in Figure 6.5 the convergence of the scattered field u_s. This was done by first computing an approximate solution u_s for $N = 100$ and then approximating the L^∞ error $\|u_s^{(N)} - u_s\|_{L^\infty(\mathbb{R}^2 \setminus \Omega)}$, where $u_s^{(N)}$ is the approximate scattered field in step N, by evaluating it on the same grid that was used to plot the solution in Figure 6.4. Notice that the L^∞ errors are always at least a factor of 50 smaller than $t[v_{LS}]$. Therefore this plot supports the stronger conjecture

$$\|u - \hat{u}_{inc} - v\|^2_{L^\infty(\mathbb{R}^2 \setminus \Omega)} \leq CJ(v),$$

where $\hat{u}_{inc}$ is as defined in (2.5) and $C > 0$ is a constant, which is, however, expected to be large if exterior resonance effects occur.

6.3. A cavity domain. To conclude this section we demonstrate the example of scattering from a non-star-shaped cavity domain, which allows strong resonances to occur. The domain decomposition is shown in the left plot of Figure 6.6. Note that the leftmost element does not share a corner with the cavity domain; in this element we use a regular Fourier–Bessel expansion, that is, (2.1) with $\alpha_i = 1$ and both sin and cos terms, whose origin is chosen inside the element. The reason for using this new type of element is to reduce the size of the elements at the corners of the entrance to the cavity, which increases the relative distance to neighboring singularities and thereby increases their rate of exponential convergence. Even though an extra basis set is required, at high accuracy the total basis size is reduced, and hence efficiency is improved. The previous convergence analysis would easily extend to this case.

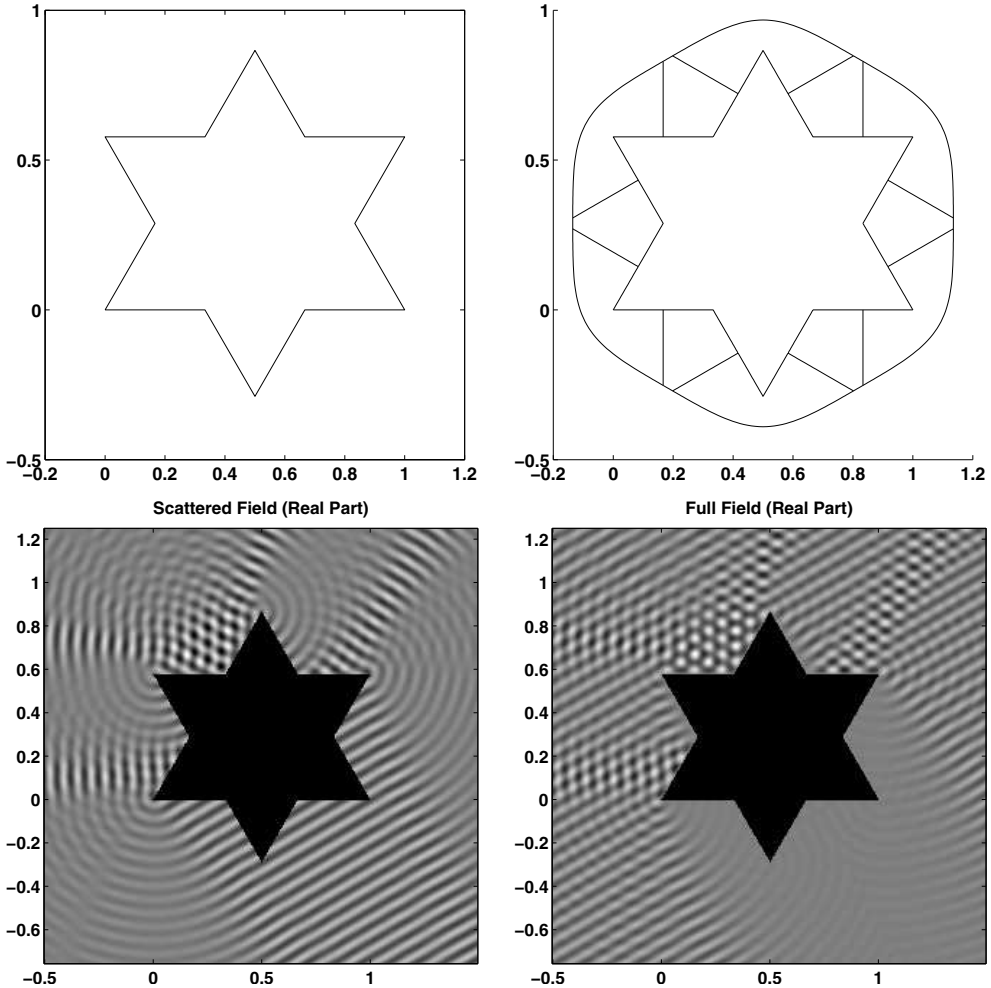


FIG. 6.4. *Top plots: A snowflake-shaped domain and the corresponding domain decomposition. Bottom plots: The scattered (left) and full (right) fields for sound-soft scattering on the snowflake domain with* $k = 100$.

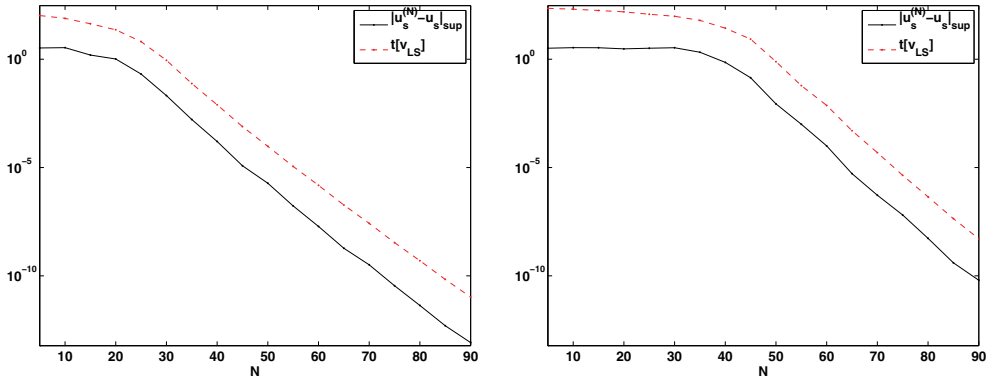


FIG. 6.5. *Convergence of* $t[v_{LS}]$ *(dashed line) for* $k = 50$ *(left plot) and* $k = 100$ *(right plot). The plots also show an approximation to the convergence versus* N *of the* L^∞ *norm of the approximate scattered field* $u_s^{(N)}$ *to the exact scattered field* u_s *(solid line).*

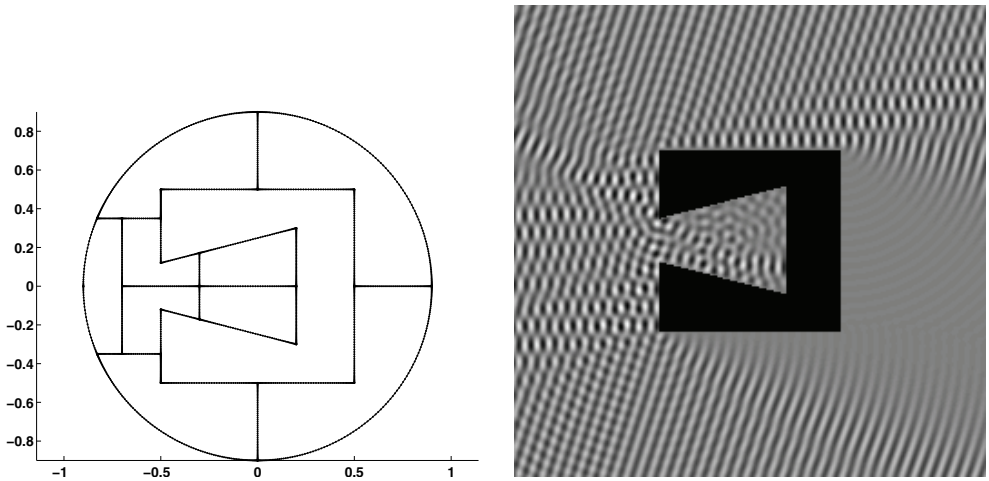


FIG. 6.6. *Sound-soft scattering from a cavity domain.*

In the right plot of Figure 6.6 we show the solution of sound-soft scattering by an incoming plane wave with direction $-\pi/10$ for $k = 100$, i.e., 16 wavelengths on the longest edge. The setup and solution of the associated least-squares problem takes 14 seconds (matrix size 1938×1521) for an error of $t[v_{LS}] \approx 3 \cdot 10^{-8}$. To evaluate and plot the solution on $6 \cdot 10^4$ points (a square grid of spacing 0.01) takes an additional 46 seconds.

7. Numerical stability of the method. The method discussed in this paper requires the solution of least-squares problems of the form (3.2), where $A \in \mathbb{C}^{m \times N_V}$ is highly ill-conditioned and W is a diagonal matrix containing quadrature weights. Sources of the ill-conditioning in A are the large dynamic range of the Fourier–Bessel basis functions and the ill-conditioning of fundamental solutions bases. The ill-conditioning of the Fourier–Bessel functions can be improved by rescaling these functions so that they have unit value at the subdomain's maximum radius from the corner, which is what we have done in all example calculations. By contrast, ill-conditioning of the fundamental solutions is inextricably linked to the fast exponential decay of the eigenvalues $\hat{s}_\eta(m)$ of the associated layer potential operator S_η, and thus cannot be removed by scaling.

Denote by c_{LS} the solution of the least-squares problem (3.2). If we have sufficiently many quadrature points, then $t[v_{LS}]$ is well approximated by the discrete error $t_d[v_{LS}] := \|W(Ac_{LS} - b)\|_2$. Solving (3.2) using a backward stable least-squares solver—such as `mldivide` in MATLAB—returns a solution $\tilde{c}_{LS}$, which is the exact minimizer of

$$\min_{c \in \mathbb{C}^{N_V}} \|(WA + E)c - (Wb + f)\|_2,$$

where E and f are small perturbations bounded by

$$\|E\|_2, \|f\|_2 \leq C\epsilon_{mach}, \tag{7.1}$$

where the constant $C > 0$ is small and the machine precision is ϵ_{mach} [2]. Because of ill-conditioning, $\tilde{c}_{LS}$ may differ drastically from c_{LS}. However, the following shows

that the corresponding discrete least-squares functional can deviate from its true minimum only under certain conditions.

LEMMA 7.1. *Let $\tilde{c}_{LS}$ be defined as above, and let $\tilde{v}_{LS}$ be the corresponding solution field. Then,*

$$t_d[v_{LS}] \;\le\; t_d[\tilde{v}_{LS}] \;\le\; t_d[v_{LS}] + C(2 + \|c_{LS}\|_2 + \|\tilde{c}_{LS}\|_2)\epsilon_{mach}. \tag{7.2}$$

Proof. Using the property that c_{LS} minimizes the unperturbed least-squares problem, and also using the triangle inequality, results in

$$\begin{aligned}\|WAc_{LS} - Wb\|_2 &\le \|WA\tilde{c}_{LS} - Wb\|_2 \\ &\le \|(WA+E)\tilde{c}_{LS} - (Wb+f)\|_2 + C(1 + \|\tilde{c}_{LS}\|_2)\epsilon_{mach}.\end{aligned}$$

Exchanging perturbed and unperturbed quantities in the above gives similarly

$$\|(WA+E)\tilde{c}_{LS} - (Wb+f)\|_2 \le \|WAc_{LS} - Wb\|_2 + C(1 + \|c_{LS}\|_2)\epsilon_{mach}.$$

Combining the two estimates gives (7.2). ∎

Thus, if the basis coefficient 2-norms at the approximate and exact minima are small, the numerical least-squares solution must converge at the same exponential rate (Theorem 5.4) as the exact least-squares solution. To rigorously apply this lemma we need bounds on the norms of $\|c_{LS}\|_2$ and $\|\tilde{c}_{LS}\|_2$. A partial result towards deriving such bounds is given in the following theorem, which establishes under which conditions on Γ_F the coefficients of the fundamental solutions basis functions in Ω_e are bounded independently of N. A similar result for approximating interior Helmholtz problems with fundamental solutions was given in [4].

THEOREM 7.2. *Consider a sequence of fundamental solutions approximations, each of the form $v(\boldsymbol{x}) = \sum_{j=1}^{N_e} c_j^{(e)}\big(\frac{i}{4}\frac{\partial}{\partial\nu(\boldsymbol{y}_j)}H_0^{(1)}(k|\boldsymbol{x}-\boldsymbol{y}_j|) - \frac{\eta}{4}H_0^{(1)}(k|\boldsymbol{x}-\boldsymbol{y}_j|)\big)$, with growing numbers N_e (N_e even) of charge points, that attains the error bound from Theorem 5.2 as $N_e \to \infty$. Let each coefficient vector be written as $c^{(e)} := \big[c_1^{(e)}, \ldots, c_{N_e}^{(e)}\big]^T$. If $R > \max_i |p_i|$, then the sequence of norms $\|c^{(e)}\|_2$ is bounded independently of N_e.*

Proof. In the following, $C > 0$ denotes an unspecified constant that depends on k, R, r, and ϵ but not on N_e, and may change throughout the proof. By assumption $\max_i |p_i| < R < r$.

Let $g(\theta) = \sum_{j=1}^{N_e} c_j^{(e)}\delta(\phi - \phi_j)$ and let $\hat{g}(m) = \frac{N_e}{2\pi}\hat{c}^{(i)}_{(m \mod N_e)}$ be the Fourier coefficients of g (see section 5.1). The proof follows by showing that $\hat{g}(m)$ is bounded for m in the interval $[-N_e/2+1, \ldots, N_e/2]$ and by applying the discrete version of Parseval's identity. For the specific choice of the coefficients $c^{(e)}$, such that $\hat{g}(m) = \frac{\hat{u}_s(m)}{\hat{s}_\eta(m)}$, which was used for the convergence analysis in section 5.1, boundedness already follows for sufficiently small ϵ from

$$|\hat{g}(m)| \le C\left(\frac{r}{R\rho_\epsilon}\right)^{|m|}$$

and the assumption $R > \max_i |p_i|$.

To establish boundedness of $|\hat{g}(m)|$ for any sequence $c^{(e)}$ that attains the optimal rate of convergence from Theorem 5.2 we proceed as follows. From (5.2) and (5.5) we have that

$$k\sqrt{2\pi r}|\hat{u}_s(m) - \hat{s}_\eta(m)\hat{g}(m)| \le C\tau^{-N_e}, \quad m \in \mathbb{N},$$

where $\tau = \min\{\left(\frac{r}{R}\right), \rho_\epsilon^{1/2}\}$ (note that by Remark 5.3, $\tilde{\epsilon} = 0$ in Theorem 5.2 is possible since we use only the estimate for the function and not for the normal derivative). It follows that

$$|\hat{g}(m)| \leq |\hat{s}_\eta(m)|^{-1} \left(C\tau^{-N_e}(2\pi r)^{-1/2} + |\hat{u}_s(m)| \right), \quad m \in \mathbb{N}.$$

Now restrict m to the interval $[-N_e/2+1, \ldots, N_e/2]$. If $R > \sqrt{r \max_i |p_i|}$, then for sufficiently small ϵ it follows that $\tau = \frac{r}{R}$. Using (4.4) and (5.4) we have

$$|\hat{g}(m)| \leq C \left(\frac{r}{R}\right)^{|m|} \left[\left(\frac{r}{R}\right)^{-N_e} + \rho_\epsilon^{-|m|} \right] \leq C \left(\frac{r}{R}\right)^{|m|} \left[\left(\frac{r}{R}\right)^{-|m|} + \rho_\epsilon^{-|m|} \right] \tag{7.3}$$

since $r > R$ and $|m| \leq \frac{N_e}{2}$. From $\max_i |p_i| < R$ it follows for sufficiently small ϵ that $\rho_\epsilon^{-|m|} \leq \left(\frac{r}{R}\right)^{-|m|}$. Inserting this into (7.3) gives

$$|\hat{g}(m)| \leq C,$$

where all constants are absorbed into C. If $R < \sqrt{r \max_i |p_i|}$, then for sufficiently small ϵ it holds that $\tau = \rho_\epsilon$. We obtain, absorbing constants into C,

$$\begin{aligned} |\hat{g}(m)| &\leq C \left(\frac{r}{R}\right)^{|m|} \left[\rho_\epsilon^{-N_e/2} + \rho_\epsilon^{-|m|} \right] \\ &\leq C \left(\frac{r}{R}\right)^{|m|} \left[\rho_\epsilon^{-|m|} + \rho_\epsilon^{-|m|} \right] \leq C \left(\frac{r}{R\rho_\epsilon}\right)^{|m|} \leq C \end{aligned}$$

for ϵ sufficiently small, since $\max_i |p_i| < R$, and therefore $\rho_\epsilon > \frac{r}{R}$.

From (4.8) it follows that $\hat{g}(m) = \frac{N_e}{2\pi}\hat{c}^{(e)}(m)$ for $m \in [-N_e/2+1, \ldots, N_e/2]$, where the vector $\hat{c}^{(e)} \in \mathbb{C}^{N_e}$ is the discrete Fourier transform of $c^{(e)}$. Hence, using the boundedness of $\hat{g}(m)$ for $m \in [-N_e/2+1, \ldots, N_e/2]$ together with the discrete version of Parseval's identity $\|c^{(e)}\|_2^2 = N_e \|\hat{c}^{(e)}\|_2^2$, we obtain

$$\|c^{(e)}\|_2^2 = N_e \|\hat{c}^{(e)}\|_2^2 = \frac{(2\pi)^2}{N_e} \sum_{j=-\frac{N_e}{2}+1}^{\frac{N_e}{2}} |\hat{g}(m)|^2 \leq (2\pi)^2 C^2,$$

which concludes the proof. ∎

Hence, the coefficient vector of the MFS approximations stays bounded in norm as long as $\max_i |p_i| < R$. In [4] this was studied numerically for interior Helmholtz problems, in the general case of analytic MFS and boundary curves. In that work we also showed that a large coefficient norm forces a corresponding loss of accuracy in solution evaluation due to round-off error.

Consider again the problem of sound-soft scattering on the square (section 6.1). We have $|p_i| = \frac{\sqrt{2}}{2}$ for all corners of the square $\Omega = (-1/2, 1/2)^2$. If the radius of the source points $\mathbf{y}_j$ is $R = 0.4$, then for $k = 100$ and $N = 150$ we find that the computed coefficient vector $\tilde{c}_{LS}$ has $\|\tilde{c}_{LS}\| \approx 3 \cdot 10^9$. The measured boundary error $t_d[\tilde{v}_{LS}] \approx 3.6$ is terrible. On the other hand, if $R = 0.8 > \max_i |p_i|$, then $\|\tilde{c}_{LS}\| \approx 6.2$, and $t_d[\tilde{v}_{LS}] \approx 7 \cdot 10^{-10}$ is excellent. Note that in both cases the matrix WA is numerically singular with a condition number of around 10^{20}.

To summarize this section, the crucial factor limiting the accuracy of the method is not the condition number of WA, but the norm of the basis coefficients vector. This can be kept small by rescaling Fourier–Bessel functions and by an informed choice of the fundamental solutions curve.

8. Conclusions. In this paper we have demonstrated that efficient and highly accurate computation of scattering from a variety of convex and nonconvex polygons can be achieved by using a small number of elements, one for each corner, with the right kind of basis functions on each element. This allows one to tackle problems tens of wavelengths across quite rapidly, even with dense linear algebra. An advantage of the latter is that a dense factorization would allow multiple right-hand sides (e.g., incident wave directions) to be solved at minimal extra cost. Once the coefficients have been solved for, the user may choose where to evaluate the solution at a constant cost per point—this may be slow on large grids, but is trivially parallelizable.

For high-frequency sound-soft scattering of plane waves on convex polygons, the method analyzed in this paper cannot compete with specialized high-frequency boundary element methods (see, e.g., [10]). However, the approach in this paper is much more general. It allows for the computation of convex and nonconvex polygons, sound-soft or sound-hard boundary conditions, and non–plane wave incident fields. Due to its fast exponential convergence, up to medium frequencies (≤ 50 wavelengths in size), it is very competitive.

Our implementation is also much simpler than boundary element or boundary integral methods. At high frequency these usually need a careful treatment of oscillatory integrals with complicated basis functions and weakly oscillatory kernels, or spectral quadrature schemes for (hyper)singular kernels. In addition they require careful mesh-refinement or quadrature schemes for corner singularities. None of this complexity is needed in our approach. Moreover, with `MPSpack` we have created a simple-to-use software tool that allows rapid and flexible implementation of our methods, and more (see Appendix A).

The approach in this paper can easily be generalized to multiple scattering problems. For example, Figure 8.1 shows the solution for sound-soft scattering with incident angle $-\frac{\pi}{3}$ from a disk near a snowflake domain, at wavenumber $k = 100$. Setting up and solving the matrix problem takes around 7.5 minutes with an error of $t[v_{LS}] \approx 4 \cdot 10^{-7}$. Here we used an additional basis of fundamental solutions with charge points lying on a circle inside the disk domain.

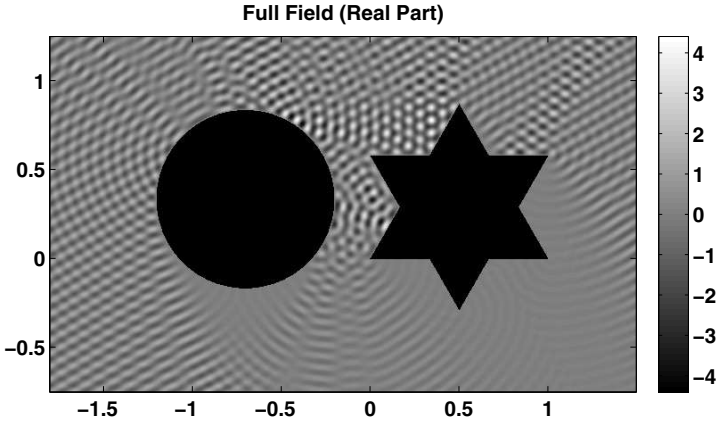


FIG. 8.1. *Multiple scattering from a disk and a snowflake domain.*

The exponential convergence analyzed in this paper relies on asymptotic series for Helmholtz solutions at corners, and their convergence properties, which are known analytically for sound-soft and sound-hard boundary conditions on straight bound-

aries. For impedance, transmission, and curved-wall corner problems, the solution is not separable close to a corner, and an open problem is what kind of basis functions to choose here to obtain fast exponential convergence.

Another open problem is optimal mesh (subdomain) generation. The mesh should be chosen in such a way that fast exponential convergence on each element is guaranteed by keeping the relative distance of mesh elements to neighboring singularities large. But at the same time the mesh should not contain too many elements, so that the overall numbers of basis functions and matching points remain small.

Finally, the efficient solution of the least-squares linear algebra problem needs to be further investigated. At the moment we use a dense, backward stable solver, which works well for medium wavenumbers. However, it would be desirable to have iterative solvers for large, structured, ill-conditioned least-squares problems in order to remove the $O(k^3)$ bottleneck at high wavenumbers.

Appendix A. `MPSpack` example: Scattering from a square. `MPSpack` is an object-oriented MATLAB toolbox developed by the authors to solve two-dimensional Helmholtz/Laplace PDE problems with particular and fundamental solutions methods. The code can be obtained from http://code.google.com/p/mpspack. The download section of that Web page also includes an extensive tutorial and a technical manual.

```
% Exponentially accurate sound-soft time-harmonic scattering from the square
k = 50;                  % Wavenumber
r = 1.0;                 % Radius of outer artificial circle
M = 100;                 % Number of quadrature points on each segment
N = 90;                  % Number of basis funcs in each corner subdomain
a = 0.5;                 % Half-size of the square
R = sqrt(0.5);           % Radius of the fundamental solutions curve

% Define segments...
s = segment.polyseglist(M, [1i*r 1i*a a+1i*a a r]);         % straight
s = [s(1:3) segment(3*M, [0 r 0 pi/2])];                    % add arc
s = [s rotate(s, pi/2) rotate(s, pi) rotate(s, 3*pi/2)];    % add 3 copies
sart = s([1 4 5 8 9 12 13 16]);        % list of all artificial boundaries
sext = s([4 8 12 16]);                 % segments forming outer circle

% Define domains...
for j=1:4, d(j) = domain(s(1+mod(4*(j-1)+[0 1 2 12 3],16)),[1 1 1 -1 1]); end
ext = domain([], [], sext(end:-1:1), -1);
sart.setmatch([k -k], [1 -1]);         % matching conditions between elements

% Basis functions...
nuopts = struct('type','s', 'cornermultipliers',[0 0 1 0 0], 'rescale_rad',1);
for j=1:4, d(j).addcornerbases(N,nuopts); end               % frac-order FB
Z = @(t) R*exp(2i*pi*t); Zp = @(t) 2i*pi*R*exp(2i*pi*t);    % fund soln curve
ext.addmfsbasis({Z, Zp}, N, struct('eta',k, 'fast',2, 'nmultiplier',2.1));

p = scattering(ext, d);                % now set up problem, solve, and plot...
p.setoverallwavenumber(k);
p.setincidentwave(-pi/6);
p.solvecoeffs;
fprintf('least-square err = %g, coeff norm = %g\n', p.bcresidualnorm, norm(p.co))
p.showfullfield(struct('bb',[-1.5 1.5 -1.5 1.5], 'dx',0.02));
```

FIG. A.1. *MATLAB code for sound-soft scattering from a square using* `MPSpack` *toolbox.*

The short, human-readable code in Figure A.1 implements scattering from the square as in section 6.1, giving least-squares error 8×10^{-11} in 6 seconds of CPU time, and a further 11 seconds to plot the solution. The code is explained in more detail in the tutorial. Here we will mention only that segments are first created as in Figure 2.1, and then subdomains E_1 through E_4 and Ω_e^+ are created. Fractional-order Bessel function bases are added to subdomains with the command `addcornerbases`, and fundamental solutions are added with `addmfsbasis`. Finally, a `scattering` object is created which contains methods to construct the matrix WA from section 3, solve the least-squares linear system, and plot the solution.

Acknowledgments. Timo Betcke wishes to thank Lehel Banjai (MPI Leipzig), Mahadevan Ganesh (Colorado School of Mines), and Ralf Hiptmair (ETH Zürich) for helpful discussions. Alex Barnett thanks Leslie Greengard (Courant) for helpful discussions. Both authors wish to thank the anonymous referees for helpful comments that improved the presentation of the paper.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 10th ed., Dover, New York, 1964.

[2] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[3] I. Babuška and J. M. Melenk, *The partition of unity method*, Internat. J. Numer. Methods Engrg., 40 (1997), pp. 727–758.

[4] A. H. Barnett and T. Betcke, *Stability and convergence of the method of fundamental solutions for Helmholtz problems on analytic domains*, J. Comput. Phys., 227 (2008), pp. 7003–7026.

[5] T. Betcke, *A GSVD formulation of a domain decomposition method for planar eigenvalue problems*, IMA J. Numer. Anal., 27 (2007), pp. 451–478.

[6] A. Bogomolny, *Fundamental solutions method for elliptic boundary value problems*, SIAM J. Numer. Anal., 22 (1985), pp. 644–669.

[7] O. P. Bruno, C. A. Geuzaine, J. J. A. Monro, and F. Reitich, *Prescribed error tolerances within fixed computational times for scattering problems of arbitrarily high frequency: The convex case*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 362 (2004), pp. 629–645.

[8] O. Cessenat and B. Despres, *Application of an ultra weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problem*, SIAM J. Numer. Anal., 35 (1998), pp. 255–299.

[9] S. N. Chandler-Wilde and I. Graham, *Boundary integral methods in high frequency scattering*, in Highly Oscillatory Problems: Computation, Theory and Applications, B. Engquist, T. Fokas, E. Hairer, and A. Iserles, eds., Cambridge University Press, Cambridge, UK, 2009.

[10] S. N. Chandler-Wilde and S. Langdon, *A Galerkin boundary element method for high frequency scattering by convex polygons*, SIAM J. Numer. Anal., 45 (2007), pp. 610–640.

[11] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Appl. Math. Sci. 93, Springer-Verlag, Berlin, 1998.

[12] J. Descloux and M. Tolley, *An accurate algorithm for computing the eigenvalues of a polygonal membrane*, Comput. Methods Appl. Mech. Engrg., 39 (1983), pp. 37–53.

[13] V. Domínguez, I. G. Graham, and V. P. Smyshlyaev, *A hybrid numerical-asymptotic boundary integral method for high-frequency acoustic scattering*, Numer. Math., 106 (2007), pp. 471–510.

[14] S. C. Eisenstat, *On the rate of convergence of the Bergman–Vekua method for the numerical solution of elliptic boundary value problems*, SIAM J. Numer. Anal., 11 (1974), pp. 654–680.

[15] G. Fairweather and A. Karageorghis, *The method of fundamental solutions for elliptic boundary value problems*, Adv. Comput. Math., 9 (1998), pp. 69–95.

[16] G. Fairweather, A. Karageorghis, and P. A. Martin, *The method of fundamental solutions for scattering and radiation problems*, Eng. Anal. Bound. Elem., 27 (2003), pp. 759–769.

[17] P. Gamallo and R. J. Astley, *A comparison of two Trefftz-type methods: The ultraweak variational formulation and the least-squares method, for solving shortwave* 2*-d Helmholtz problems*, Internat. J. Numer. Methods Engrg., 71 (2007), pp. 406–432.

[18] C. Geuzaine, O. Bruno, and F. Reitich, *On the O(*1*) solution of multiple-scattering problems*, IEEE Trans. Magnetics, 41 (2005), pp. 1488–1491.

[19] P. Henrici, *A survey of I. N. Vekua's theory of elliptic partial differential equations with analytic coefficients*, Z. Angew. Math. Phys., 8 (1957), pp. 169–203.

[20] P. Henrici, *Applied and Computational Complex Analysis. Vol.* 3, Pure Appl. Math., John Wiley & Sons Inc., New York, 1986.

[21] R. Hiptmair, A. Moiola, and I. Perugia, *Plane Wave Discontinuous Galerkin Methods for the* 2*D Helmholtz Equation: Analysis of the p-Version*, Tech. Report 2009-20, SAM, ETH Zürich, Switzerland, 2009.

[22] U. Kangro, *Convergence of collocation method with delta functions for integral equations of first kind*, Integral Equations Operator Theory, 66 (2010), pp. 265–282.

[23] J. R. Kuttler and V. G. Sigillito, *Eigenvalues of the Laplacian in two dimensions*, SIAM Rev., 26 (1984), pp. 163–193.

[24] P. Monk and D.-Q. Wang, *A least-squares method for the Helmholtz equation*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 121–136.

[25] F. W. J. Olver, *Asymptotics and Special Functions*, AKP Classics, A K Peters, Ltd., Wellesley, MA, 1997. Reprint of the 1974 original [Academic Press, New York].

[26] E. Perrey-Debain, *Plane wave decomposition in the unit disc: Convergence estimates and computational aspects*, J. Comput. Appl. Math., 193 (2006), pp. 140–156.

[27] M. Stojek, *Least-squares Trefftz-type elements for the Helmholtz equation*, Internat. J. Numer. Methods Engrg., 41 (1998), pp. 831–849.

[28] I. N. Vekua, *New Methods for Solving Elliptic Equations*, OGIZ, Moscow, Leningrad, 1948 (in Russian); North–Holland, Amsterdam, 1967 (in English).

SIAM J. NUMER. ANAL.
Vol. 47, No. 3, pp. 1952–1970
© 2009 Society for Industrial and Applied Mathematics

PERTURBATIVE ANALYSIS OF THE METHOD OF PARTICULAR SOLUTIONS FOR IMPROVED INCLUSION OF HIGH-LYING DIRICHLET EIGENVALUES*

A. H. BARNETT†

Abstract. The Dirichlet eigenvalue or "drum" problem in a domain $\Omega \subset \mathbb{R}^2$ becomes numerically challenging at high eigenvalue (frequency) E. In this regime the method of particular solutions (MPS) gives spectral accuracy for many domain shapes. It requires a number of degrees of freedom scaling as $\sqrt{E}$, the number of wavelengths on the boundary, in contrast to direct discretization for which this scaling is E. Our main result is an inclusion bound on eigenvalues that is a factor $O(\sqrt{E})$ tighter than the classical bound of Moler–Payne and that is optimal in that it reflects the true slopes of curves appearing in the MPS. We also present an MPS variant that cures a normalization problem in the original method, while evaluating basis functions only on the boundary. This method is efficient at high frequencies, where we show that, in practice, our inclusion bound can give three extra digits of eigenvalue accuracy with no extra effort.

Key words. particular solutions, eigenvalue, Dirichlet, drum, Laplacian, inclusion

AMS subject classifications. 35P15, 47A55, 65N35, 65N15

DOI. 10.1137/080724022

1. Introduction and main results.

Computing the eigenvalues and eigenmodes of the Laplace operator with homogeneous boundary condition (BC) in a bounded domain $\Omega \subset \mathbb{R}^d$ for $d = 2, 3, \ldots$ is a classical problem with a 150-year-long history; for reviews, see [30, 43]. It has a wealth of applications in physics and engineering including the vibration of membranes (drums), resonances and modes of acoustic, electromagnetic, and optical cavities and waveguides, energy states of quantum particles, and data analysis [40]. We focus on the case of a Dirichlet BC, i.e., acoustically sound-soft walls. The spectrum $\sigma_D := \{E_j\}_{j=1}^{\infty}$ and modes $\{\phi_j\}_{j=1}^{\infty}$ are defined by

$$-\Delta\phi_j = E_j\phi_j \quad \text{in } \Omega, \tag{1.1}$$

$$\phi_j = 0 \quad \text{on } \partial\Omega\ . \tag{1.2}$$

Physically the eigenvalues can be interpreted as frequencies; we order them (including multiplicities) $E_1 < E_2 \leq E_3 \leq \cdots$. We may choose orthonormal modes, that is, $(\phi_i, \phi_j)_\Omega := \int_\Omega \overline{\phi_i}\phi_j \, d\mathbf{x} = \delta_{ij}$ for all $i, j = 1, 2, \ldots$, where δ is the Kronecker delta.

There has been a growing interest in the computationally demanding case of high eigenvalue (high mode number j), where many wavelengths span the domain. One example is dielectric microcavity lasers, where the use of asymmetric shapes has led to much higher output powers; to predict lasing [45], a large number of cavity resonances (which can be approximated by Dirichlet modes) are needed at $j > 10^3$. More broadly, the field of "quantum chaos" (short-wavelength asymptotic study of modes of systems with chaotic ray dynamics) has blossomed in the last 20 years in both physics [22] (with applications such as atomic physics and quantum dots) and mathematics (e.g.,

*Received by the editors May 13, 2008; accepted for publication (in revised form) January 5, 2009; published electronically May 27, 2009. This work was supported in part by a Courant Instructorship at NYU and by NSF grant DMS-0507614.
http://www.siam.org/journals/sinum/47-3/72402.html
†Department of Mathematics, Dartmouth College, Hanover, NH 03755 (ahb@math.dartmouth.edu).

quantum ergodicity [47] and arithmetic manifolds). Numerical investigation has been a vital part of this endeavor [1, 4] and has led to discoveries such as "scars" of periodic ray orbits in the modes [25].

The method of particular solutions (MPS), also known as collocation or "point matching," uses basis functions that satisfy the Helmholtz equation (1.1) but not necessarily (1.2). It was originally used by Fox, Henrici, and Moler [16] to compute low eigenvalues of an L-shaped domain to 8 digit accuracy, and has been improved by many, including Kuttler and Sigillito [29], and recently by Betcke and Trefethen [5]. Much of the key early analysis of the MPS appeared in SINUM [16, 33, 41, 10]. For geometries with many corners, a domain decomposition version was pioneered by Descloux and Tolley [8], modified by Driscoll [9] to find eigenvalues of the well-known pair of Gordon–Wolpert–Webb isospectral drums [21] to high accuracy, and further improved by Betcke [6]. In physics, where the emphasis is on high frequencies, related methods of Heller [25, 26] and the very efficient "scaling method" of Vergini and Saraceno [46, 3, 4] have enjoyed unrivaled success. In this regime the advantage of the MPS is that the basis size scales (in $d = 2$) as the wavenumber $\sqrt{E}$ (e.g., see [2]), rather than E as with finite difference or finite element methods. (This advantage is shared by boundary integral methods [1]).

In the MPS a solution to (1.1) is approximated by a linear combination of basis functions $\{\xi_n\}_{n=1,\ldots,N}$, each of which satisfies $-\Delta\xi_n = E\xi_n$ in Ω at some trial frequency parameter E but which do not individually satisfy the Dirichlet BC. Popular basis functions are plane waves [26], corner-adapted Fourier–Bessel functions [16, 9, 5, 6], and fundamental solutions [12, 27, 4]. If a trial solution $u \in \mathrm{Span}\{\xi_n\}$ satisfies $u = 0$ on $\partial\Omega$ while $u \neq 0$ in Ω, then it must be a multiple of a mode ϕ_j for some j, and we must have $E = E_j$. In practice, this is rarely achieved exactly. However, by defining the boundary error measure (or "tension")

$$t = t[u] := \frac{\|u\|_{\partial\Omega}}{\|u\|_{\Omega}} , \tag{1.3}$$

where $\|u\|_{\partial\Omega}$ and $\|u\|_{\Omega}$ are the usual L^2-norms on the boundary and interior, respectively, we have the following inclusion bound from a theorem of Moler–Payne [33, 29]. Let $-\Delta u = Eu$ in Ω, then there exists an eigenvalue E_j satisfying

$$\frac{|E - E_j|}{E_j} \;\leq\; C'_\Omega t[u] , \tag{1.4}$$

where C'_Ω is a constant depending only on the domain. (C'_Ω is an a priori upper bound on $\|h\|_\Omega/\|h\|_{\partial\Omega}$ for harmonic functions h [29], hence (1.4) is known as an a priori-a posteriori inequality [30].) Eigenvalues may then be identified by searching the frequency (E) axis for local minima of the function

$$t_m(E) := \min_{u\in\mathrm{Span}\{\xi_n\}} t[u] , \tag{1.5}$$

as shown in Figure 1.1. Similar curves appear in [5, 7, 11]. Equation (1.4) states that the eigenvalue accuracy which may be claimed is controlled by the size of the minimum found.

However, observations of the slopes of curves such as Figure 1.1 led the author to notice that the Moler–Payne bounds are not optimal at high E. It turns out that, for E in the neighborhood of each eigenvalue E_j, the bounds can be made a factor of $\sqrt{E}$ tighter, as follows.

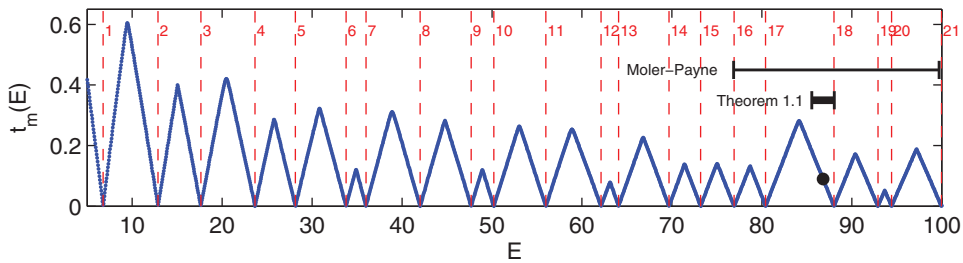


FIG. 1.1. *Graph of $t_m(E)$ versus E for the domain (S) shown and defined in Figure* 1.2 *and weight $w = 1$. Dirichlet eigenvalues E_j are shown by vertical dotted lines labelled by j. At the single frequency $E = 86.8$ (black dot), the inclusion bounds given by* (1.4) *(thin solid interval) and Theorem* 1.1 *with the choice $\epsilon = 0.05$ (thick solid interval) are shown. The latter bounds are $\sqrt{E} \approx 10$ times tighter and optimal in the sense that they just include E_{18} and correspond to the true slope of the $t_m(E)$ curve.*

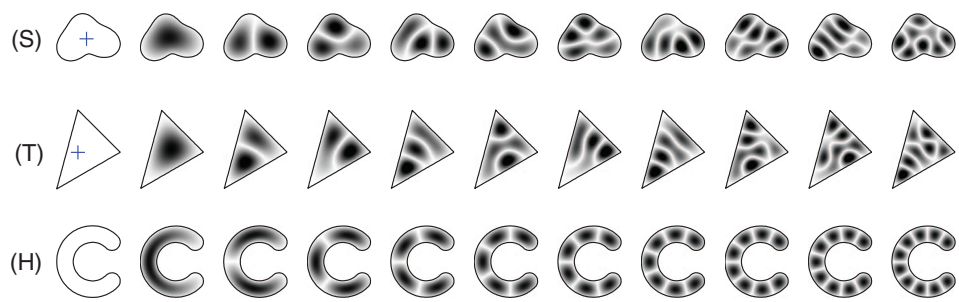


FIG. 1.2. *Three domains used in verifying the utility of the main theorem and (right) their first ten eigenmodes (white is zero and darker represents larger absolute values $|\phi_j(\mathbf{x})|$). (S) smooth (in fact analytic) and strictly star-shaped, defined by the polar function $r(\theta) = 1 + 0.2\cos 2\theta + 0.2\cos(3\theta + \pi/3)$. (T) triangle, with angles $\pi/4$, $\pi/3$, and $5\pi/12$. (H) horseshoe, not star-shaped about any point, C^1-continuous boundary, the union of a $3\pi/2$ sector of the annulus of inner and outer radii* 0.7 *and* 1.3 *with two semicircles. In (S) and (T) the origin is shown by a + sign.*

THEOREM 1.1. *Let $\Omega \subset \mathbb{R}^d$, $d \geq 2$, be a smooth bounded domain. Fix $\epsilon > 0$, then there is an open, possibly disconnected, set $\Sigma_\epsilon \subset (0, \infty)$ which contains the spectrum σ_D and has the following property. For any $E \in \Sigma_\epsilon \backslash \sigma_D$ and any u obeying $-\Delta u = Eu$ in Ω, there exists an eigenvalue E_j satisfying*

$$\frac{|E - E_j|}{\sqrt{E_j}} \;\leq\; C_\Omega(1+\epsilon)t[u] \;, \tag{1.6}$$

where C_Ω is a constant depending on the domain but not on ϵ, Σ_ϵ, E, or u.

Comparing (1.6) to (1.4), we see that for high-frequency applications, the MPS is, in fact, more accurate, by a factor of the wavenumber, than has been realized before. Although the theorem does not indicate how large the set Σ_ϵ is (and therefore, how useful the bound is), we will show that in practice the set fills almost all of $(0, \infty)$ even for ϵ much smaller than 1. We explore this in section 4.2 for three example planar ($d = 2$) domains, and we demonstrate in section 4.3 that at $E \approx 10^6$ the bound on eigenvalue accuracy is improved by 3 digits without extra effort.

Remark 1.1. The error measure (1.3), a Rayleigh quotient, was used by the author in [3, Chapter 5] (as a special case, it is implicit in [29]) and cures a normalization problem inherent in the original MPS formulation [16]. This problem, and a very

similar but independently found cure, is presented in [5]. Recently, Betcke [7] has united the various approaches within a single framework.

One of the main goals of this paper is to prove Theorem 1.1, which we do in sections 2 and 3. This will rely on the following: Lemma 2.1, which bounds $t[u]$ by a frequency (parameter) dependent eigenvalue of an operator $A(E)$ defined by (2.3) acting in $L^2(\partial\Omega)$; Theorem 2.2, which describes the quadratic parameter-dependence of such eigenvalues in the neighborhood of each E_j; and Lemma 2.3, an upper bound on the norm of eigenmode normal derivatives on the boundary. A deeper purpose of our work is to understand the structure of MPS curves such as Figures 1.1 and 5.1. We believe this can lead to better numerical methods and are particularly motivated by the variants of such curves which form the heart of the scaling method [46, 3, 4]. Section 3, where we apply perturbation theory in the frequency parameter to prove Theorem 2.2, is a step in this direction. We actually do this for a more general boundary norm

$$\|u\|_{w,\partial\Omega} := \left(\int_{\partial\Omega} w\,|u|^2 ds\right)^{1/2} , \tag{1.7}$$

where w is some strictly positive bounded weight function on $\partial\Omega$ and ds is the usual surface element. One point of introducing w is that it allows the following useful special case, where $\mathbf{x}(s) \in \partial\Omega$ is position vector and $\mathbf{n}(s)$ the outwards unit normal.

PROPOSITION 1.2. *Let Ω be strictly star-shaped about the origin, and let $w = (\mathbf{x}\cdot\mathbf{n})^{-1}$. If we replace the norm $\|u\|_{\partial\Omega}$ by $\|u\|_{w,\partial\Omega}$ in* (1.3)*, then in Theorem* 1.1 *we may choose the constant $C_\Omega = \sqrt{2}$, independent of Ω or d.*

Note that this geometric restriction on Ω is needed if w is to be positive and bounded. We demonstrate the utility of this explicit bound in section 4.

A secondary goal is to describe (in section 4.1) an implementation of the MPS, in particular, how to compute $t_m(E)$ using a generalized matrix eigenvalue problem and basis evaluations on the boundary alone. The latter makes our method more efficient than existing MPS methods [29, 5] at high frequency. Finally, we conclude in section 5.

2. Parameter-dependent boundary operator and proof of the theorem. Consider the interior Dirichlet boundary-value problem (BVP) for the homogeneous Helmholtz equation at frequency E,

$$\begin{aligned} (\Delta + E)u &= 0 \quad \text{in } \Omega\ , \\ u &= U \quad \text{on } \partial\Omega\ . \end{aligned} \tag{2.1}$$

We use small letters to indicate functions on Ω and the corresponding capital letter for their restriction to $\partial\Omega$ (boundary trace). Given $E \notin \sigma_D$, the BVP (2.1) with boundary data $U \in L^2(\partial\Omega)$ has a unique solution in $L^2(\Omega)$ [32, Theorem 4.25]. Henceforth, we work in a w-weighted inner product

$$\langle U, V\rangle := \int_{\partial\Omega} \overline{U} V\, w\, ds\ , \tag{2.2}$$

where $0 < c \le w(s) \le C$, for all $s \in \partial\Omega$ for some constants c, C. Thus (1.7) may be written $\sqrt{\langle U, U\rangle}$. The point of this weight is to allow choices such as that in Proposition 1.2, which may be more useful than the case $w = 1$, which is common in the literature.

The parameter-dependent operator $A = A(E)$ is defined by

$$\langle U, A(E)V\rangle \;=\; (u,v)_\Omega \;, \qquad \text{for all } U,V \in L^2(\partial\Omega). \tag{2.3}$$

(By v we mean the unique interior solution to (2.1) with data V). A unique such operator in $L^2(\partial\Omega)$ exists by the Riesz representation theorem. By (2.3), it is positive and self-adjoint (in the w-weighted norm) for real E. Treating (1.4) as an upper bound on $\|u\|_\Omega$, it follows that $A(E)$ is bounded for all $E \notin \sigma_D$. An interpretation of A is that its sesquilinear form encodes the domain inner product of the interior Helmholtz extensions of two given boundary functions.

An inverse eigenvalue $\lambda(E)$ of A is defined at each $E \notin \sigma_D$ by

$$A(E)X(E) = \frac{1}{\lambda(E)}X(E) \;, \tag{2.4}$$

with corresponding eigenfunction $0 \neq X(E) \in L^2(\partial\Omega)$. We are interested in how the (inverse) spectrum of $A(E)$ depends on E. At this point we might stop to ask whether A even possesses a point spectrum. (2.1) has a solution operator $\mathcal{K}$ which maps U to u. For smooth domains (C^∞ boundary), $\mathcal{K}$ is bounded from $L^2(\partial\Omega)$ to $H^{1/2}(\Omega)$ (e.g., see [31]), thus by standard Sobolev embedding theorems, is compact from $L^2(\partial\Omega)$ to $L^2(\Omega)$. Since $A = \mathcal{K}^*\mathcal{K}$, where $\mathcal{K}^*$ is the adjoint with respect to the appropriate inner products, A is compact. Thus A has pure point spectrum, and the sequence $\{\lambda_i(A)\}_{i=1,2,\ldots}$ has no accumulation point. If Ω has corners, we expect (and have numerical evidence) that A remains compact but do not yet know a proof.

The w-weighted version of (1.3) is (we retain the same symbol)

$$t[u] \;:=\; \frac{\|u\|_{w,\partial\Omega}}{\|u\|_\Omega} \;. \tag{2.5}$$

For general w, we have the following simple lower bound on $t[u]$ in terms of the parameter-dependent spectrum of $A(E)$.

LEMMA 2.1. *Let $\lambda_1(E)$ be the lowest inverse eigenvalue of $A(E)$ defined by* (2.4). *Fix $E > 0$, then for all u obeying $-\Delta u = Eu$ in Ω, it holds that*

$$t[u] \;\geq\; \sqrt{\lambda_1(E)} \;, \tag{2.6}$$

where $t[u]$ is defined by (2.5).

Proof. Since A is self-adjoint and bounded,

$$\|u\|_\Omega^2 = \langle U, AU\rangle \leq \frac{1}{\lambda_1}\|U\|_{w,\partial\Omega}^2 \;, \tag{2.7}$$

and the proof follows from the definition (2.5). ∎

We now describe the behavior of this lowest inverse eigenvalue for parameter values E in the neighborhood of the Dirichlet spectrum of Ω. We will make frequent use of the weighted eigenmode boundary functions

$$\psi_j := \frac{1}{w}\partial_n\phi_j \;, \tag{2.8}$$

where $\partial_n\phi_j := \mathbf{n}\cdot\nabla\phi_j|_{\partial\Omega}$ is the usual normal derivative. These turn out to be the natural boundary representation of modes associated with the norm (1.7). We also need the infinite matrix $Q = (Q_{ij})$ whose entries are the weighted mode boundary inner products

$$Q_{ij} \;:=\; \langle\psi_i, \psi_j\rangle \;, \qquad i,j = 1,2,\ldots \;. \tag{2.9}$$

The proof of the following theorem is more involved and is the purpose of section 3.

THEOREM 2.2. *Fix $j \in \mathbb{N}$, and let p be the (finite) multiplicity of E_j. Without loss of generality choose j to be the first in the list of degenerate eigenvalues, that is, $E_j = E_{j+1} = \cdots = E_{j+p-1}$. Then in the limit as E tends to E_j, precisely p inverse eigenvalues of $A(E)$ vanish in the following fashion:*

$$\lambda_i(E) \;=\; c_j^{(i)}(E-E_j)^2 + O\left((E-E_j)^4\right) \;, \qquad i = 1, \ldots, p, \tag{2.10}$$

where the curvature coefficients $c_j^{(1)} \le c_j^{(2)} \le \cdots \le c_j^{(p)}$ are given by the inverse eigenvalues of the $p \times p$ submatrix $\tilde{Q} := (Q_{ik})_{i,k=j,\ldots,j+p-1}$. The smallest coefficient satisfies

$$\frac{1}{c_j^{(1)}} \;=\; \sup_{\phi\in\Phi, \|\phi\|_\Omega = 1} \int_{\partial\Omega} \frac{1}{w} |\partial_n \phi|^2 ds \;, \tag{2.11}$$

where $\Phi := \mathrm{Span}\{\phi_j, \ldots, \phi_{j+p-1}\}$ is the E_j-eigenspace.

This locally quadratic behavior of the smallest inverse eigenvalue $\lambda_1(E)$ versus frequency E is visible in Figure 1.1: $t_m(E)$ is a good approximation to $\sqrt{\lambda_1(E)}$ and appears locally linear. Note that when E_j is a simple eigenvalue ($p = 1$), this gives

$$\frac{1}{c_j^{(1)}} = \|\psi_j\|^2_{w,\partial\Omega} = \int_{\partial\Omega} \frac{1}{w} |\partial_n \phi_j|^2 ds \;. \tag{2.12}$$

The right-most expressions in (2.11) and (2.12) are now bounded above because of the following, which puts a limit on how "shallow" the parabola $\lambda_1(E)$ may be.

LEMMA 2.3 (Rellich). *Given a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, there are constants $C_{w,\Omega} \ge c_{w,\Omega} > 0$ depending only on Ω and the weight w, such that*

$$c^2_{w,\Omega} E_j \;\le\; \int_{\partial\Omega} \frac{1}{w} |\partial_n \phi_j|^2 ds \;\le\; C^2_{w,\Omega} E_j \qquad \textit{for all } j = 1, 2, \ldots \;. \tag{2.13}$$

For the special case $w = (\mathbf{x}\cdot\mathbf{n})^{-1}$, one may choose $c_{w,\Omega} = C_{w,\Omega} = \sqrt{2}$, independent of Ω, in which case the inequalities become an equality.

The inequalities are proved in Hassell–Tao [24, section 2] in the case $w = 1$ and $\partial\Omega$ smooth but carry over trivially to strictly positive bounded w. The lower bound uses (2.14) below, thus holds for Lipschitz domains. They derived the upper bound using commutator estimates involving a smooth vector field which points normally outwards on $\partial\Omega$. (Also see [19, Lemma 2.1] for the case of domains with a Lipschitz normal vector field.) A simple generalization of this vector field construction to be smooth with merely positive normal component on $\partial\Omega$ allows the upper bound to carry over to Lipschitz domains. For instance, the vector field $\mathbf{x}$ works for strictly star-shaped Lipschitz domains [23]. The special-weight case in Lemma 2.3 is due to Rellich [38], who proved

$$\int_{\partial\Omega} \mathbf{x}\cdot\mathbf{n}\, |\partial_n \phi|^2 ds \;=\; 2E_j \tag{2.14}$$

for any bounded Ω and dimension d (regardless of whether Ω is star-shaped).

Remark 2.1. The argument and Remark of [24, section 2] carries over to the case of Ω possessing a boundary of only C^1 continuity, since the first derivatives of the constructed vector field remain bounded. Therefore, with $w = 1$, we get that $C_{w,\Omega}$ is the inverse of the largest δ such that the function $r(\mathbf{x}) = \mathrm{dist}(\mathbf{x}, \partial\Omega)$, $\mathbf{x} \in \Omega$, is C^1 continuous wherever $r \in [0, \delta]$. For domain (H) of Figure 1.2, this gives $C_{w,\Omega} = 10/3$, the inverse of the minimum radius of curvature.

2.1. Proof of main theorem. We now have all the pieces to prove Theorem 1.1, more precisely, its generalization to general strictly positive bounded weight w. For a given eigenmode number j, Theorem 2.2 implies that $\lim_{E\to E_j} c_j^{(1)}(E-E_j)^2/\lambda_1(E) = 1$, thus there is a $\delta_j > 0$ such that

$$\sqrt{\frac{c_j^{(1)}}{\lambda_1(E)}}|E-E_j| \;\le\; 1+\epsilon \tag{2.15}$$

holds for all E in the punctured interval $0 < |E-E_j| < \delta_j$. By choosing $\Sigma_\epsilon = \bigcup_{j=1}^{\infty}\{E : |E-E_j| < \delta_j\}$, then for each $E \in \Sigma_\epsilon \setminus \sigma_D$, there exists a j such that (2.15) holds. Substituting into (2.15), the bound from Lemma 2.1 and the upper bound in Lemma 2.3 completes the proof of Theorem 1.1, with the constant C_Ω being equal to the upper constant $C_{w,\Omega}$ in Lemma 2.3.

Remark 2.2. Theorem 2.2 is stronger than necessary for this proof, since only the existence of $\lim_{E\to E_j} c_j^{(1)}(E-E_j)^2/\lambda_1(E)$ is needed. In fact, the full analyticity of $\lambda_1(E)$ will be proved in section 3.

3. Proof of quadratic behavior of inverse eigenvalues of A. Here (with Appendices A and B) we prove Theorem 2.2. The first part involves a modal expansion, which allows us to split the operator $A(E)$ into a finite-rank part associated with the jth eigenspace, and a remainder, which we show is analytic in the parameter E in the neighborhood of E_j. We then apply Rellich–Kato analytic operator perturbation theory to show that the inverse eigenvalues $\{\lambda_i(E)\}_{i=1,\dots,p}$ have quadratic behavior about E_j. However, there is a twist: the needed eigenvalue(s) of $A(E)$ have poles at $E = E_j$. This we may cancel by scalar multiplication, but the price to pay is that analytic perturbation theory is no longer applicable to the resulting infinite multiplicity zero eigenvalue. Nevertheless, the Cauchy interlacing property will recover what is needed.

3.1. Mode expansion of the boundary operator. The key to our proof is to express our boundary operator A as an infinite sum of rank-1 operators associated with each eigenmode boundary function.

LEMMA 3.1.

$$A(E) = \sum_{j=1}^{\infty} \frac{\psi_j\langle\psi_j,\cdot\rangle}{(E-E_j)^2}. \tag{3.1}$$

Remark 3.1. A similar formula (for the case $w = 1$) is known [34] for the Dirichlet-to-Neumann map for (2.1). In fact, when $w = 1$ it can be shown directly that $-A(E)$ is the E-derivative of the Dirichlet-to-Neumann map [17, equation (2.5)].

Proof. The Helmholtz equation Green's function G in the domain Ω is defined by [18]:

$$\begin{aligned} (-\Delta_{\mathbf{x}} - E)G(\mathbf{x},\mathbf{y}) &= \delta(\mathbf{x}-\mathbf{y}) && \text{for } \mathbf{x},\mathbf{y} \in \Omega,\\ G(\mathbf{x},\mathbf{y}) &= 0 && \text{for } \mathbf{x} \in \partial\Omega \text{ or } \mathbf{y} \in \partial\Omega. \end{aligned} \tag{3.2}$$

G depends on E, although we shall not indicate this explicitly. δ is the Dirac delta distribution in $\mathbb{R}^d$. Then the solution operator $\mathcal{K}$ for the BVP (2.1) is

$$u(\mathbf{x}) = (\mathcal{K}U)(\mathbf{x}) = \int_{\partial\Omega} \mathcal{K}(\mathbf{x},s)U(s)\,ds \qquad \text{for } \mathbf{x} \in \Omega\ , \tag{3.3}$$

where $s \in \partial\Omega$ is a boundary coordinate and the Poisson kernel for the Helmholtz equation is

$$\mathcal{K}(\mathbf{x}, s) = \partial_{n(s)} G(\mathbf{x}, s), \tag{3.4}$$

where here $\partial_{n(s)}$ indicates the normal derivative with respect to the second argument, evaluated at the boundary point s. From (3.3) we get

$$(u, v)_\Omega \;=\; \int_{\partial\Omega}\int_{\partial\Omega} \overline{U(s)} V(s') \int_\Omega \partial_{n(s)} \overline{G(\mathbf{x}, s)} \partial_{n(s')} G(\mathbf{x}, s')\, d\mathbf{x}\, ds ds' \;. \tag{3.5}$$

Since (2.3) holds for arbitrary $U, V \in L^2(\partial\Omega)$, we may write the action of A on any $V \in L^2(\partial\Omega)$ as the integral operator $(AV)(s) = \int_{\partial\Omega} A(s, s') V(s')\, ds'$ whose kernel is

$$A(s, s') = \frac{1}{w(s)} \int_\Omega \partial_{n(s)} \overline{G(\mathbf{x}, s)} \partial_{n(s')} G(\mathbf{x}, s')\, d\mathbf{x}. \tag{3.6}$$

Inserting the usual eigenfunction expansion [18]

$$G(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{\infty} \frac{\overline{\phi_j(\mathbf{x})} \phi_j(\mathbf{y})}{E - E_j} \tag{3.7}$$

into (3.6), orthonormality on Ω kills the domain integral and one of the sums, leaving

$$A(s, s') = \frac{1}{w(s)} \sum_{j=1}^{\infty} \frac{\partial_n \overline{\phi_j(s)}\, \partial_n \phi_j(s')}{(E - E_j)^2}. \tag{3.8}$$

The apparent lack of symmetry of the kernel is deceptive and is merely a result of the weighted inner product (2.2). It is simple to check that this kernel is equivalent to the desired symmetric formula (3.1). ∎

Remark 3.2. Despite its suggestive form, (3.1) is not a sum of orthogonal projections, nor a spectral representation, since $\{\psi_j\}$ are not unit norm and, although complete in $L^2(\partial\Omega)$, they are generally not orthogonal.

Remark 3.3. Although each term in (3.1) is a bounded operator, with bound $C^2_{w,\Omega} E_j/(E - E_j)^2$ following from Lemma 2.3, the sum of these bounds is divergent for all $d \geq 2$. This follows from Weyl's law [18, Chapter 11], which states that asymptotically, the number of Laplacian eigenvalues below E grows like $E^{d/2}$. Therefore, (3.1) *cannot* directly be used to show that A is the norm limit of finite-rank operators and hence compact.

3.2. Perturbation theory for inverse eigenvalues. From now on we fix the mode number j, with multiplicity p. Recall if $p > 1$, then for convenience, j is chosen to be the lowest in the degenerate set $j, \ldots, j + p - 1$. It is a standard result that p is finite, since the Laplacian has compact inverse [13, section 6.5.1]. We separate the terms in the sum (3.1) associated with this set, giving

$$(E - E_j)^2 A(E) \;=\; B_j \;+\; (E - E_j)^2 A_j(E) \;, \tag{3.9}$$

where the fixed rank-p operator associated with the E_j-eigenspace is

$$B_j := \sum_{m=1}^{p} \psi_{j+m-1} \langle \psi_{j+m-1}, \cdot \rangle \tag{3.10}$$

and the remainder is a scalar multiple of the bounded operator

$$A_j(E) := \sum_{m\in\mathbb{N}, m\notin[j,j+p-1]} \frac{\psi_m\langle\psi_m,\cdot\rangle}{(E-E_m)^2} \ . \tag{3.11}$$

With respect to the parameter E, the latter operator has second-order poles at all eigenvalues not equal to E_j. It is also analytic near E_j (the proof is in Appendix A).

LEMMA 3.2. *$A_j(E)$ is a bounded regular (analytic) operator in the sense of Rellich* [39, p. 55] *for E in a neighborhood of E_j. That is, for each $f \in L^2(\partial\Omega)$, the element $A_j(E)f$ exists as a power series $f_0 + (E-E_j)f_1 + (E-E_j)^2 f_2 + \cdots$ convergent in $L^2(\partial\Omega)$ for E in a neighborhood of E_j.*

Now define $\{\mu_i(E)\}$ to be the set of parameter-dependent eigenvalues of the operator $(E-E_j)^2 A(E)$. We wish to understand their behavior when E is close to E_j, since from them we can simply recover the desired $\lambda(E)$ as defined in (2.4). We first characterize the eigenvalues of B_j in terms of $\tilde{Q}$, the mode boundary inner product matrix for the eigenspace.

LEMMA 3.3. *B_j has (counting multiplicity) precisely p nonzero eigenvalues $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_p$, which are identical to the set of eigenvalues (counting multiplicity) of the matrix $\tilde{Q}$ defined by Theorem* 2.2 *and* (2.9). *B_j also has a zero eigenvalue of infinite multiplicity.*

Proof. Given $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p)^T$, define $\phi = \sum_{m=1}^p \beta_m \phi_{j+m-1}$, then by orthonormality $\|\phi\|_\Omega = |\boldsymbol{\beta}|$, the usual Euclidean norm in $\mathbb{C}^p$. Thus ϕ is an $L^2(\Omega)$-normalized eigenmode whenever $|\boldsymbol{\beta}| = 1$. Its weighted normal derivative is $\psi = \frac{1}{w}\partial_n\phi$. Then, using (2.9) and (2.13) we have

$$\inf_{|\boldsymbol{\beta}|=1} \boldsymbol{\beta}^H \tilde{Q}\boldsymbol{\beta} = \inf_{|\boldsymbol{\beta}|=1} \sum_{m,n=1}^p \overline{\beta_m}\beta_n \langle\psi_{j+m-1}, \psi_{j+n-1}\rangle = \|\psi\|^2_{w,\partial\Omega} \geq c^2_{w,\Omega} E_j > 0. \tag{3.12}$$

Thus all eigenvalues of $\tilde{Q}$ are positive, and also the vectors $\{\psi_{j+m-1}\}_{m=1,\ldots,p}$ are linearly independent. From the definition (3.10) any eigenvector of B_j with nonzero eigenvalue must lie in $\mathrm{Span}\{\psi_{j+m-1}\}_{m=1,\ldots,p}$. Suppose $\sum_{m=1}^p \beta_m\psi_{j+m-1}$ is such an eigenvector with eigenvalue ν. Using (2.9), this is equivalent to the statement

$$\sum_{m,n=1}^p \beta_m\psi_n\tilde{Q}_{mn} = \nu \sum_{n=1}^p \beta_n\psi_n \ . \tag{3.13}$$

By linear independence, we equate coefficients of ψ_n, so (3.13) is equivalent to $\tilde{Q}\boldsymbol{\beta} = \nu\boldsymbol{\beta}$, and these two problems have the same set of nonzero eigenvalues. The infinite multiplicity zero eigenvalue exists, since B_j has finite rank (degenerate kernel). ∎

We would like to apply perturbation theory in the small parameter $(E-E_j)$ to (3.9). Unfortunately, its infinite multiplicity prevents the results of analytic perturbation theory from being applied directly to the zero eigenvalue of B_j. We circumvent this problem[1] by appealing to the following classical property of the point spectrum under a rank-1 perturbation (for convenience, we include a proof in this operator setting in Appendix B).

LEMMA 3.4 (Cauchy interlacing). *Let S be a self-adjoint operator in a Hilbert space, bounded from below, with eigenvalues $\sigma_1 \leq \sigma_2 \leq \cdots$. Let $\psi\langle\psi,\cdot\rangle$ define a*

[1] The author is indebted to Percy Deift for this argument.

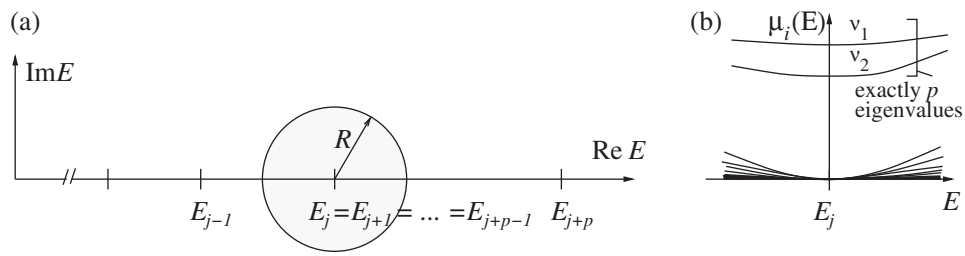


FIG. 3.1. (a) *Complex E plane showing disc around the current eigenvalue E_j used in the proof of Lemma* 3.2. *The operator $A(E)$ has second-order poles at each of the E_j.* (b) *Sketch of the spectrum $\{\mu_i(E)\}$ of the operator $(E-E_j)^2A(E)$ used to prove Theorem* 2.2.

*(nonnegative) rank-*1 *operator, and let the eigenvalues of $S-\psi\langle\psi,\cdot\rangle$ be $\tau_1 \le \tau_2 \le \cdots$. Then $\tau_1 \le \sigma_1 \le \tau_2 \le \sigma_2 \le \tau_3 \le \cdots$.*

For real $E \neq E_j$ in the domain of analyticity of $A_j(E)$, we proceed inductively as follows. If $p=1$, this lemma applies with $S=-(E-E_j)^2A_j(E)$ and $\psi=\psi_j$. The result is that at most one eigenvalue $\mu_1(E)$ may exceed the largest eigenvalue of $(E-E_j)^2A_j(E)$. For $p>1$, we may apply the interlacing lemma p times and conclude that at most p eigenvalues $\mu_1,\mu_2,\dots$ exceed the largest eigenvalue of $(E-E_j)^2A_j(E)$. The eigenvalues of $A_j(E)$ are analytic by the theorem of Rellich [39, Theorem II.2.1], which applies due to Lemma 3.2. Therefore, the eigenvalues of $(E-E_j)^2A_j(E)$ have convergent power series beginning with $(E-E_j)^2$ and hence vanish as $E\to E_j$. This leaves a set of at most p eigenvalues $\mu_i(E)$, which may have nonzero limits as $E\to E_j$. This behavior is illustrated in Figure 3.1b. So we have proved that there is, after all, no complication introduced by the infinite multiplicity of the zero eigenvalue. Since by (3.9), $(E-E_j)^2A(E)$ is also bounded analytic in a neighborhood of E_j, the Rellich theorem tells us that members of the above-mentioned set $\mu_i(E)$ are analytic in a neighborhood of E_j, so at $E=E_j$ their limits must coincide with the set $\{\nu_i\}_{i=1,\dots,p}$ of nonzero eigenvalues of B_j. Thus we have proved that there are p eigenvalues of the operator $(E-E_j)^2A(E)$ behaving as

$$\mu_i(E)=\nu_i+O\left((E-E_j)^2\right), \qquad i=1,\dots,p. \tag{3.14}$$

Note that there is no first-order term in $(E-E_j)$ because the perturbation term in (3.9) has a power series beginning with $(E-E_j)^2$.

Eigenvalues $\mu_i(E)$ and $\lambda_i(E)$ are connected for $E\neq E_j$ via (2.4) and (3.9), so

$$\lambda_i(E)=\frac{(E-E_j)^2}{\mu_i(E)}\ . \tag{3.15}$$

Combining this with (3.14) shows that there are precisely p inverse eigenvalues of $A(E)$ which vanish as $E\to E_j$, proving the first part of Theorem 2.2:

$$\lambda_i(E)=\frac{1}{\nu_i}(E-E_j)^2+O\left((E-E_j)^4\right)\ , \qquad i=1,\dots,p. \tag{3.16}$$

Finally, we now place a lower bound on the curvature coefficient of the asymptotically smallest eigenvalue $\lambda_1(E)$. Comparing the above to (2.10) and using Lemma 3.3 yields $c_j^{(1)}=\nu_1^{-1}=(\sup_{|\boldsymbol{\beta}|=1}\boldsymbol{\beta}^H\tilde{Q}\boldsymbol{\beta})^{-1}$. Noticing that $\boldsymbol{\beta}^H\tilde{Q}\boldsymbol{\beta}=\|\psi\|^2_{w,\partial\Omega}=\int_{\partial\Omega}\frac{1}{w}|\partial_n\phi|^2ds$, where ϕ and ψ are defined as in the proof of Lemma 3.3, completes the proof.

4. Implementation and performance of eigenvalue inclusion. Here we discuss an efficient two-dimensional ($d = 2$) implementation of the MPS, then show that, in practice, Theorem 1.1 gives useful eigenvalue inclusion bounds even for ϵ much smaller than 1. As illustrated in Figure 1.1, these supercede the Moler–Payne bounds.

4.1. Computation of minimum boundary error $t_m(E)$. At each given frequency E, inserting the basis representation $u = \sum_{n=1}^{N} \alpha_n \xi_n$ converts the minimum boundary error (1.5) into the Rayleigh quotient (here, $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_N)^T$ and ^{H} indicates conjugate transpose)

$$t_m(E) = \min_{u \in \mathrm{Span}\{\xi_n\}} \frac{\|u\|_{w,\partial\Omega}}{\|u\|_\Omega} = \min_{\boldsymbol{\alpha} \neq \mathbf{0}} \sqrt{\frac{\boldsymbol{\alpha}^H F(E) \boldsymbol{\alpha}}{\boldsymbol{\alpha}^H G(E) \boldsymbol{\alpha}}} = \sqrt{\hat{\lambda}_1(E)} \ , \tag{4.1}$$

where $\hat{\lambda}_1(E)$ is the smallest generalized eigenvalue of the matrix pencil defined by

$$F\boldsymbol{\alpha} = \hat{\lambda} G \boldsymbol{\alpha} \ . \tag{4.2}$$

Note that here, and from now on, we use the weighted boundary norm (1.7) in place of $\|u\|_{\partial\Omega}$. The matrix elements are (for $m, n = 1, \ldots, N$)

$$F_{mn}(E) := \int_{\partial\Omega} w(s) \overline{\xi_m(s)} \xi_n(s) \, ds \ , \qquad G_{mn}(E) := (\xi_m, \xi_n)_\Omega \ . \tag{4.3}$$

With a good choice of basis set (e.g., see section 4.2), $\hat{\lambda}_1(E)$ is very close to its minimum achievable value $\lambda_1(E)$.

We now discuss how we fill F and G efficiently, which is not trivial at high frequency. Naively, we might expect that N^2 separate boundary integrals are required to fill F, where N is the basis size. As we discuss below (and in [2]) to achieve reasonable approximations of eigenmodes, N must scale like $\sqrt{E}$ at high frequency. Each such integral must be approximated using a set of M quadrature points $\mathbf{x}_i \in \partial\Omega$. Since integrands oscillate on the wavelength scale, it turns out that M must also scale like $\sqrt{E}$, so M is the same order as N. Thus $O(N^3)$ basis function evaluations would be required to fill F. However, we can reduce this to $O(N^2)$ by writing $F \approx A^H W A$, where the $M \times N$ rectangular matrix A has elements $A_{in}(E) := \xi_n(\mathbf{x}_i)$ and W is a diagonal matrix containing the products of $w(\mathbf{x}_i)$ and the quadrature weights. The dense matrix-matrix product remains $O(N^3)$, but now the prefactor is much smaller (since basis functions are typically trigonometric or Bessels and may require 10^2 flops or more per evaluation).

The naive cost of filling G is worse: each domain integral would require $O(N^2)$ quadrature points to handle the oscillatory integrand; even using the factorization above would bring this only down to $O(N^3)$ basis evaluations. However, we may make use of the following identity [4, Lemma 3.1] (also see [3, Appendix H]). Let $-\Delta u = Eu$ and $-\Delta v = Ev$ hold in any Lipschitz domain $\Omega \in \mathbb{R}^2$, then

$$(u, v)_\Omega = \frac{1}{2E} \int_{\partial\Omega} (\mathbf{x} \cdot \mathbf{n})(E\overline{u}v - \nabla\overline{u} \cdot \nabla v) + (\mathbf{x} \cdot \nabla\overline{u})(\mathbf{n} \cdot \nabla v) + (\mathbf{n} \cdot \nabla\overline{u})(\mathbf{x} \cdot \nabla v) \, ds \tag{4.4}$$

expresses the domain inner product as a sum of four boundary inner products. Substituting this into (4.3) using $u = \xi_m$ and $v = \xi_n$, we see that G can be written as a sum of four boundary integral matrices of similar form to F. By computing each

of the four terms using a product of rectangular matrices as with F, we may, therefore, fill G using only $O(N^2)$ basis (and first-derivative) evaluations plus a few dense matrix-matrix products which are $O(N^3)$ with a small prefactor.

Solving the generalized eigenvalue problem (4.2) has an interesting twist: it turns out to become numerically singular, with F and G effectively acquiring a common nullspace whenever N is chosen large enough to achieve high accuracy. This is due to the fact that as N increases, there arise unit-norm linear combinations of basis functions which have exponentially small values in the closure of Ω; this ill-conditioning of the basis appears to be a general feature of MPS type methods [12, 5, 2]. As a consequence, standard methods, such as Cholesky factorization of the (formally) positive-definite matrix G [20, section 8.7] or QZ decomposition [20, section 7.7], do not produce reliable eigenvalues of (4.2).[2] However, by using the following regularization method, we project out the numerical nullspace and accurately compute the remaining "stable" (in the sense of [15]) eigenvalues. We diagonalize the Hermitian matrix $G = VDV^H$ (where V is unitary), then define $\tilde{D}$ to be the diagonal matrix containing only the subset of entries of D which exceed a cutoff ε_{reg} (typically 10^{-15}) times D's largest entry. Defining $\tilde{V}$ to contain as columns the corresponding subset of eigenvectors and

$$\tilde{F} := \tilde{D}^{-1/2}\tilde{V}^H F \tilde{V}\tilde{D}^{-1/2} \ , \tag{4.5}$$

we diagonalize $\tilde{F} = U\Lambda U^H$. The diagonal matrix Λ contains approximations to the stable eigenvalues $\hat{\lambda}$ of the pencil (4.2), with eigenvectors as the columns of $\tilde{V}\tilde{D}^{-1/2}U$.

The above method was used by Vergini–Saraceno [46] and is similar to a more general procedure of Fix–Heiberger [15], who show that the error introduced into the stable eigenvalues is $O(\varepsilon_{\text{reg}})$. In practice, we find that the errors are bounded by roughly $10^3\varepsilon_{\text{reg}}$. We believe that this prefactor reflects the closeness of the nullspaces of F and G; however, this method deserves a full analysis, which we will not attempt here. We also find (by extrapolating in ε_{reg}) that the approximate $\hat{\lambda}_1$ overestimates the true value. Note that Betcke's recent GSVD method [7] includes a regularization similar to the above and can give higher accuracy; we discuss this in comment 2 of section 5. Finally, since only the lowest (or lowest few) eigenvalues are needed, one might expect that an iterative method could improve upon the above $O(N^3)$ effort.

4.2. Testing the applicability of the main theorem. Theorem 1.1 is an asymptotic result: given an $\epsilon > 0$, it does not tell us in how large a (punctured) neighborhood of each E_j the inequality (1.6) holds, only that some neighborhood exists. We now show that, in practice, for a small choice of ϵ, the valid neighborhoods around each E_j coalesce so that the theorem may be applied to essentially the whole positive real axis, i.e., $\Sigma_\epsilon = [a, \infty)$ for some small a. Our three test domains are defined and shown in Figure 1.2 and represent three classes of interest: (S) is smooth, (T) is a triangle with one singular ($5\pi/12$) corner, and (H) is a "horseshoe" with C^1 continuous boundary. Recall that a singular corner is a corner with angle not equal to π/n for some $n \in \mathbb{N}$ [5]. (S) and (T) are strictly star-shaped, but (H) is not.

Before presenting results, we discuss implementation details. To compute $t_m(E)$, we used the method of the previous section. For (S) and (H), we use as basis functions ξ_n fundamental solutions with origins (source points) placed along an exterior curve a distance of order one wavelength from Ω. Such bases are efficient in many domains [14, 12] and show spectral convergence in analytic domains [2]. For (T), we use a cylindrical wave expansion (generalized harmonic polynomials), in polar coordinates

[2] We implement these with the `eig` command of MATLAB version 2007b.

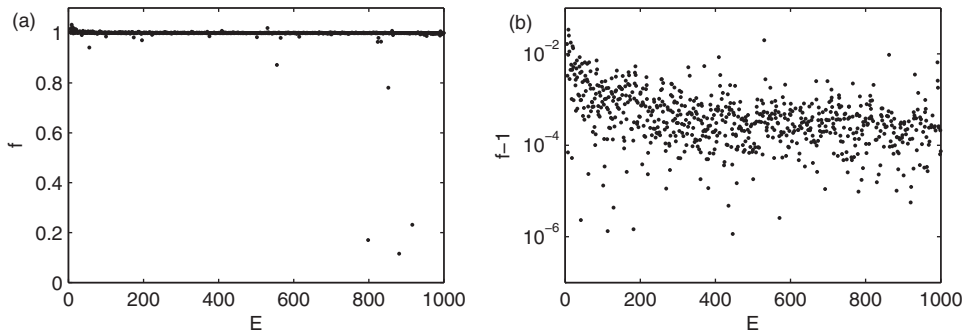


FIG. 4.1. (a) *Relative slope measure* $f(E)$ *defined by* (4.6)*, for the smooth domain (S) with weight* $w = (\mathbf{x}\cdot\mathbf{n})^{-1}$*, plotted at* $E = 5, 6, 7, \ldots, 1000$. (b) *Logarithmic plot of* $f(E) - 1$ *(no point is plotted if* $f(E) < 1$*). The value* $f(E) - 1$ *is the smallest value of* ϵ *in Theorem* 1.1 *such that* $E \in \Sigma_\epsilon$.

$\xi_n(r,\theta) = e^{in\theta} J_n(\sqrt{E}r)$ for $-N/2 \le n < N/2$ and N even. This basis is convergent in any bounded Ω with a finite number of corners [41, Corollary to Theorem I6.1]. It is extremely fast to evaluate (at each r, all n values may be computed via a single Miller's downward recurrence [35, section 5.4]), and we found that it produced values of $t_m(E)$ for $E = E_j$, $j < 10^3$, no larger than 10^{-3}, adequate for our needs. (Note that this could be improved upon by adding functions adapted to the singular corner [10, 5].) In all three domains the scaling of basis size with wavenumber was linear: we typically chose between 3 to 6 degrees of freedom per wavelength on the boundary. Boundary quadrature was equispaced in angle for (S), and Clenshaw–Curtis [44] on each side of (T) and each arc of (H). In all cases, the number of quadrature points M was chosen to be large enough to make the results insensitive to M (as discussed in [2]); generally, M was between N and $2N$. For (S) and (T), Dirichlet eigenvalues E_j were independently located with the scaling method [46, 3, 4] to an absolute error of roughly 10^{-3}; for (H), these were located by finding minima of $t_m(E)$. In all cases, the first two terms of Weyl's law for the asymptotics of eigenvalues [22] was used to eliminate the possibility of missing eigenvalues.

Motivated by (1.6), given some constant C_Ω, we will study the "relative slope measure"

$$f(E) := \frac{1}{C_\Omega t_m(E)} \min_j \frac{|E - E_j|}{\sqrt{E_j}} \ . \tag{4.6}$$

The interpretation is that $f(E) - 1$ gives a strict lower bound on ϵ such that $E \in \Sigma_\epsilon$. A value of f close to 1 shows that the perturbative result Theorem 2.2 that we used to prove Theorem 1.1 is accurately predicting the slopes in curves such as Figure 1.1.

We start with domain (S), and choose $w = (\mathbf{x}\cdot\mathbf{n})^{-1}$, thus use $C_\Omega = \sqrt{2}$. In Figure 4.1 we show $f(E)$ computed at $E = 5, 6, 7, \ldots, 1000$. Note that the lowest 242 Dirichlet eigenvalues E_j lie in the interval $[5, 1000]$ and have no known correlation with the integers (we have checked that random sampling of E gives the same behavior). It is clear that f is almost always close to 1, in fact, the largest value is $f(9) = 1.0335$. Panel (b) shows the overall tendency that f values approach 1 from above as E grows. Occasionally, f values are significantly smaller than 1: this happens only when E is very close to E_j for some j and limitations of the basis set prevent $t_m(E)$ from reaching its theoretical minimum. Occasionally there is an anomalously large f value at higher E: in all cases these spikes in $f(E)$ occur for E very close to E_j for some j and are due to the errors in the computed E_j. For instance, the largest such value shown is

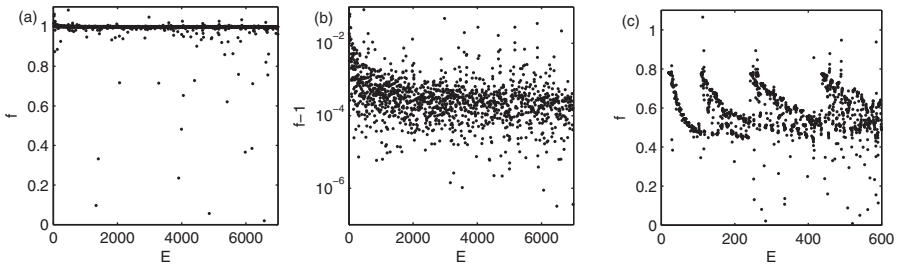


FIG. 4.2. (a,b) *Same as Figure* 4.1 *except for the nonsmooth triangle domain (T) with weight* $w = (\mathbf{x} \cdot \mathbf{n})^{-1}$*, plotted at* $E = 15, 18, \ldots, 6999$. (c) *Same as* (a) *except for horseshoe domain (H) with weight* $w = 1$*, plotted at* $E = 20, 20.3, \ldots, 599.9$.

$f(530) = 1.0198$, where E is very close to $E_j = 530.023$. The same tendency persists at higher frequencies: computing 285 samples of $f(E)$ at $E = 10007, 10014, \ldots, 11995$, with $N = 662$, $M = 1277$, and using the 517 eigenvalues E_j lying in $[10000, 12000]$, we found that the largest value was $f(10511) = 1.030$. This again was traced to error in the nearby eigenvalue $E_j = 10510.768$, and all exact f values are smaller than this. Thus for domain (S), there is strong evidence that a small value such as $\epsilon = 0.05$ is sufficient for the theorem to hold for all parameter values $E > 5$.

For the triangle (T), we use the same weight and C_Ω. We compute $f(E)$ at 2329 values in the interval $E = [15, 6999]$, which includes the lowest 625 Dirichlet eigenvalues. As Figure 4.2(a) and (b) show, the results are essentially identical to the smooth domain: f values tend to decrease towards 1 with increasing E. The largest value found was $f(468) = 1.0840$. Finally, for the horseshoe (H), which is not star-shaped, we choose $w = 1$ and $C_\Omega = 10/3$ (see Remark 2.1). Figure 4.2(c) shows $f(E)$ computed at 1934 values in the interval $E = [20, 600]$, which includes the lowest 137 Dirichlet eigenvalues. Other than the value $f(113) = 1.065$, f is always less than 1. Periodicity in slope is evident, due to transverse mode cutoffs in the radial direction.

In conclusion, ϵ may be chosen much less than 1, thus its exact value is inconsequential for the number of significant digits which may be claimed using Theorem 1.1.

4.3. Application to accurate inclusion of high eigenvalues. Taking domain (S) as an example, we now show how, at high eigenvalue, the bounds achievable via Theorem 1.1 improve upon those of Moler–Payne (1.4). For the constant in the latter, one has $C'_\Omega = q_1^{-1/2}$, where q_1 is the lowest eigenvalue of a Stekloff eigenproblem for the domain [28, (2.11)]. In star-shaped domains $q_1 \geq E_1^{1/2} r/2R$, where r and R are, respectively, the minimum and maximum of $\mathbf{x} \cdot \mathbf{n}$ on $\partial\Omega$ [28, Table I]. For domain (S), this gives $C'_\Omega = 1.43$. One could improve this by an $O(1)$ factor by solving the Stekloff problem, however, this would scarcely affect our conclusions.

In Figure 4.3, we show an approximate mode of domain (S) with approximate eigenvalue $E = 100005.2660$, computed using $N = 1666$ and $M = 3690$. (Note that this and the following trial eigenvalue were found via the scaling method, although this is not relevant for our results.) Its boundary error $t \approx 1.8 \times 10^{-7}$ enables the following claims: via the Moler–Payne theorem, we have an inclusion interval for some true eigenvalue $E_j \in [100005.24, 100005.29]$, whereas via Theorem 1.1, we get $E_j \in [100005.2659, 100005.2661]$, i.e., a relative error of 10^{-9} and an improvement by 2.5 digits over Moler–Payne.

At higher frequency still $E = 1000019.50$, about 300 wavelengths across the domain, we find an approximate mode with boundary error $t \approx 3 \times 10^{-5}$, using

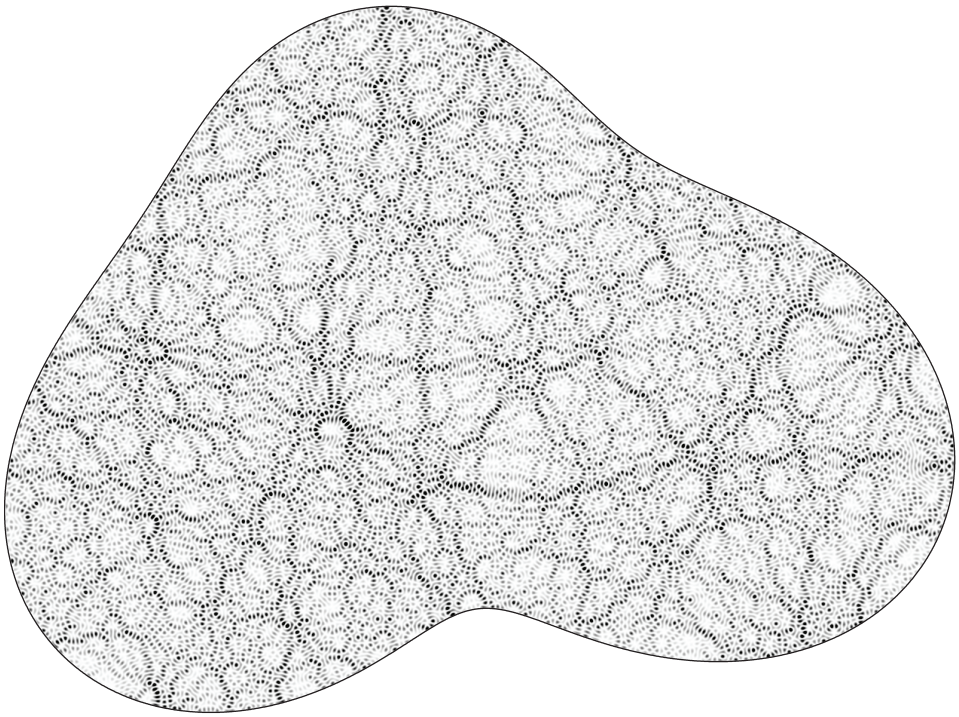


FIG. 4.3. *High-lying approximate eigenmode of domain (S) with eigenvalue E_j lying in the interval* $[100005.2659, 100005.2661]$ *whose bounds were computed from Theorem* 1.1 *with* $\epsilon = 0$. *These bounds are* 2.5 *digits (*320 *times) better than the Moler–Payne Theorem* 1.4. *The intensity* $|\phi_j(\mathbf{x})|^2$ *is shown on a greyscale of* 0 *(white) to* 2 *(black). The mode number is* $j \approx 25823$ *(estimated to within* $O(1)$ *error using* 2 *terms of Weyl's asymptotic expansion* [22]*). There are about* 100 *wavelengths across the domain.*

$N = M = 4372$, which corresponds to only three degrees of freedom per wavelength on the boundary. (To give the reader an idea of CPU time, computation of $\hat{\lambda}_1$ at a single E value took 33 minutes.) By Moler–Payne, we have that there is some eigenvalue $E_j \in [999978, 1000061]$; note that this bound is scarcely useful, since it is much wider than the domain's mean eigenvalue spacing of 3.8. From our theorem we get $E_j \in [1000019.45, 1000019.54]$, a factor 10^3 improvement in accuracy without extra effort.

How high in E may one go? We note the following limitation. At $E \approx 10^6$, we found that the lowest eigenvalue $\hat{\lambda}_1$ was sensitive to ε_{reg}, hence we independently verified the boundary error t by direct evaluation of u using the generalized eigenvector coefficients. This shows that we are starting to reach the useful limit of our regularization method; see comment 2 of section 5 for a possible solution.

5. Conclusion and discussion. By analyzing the slopes of the MPS curves, we have proved a new bound (Theorem 1.1) on Dirichlet eigenvalues E_j of smooth domains in terms of the boundary L^2 norm (1.5) of trial functions (Helmholtz solutions). This bound has better scaling at high frequency than existing bounds, improving upon them by a factor of $O(\sqrt{E})$, i.e., the wavenumber. Our main tool was analytic perturbation theory of the frequency-dependent spectrum of the natural $L^2(\partial\Omega)$ operator associated with (1.3) (or (2.5)); this gave us parabolic behavior at the bottom of the spectrum but limited the theorem to a (possibly small) neighborhood of each E_j.

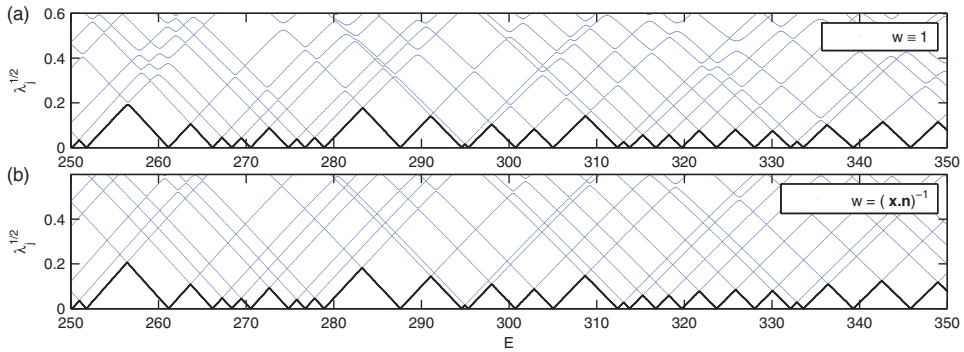


FIG. 5.1. *Higher generalized eigenvalues of* (4.2) *(the lowest* $\hat{\lambda}_1(E)$ *is shown as a thick line) in domain (S), with weight function* (a) *constant* $w = 1$, (b) *special weight* $w = (\mathbf{x} \cdot \mathbf{n})^{-1}$. *Observe that the slight mode-to-mode variations in slope and "avoided crossings" in* (a) *nearly vanish in* (b).

We provided strong numerical evidence that, in fact, these neighborhoods are large; thus we believe that, in practice, the theorem may be used with confidence for all $E \in [a, \infty)$ for some $a < E_1$, given a small choice of ϵ such as 10^{-1}. This we checked at thousands of E values for smooth, polygonal and C^1 domains. We also presented an efficient implementation of the MPS, which involves boundary evaluations alone, and we demonstrated a 3-digit accuracy gain via our theorem at very high frequencies.

We end with several open questions and research directions of interest.

1. We suspect Theorem 1.1 could be extended to Lipschitz domains (either via proving compactness of $A(E)$ or by taking into account a possible essential spectrum in section 3.2). Another natural next step is to seek corresponding bounds on eigenfunction error $\|u - \phi_j\|_{L^2(\Omega)}$ (as in [33, 29]).
2. Since the boundary error norm $t_m(E)$ is the square-root of a generalized eigenvalue, our method cannot access t values below roughly $\sqrt{\varepsilon_{\text{reg}}}$, which must exceed $\sqrt{\varepsilon_{\text{mach}}} \approx 10^{-8}$ in double-precision. Betcke's GSVD method [7], by working directly with A from section 4.1, avoids matrix squaring so is able to reach t of order $\varepsilon_{\text{mach}}$, but requires estimation via interior points. It would be valuable to seek an efficient way to construct a "square-root" of G using the boundary alone, possibly via (4.4).
3. Still [42, Theorem 4] has applied the Kato–Temple inequality to the MPS, effectively replacing $t[u]$ in (1.4) by $t[u]^2$, which is clearly advantageous at small t. The disadvantage is that a PDE must be solved; however, its utility should be compared to that of Theorem 1.1.
4. To understand why the quadratic approximation (2.10) predicts MPS eigenvalues so well, we need the 4th- and higher-order terms, via a perturbation expansion. In a similar vein, the higher generalized eigenvalues, shown by Figure 5.1, contain much tantalizing and potentially useful structure and so should be analyzed. This might lead to bounds on multiple close eigenvalues, as in recent work of Eisenstat [11].
5. Given a Lipschitz domain, how should an optimal value for $C_{w,\Omega}$ be computed for Lemma 2.3?
6. We have shown that the MPS is a powerful tool for high-lying Dirichlet eigenvalues. However, locating eigenvalues one by one by minimizing $t_m(E)$ is cumbersome, needing many evaluations of this function, each of effort $O(N^3)$, per eigenvalue found. The scaling method [46, 3, 4] for star-shaped domains is

$O(N)$ times more efficient but has no rigorous error analysis. We intend to apply tools of this paper to the scaling method.

Appendix A. Proof of analyticity of $A_j(E)$, Lemma 3.2. Each term in (3.11) is analytic in the neighborhood of E_j, but we need to show that their infinite sum is also analytic. Let $R := \frac{1}{2}\min_{m\notin[j,j+p-1]}|E_m - E_j|$ be the radius of a closed disc in the complex E plane centered at E_j. Its radius is half the distance to the nearest neighboring eigenvalue; see Figure 3.1(a). Then for each $m \notin [j, j+p-1]$ and E with $|E - E_j| \le R$, we have $|E - E_m| \ge \frac{1}{3}|R + E_j - E_m|$. Using this we have for any $f \in L^2(\partial\Omega)$ and $|E - E_j| \le R$,

$$|\langle f, A_j(E)f\rangle| = \left|\sum_{m\notin[j,j+p-1]} \frac{|\langle f,\psi_m\rangle|^2}{(E-E_m)^2}\right| \le 9\sum_{m\notin[j,j+p-1]} \frac{|\langle f,\psi_m\rangle|^2}{(R+E_j-E_m)^2}$$

$$= 9\langle f, A_j(E_j+R)f\rangle \le C\|f\|^2_{L^2(\partial\Omega)}\ , \tag{A.1}$$

since $E_j + R \notin \sigma_D$, thus A is bounded and from (3.9), A_j differs from A by a bounded operator. For each integer $N \ge j+p$, we may define the sum

$$A_j^{(N)}(E) := \left(\sum_{m=1}^{j-1} + \sum_{m=j+p}^{N}\right) \frac{\psi_m\langle\psi_m,\cdot\rangle}{(E-E_m)^2} \tag{A.2}$$

so that $A_j(E) = \lim_{N\to\infty} A_j^{(N)}(E)$. For each $f \in L^2(\partial\Omega)$ and each N, since this sum is finite and its poles lie a distance at least $2R$ from E_j, the Cauchy integral formula holds inside the disc:

$$\left\langle f, A_j^{(N)}(E)f\right\rangle = \frac{1}{2\pi i}\oint_{|z-E_j|=R} \frac{\left\langle f, A_j^{(N)}(z)f\right\rangle}{z-E}dz \qquad \text{for } |E-E_j| < R\ . \tag{A.3}$$

Taking the limit $N \to \infty$, we may apply the dominated convergence theorem to the right-hand side since the integrand is bounded by $|\langle f, A_j(z)f\rangle|/(R - |E - E_j|)$ and (A.1) holds. This gives, for each $f \in L^2(\partial\Omega)$,

$$\langle f, A_j(E)f\rangle = \frac{1}{2\pi i}\oint_{|z-E_j|=R} \frac{\langle f, A_j(z)f\rangle}{z-E}dz \qquad \text{for } |E-E_j| < R\ , \tag{A.4}$$

which proves that $\langle f, A_j(E)f\rangle$ is an analytic function of E with a radius of convergence about E_j of at least R. Using the polarization identity in complex Hilbert space

$$4\langle f, A_j g\rangle = \langle f+g, A_j(f+g)\rangle - \langle f-g, A_j(f-g)\rangle - i\langle f+ig, A_j(f+ig)\rangle + i\langle f-ig, A_j(f-ig)\rangle, \tag{A.5}$$

we see that $\langle f, A_j(E)g\rangle$ is analytic for all $f, g \in L^2(\partial\Omega)$ (weak analyticity). By [37, Theorem VI.4, p. 189], this implies $A_j(E)f$ is analytic for each $f \in L^2(\partial\Omega)$, which is the condition required to prove the lemma.

Appendix B. Proof of Cauchy interlacing, Lemma 3.4.

The nth eigenvalue of S has the max-min characterization [36, Theorem XIII.1]

$$\sigma_n = \sup_{U_1,\dots,U_{n-1}} \inf_{\substack{V\perp \mathrm{Span}\{U_1,\dots,U_{n-1}\}\\ \|V\|=1}} \langle V, SV\rangle\ , \tag{B.1}$$

where the sup is over all possible sets of $n-1$ vectors in the Hilbert space. The analogous formula for τ_n is found by replacing S with $R := S - \psi\langle\psi,\cdot\rangle$. Since for all $V \in L^2(\partial\Omega)$ it holds that $\langle V, RV\rangle = \langle V, SV\rangle - |\langle V,\psi\rangle|^2 \le \langle V, SV\rangle$, the same holds for the max-min values, hence for each $n = 1, 2, \ldots$, we have $\tau_n \le \sigma_n$. In addition, using (B.1) we have the inequalities

$$\begin{aligned}
\sigma_n &\le \sup_{U_1,\ldots,U_{n-1}} \inf_{\substack{V\perp \mathrm{Span}\{U_1,\ldots,U_{n-1},\psi\}\\ \|V\|=1}} \langle V, SV\rangle \\
&= \sup_{U_1,\ldots,U_{n-1}} \inf_{\substack{V\perp \mathrm{Span}\{U_1,\ldots,U_{n-1},\psi\}\\ \|V\|=1}} \langle V, RV\rangle \\
&\le \sup_{U_1,\ldots,U_{n}} \inf_{\substack{V\perp \mathrm{Span}\{U_1,\ldots,U_{n}\}\\ \|V\|=1}} \langle V, RV\rangle = \tau_{n+1}\ .
\end{aligned}$$

The middle equality follows since V is orthogonal to ψ; the inequalities are standard linear subspace arguments. Thus the interlacing $\tau_n \le \sigma_n \le \tau_{n+1}$ is proved.

Acknowledgments. The author is indebted to Percy Deift and Jonathan Goodman for introducing many of the ideas which made this work possible. The author also thanks Ross Barnett, Timo Betcke, Leslie Greengard, Andrew Hassell, Chris Judge, Bob Kohn, Peter Lax, Kevin Lin, Mikhail Shubin, Nick Trefethen, Eduardo Vergini, and the referees for useful input.

REFERENCES

[1] A. Bäcker, *Numerical aspects of eigenvalue and eigenfunction computations for chaotic quantum systems*, in The Mathematical Aspects of Quantum Maps, Lect. Notes Phys. 618, Springer, Berlin, 2003, pp. 91–144.

[2] A. H. Barnett and T. Betcke, *Stability and convergence of the method of fundamental solutions for Helmholtz problems on analytic domains*, J. Comput. Phys., 227 (2008), pp. 7003–7026.

[3] A. H. Barnett, *Dissipation in Deforming Chaotic Billiards*, Ph.D. thesis, Harvard University, Cambridge, MA, 2000, http://www.math.dartmouth.edu/~ahb/thesis_html/.

[4] A. H. Barnett, *Asymptotic rate of quantum ergodicity in chaotic Euclidean billiards*, Comm. Pure Appl. Math., 59 (2006), pp. 1457–88.

[5] T. Betcke and L. N. Trefethen, *Reviving the method of particular solutions*, SIAM Rev., 47 (2005), pp. 469–491.

[6] T. Betcke, *A GSVD formulation of a domain decomposition method for planar eigenvalue problems*, IMA J. Numer. Anal., 27 (2007), pp. 451–478.

[7] T. Betcke, *The generalized singular value decomposition and the method of particular solutions*, SIAM J. Sci. Comput., 30 (2008), pp. 1278–1295.

[8] J. Descloux and M. Tolley, *An accurate algorithm for computing the eigenvalues of a polygonal membrane*, Comput. Methods Appl. Mech. Engrg., 39 (1983), pp. 37–53.

[9] T. A. Driscoll, *Eigenmodes of isospectral drums*, SIAM Rev., 39 (1997), pp. 1–17.

[10] S. C. Eisenstat, *On the rate of convergence of the Bergman–Vekua method for the numerical solution of elliptic boundary value problems*, SIAM J. Numer. Anal., 11 (1974), pp. 654–680.

[11] S. C. Eisenstat, *Bounds for Eigenvalues and Eigenvectors of Self-adjoint Operators*, preprint, 2006.

[12] R. Ennenbach and H. Niemeyer, *The inclusion of Dirichlet eigenvalues with singularity functions*, Z. Angew. Math. Mech., 76 (1996), pp. 377–383.

[13] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 1998.

[14] G. Fairweather and A. Karageorghis, *The method of fundamental solutions for elliptic boundary value problems*, Adv. Comput. Math., 9 (1998), pp. 69–95.

[15] G. Fix and R. Heiberger, *An algorithm for the ill-conditioned generalized eigenvalue problem*, SIAM J. Numer. Anal., 9 (1972), pp. 78–88.

[16] L. Fox, P. Henrici, and C. Moler, *Approximations and bounds for eigenvalues of elliptic operators*, SIAM J. Numer. Anal., 4 (1967), pp. 89–102.

[17] L. Friedlander, *Some inequalities between Dirichlet and Neumann eigenvalues*, Arch. Ration. Mech. Anal., 116 (1991), pp. 153–160.

[18] P. R. Garabedian, *Partial Differential Equations*, John Wiley & Sons, New York, 1964.

[19] P. Gérard and E. Leichtnam, *Ergodic properties of eigenfunctions for the Dirichlet problem*, Duke Math. J., 71 (1993), pp. 559–607.

[20] G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd ed., Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 1996.

[21] C. Gordon, D. Webb, and S. Wolpert, *Isospectral plane domains and surfaces via Riemannian orbifolds*, Invent. Math., 110 (1992), pp. 1–22.

[22] M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics*, Interdiscip. Appl. Math., Springer, New York, 1990.

[23] A. Hassell, *personal communication*, 2005.

[24] A. Hassell and T. Tao, *Upper and lower bounds for normal derivatives of Dirichlet eigenfunctions*, Math. Res. Lett., 9 (2002), pp. 289–305.

[25] E. J. Heller, *Bound-state eigenfunctions of classically chaotic Hamiltonian systems: Scars of periodic orbits*, Phys. Rev. Lett., 53 (1984), pp. 1515–1518.

[26] E. J. Heller, *Wavepacket dynamics and quantum chaology*, in Chaos et Physique Quantique (Les Houches, 1989), North-Holland, Amsterdam, 1991, pp. 547–664.

[27] A. Karageorghis, *The method of fundamental solutions for the calculation of the eigenvalues of the Helmholtz equation*, Appl. Math. Lett., 14 (2001), pp. 837–842.

[28] J. R. Kuttler and V. G. Sigillito, *Inequalities for membrane and Stekloff eigenvalues*, J. Math. Anal. Appl., 23 (1968), pp. 148–160.

[29] J. R. Kuttler and V. G. Sigillito, *Bounding eigenvalues of elliptic operators*, SIAM J. Math. Anal., 9 (1978), pp. 768–778.

[30] J. R. Kuttler and V. G. Sigillito, *Eigenvalues of the Laplacian in two dimensions*, SIAM Rev., 26 (1984), pp. 163–193.

[31] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications, Vol.* 3, Dunod, Paris, 1970. Travaux et Recherches Mathématiques, 20.

[32] W. C. H. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, London, 2000.

[33] C. B. Moler and L. E. Payne, *Bounds for eigenvalues and eigenvectors of symmetric operators*, SIAM J. Numer. Anal., 5 (1968), pp. 64–70.

[34] A. Nachman, J. Sylvester, and G. Uhlmann, *An n-dimensional Borg-Levinson theorem*, Comm. Math. Phys., 115 (1988), pp. 595–605.

[35] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed., Cambridge University Press, Cambridge, 2002.

[36] M. Reed and B. Simon, *Methods of Modern Mathematical Physics.* IV. *Analysis of Operators*, Academic Press, New York, 1978.

[37] M. Reed and B. Simon, *Methods of Modern Mathematical Physics.* I. *Functional Analysis*, 2nd ed., Academic Press, New York, 1980.

[38] F. Rellich, *Darstellung der Eigenwerte von* $\Delta u + \lambda u = 0$ *durch ein Randintegral*, Math. Z., 46 (1940), pp. 635–636.

[39] F. Rellich, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, 1969.

[40] N. Saito, *Data analysis and representation on a general domain using eigenfunctions of Laplacian*, Appl. Comput. Harmon. Anal., 25 (2008), pp. 68–97.

[41] N. L. Schryer, *Constructive approximation of solutions to linear elliptic boundary value problems*, SIAM J. Numer. Anal., 9 (1972), pp. 546–572.

[42] G. Still, *Computable bounds for eigenvalues and eigenfunctions of elliptic differential operators*, Numer. Math., 54 (1988), pp. 201–223.

[43] L. N. Trefethen and T. Betcke, *Computed eigenmodes of planar regions*, Contemp. Math., 412 (2006), pp. 297–314.

[44] L. N. Trefethen, *Spectral Methods in MATLAB*, Vol. 10, Software, Environments, and Tools, SIAM, Philadelphia, PA, 2000.

[45] H. E. Tureci, H. G. L. Schwefel, P. Jacquod, and A. D. Stone, *Modes of wave-chaotic dielectric resonators*, Prog. Optics, 47 (2005), pp. 75–137.

[46] E. Vergini and M. Saraceno, *Calculation by scaling of highly excited states of billiards*, Phys. Rev. E, 52 (1995), pp. 2204–2207.

[47] S. Zelditch, *Quantum ergodicity and mixing of eigenfunctions*, in Elsevier Encyclopedia of Mathematical Physics, Vol. 1, Academic Press/Elsevier Science, Oxford, 2006, pp. 183–196, arXiv:math-ph/0503026.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational Physics 227 (2008) 7003–7026
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

Stability and convergence of the method of fundamental solutions for Helmholtz problems on analytic domains

A.H. Barnett [a,*], T. Betcke [b]

[a] Department of Mathematics, 6188 Kemeny Hall, Dartmouth College, Hanover, NH 03755, USA
[b] School of Mathematics, The University of Manchester, Manchester M13 9PL, UK

Received 26 August 2007; received in revised form 21 March 2008; accepted 3 April 2008
Available online 20 April 2008

Abstract

The method of fundamental solutions (MFS) is a popular tool to solve Laplace and Helmholtz boundary value problems. Its main drawback is that it often leads to ill-conditioned systems of equations. In this paper, we investigate for the interior Helmholtz problem on analytic domains how the singularities (charge points) of the MFS basis functions have to be chosen such that approximate solutions can be represented by the MFS basis in a numerically stable way. For Helmholtz problems on the unit disc we give a full analysis which includes the high frequency (short wavelength) limit. For more difficult and nonconvex domains such as crescents we demonstrate how the right choice of charge points is connected to how far into the complex plane the solution of the boundary value problem can be analytically continued, which in turn depends on both domain shape and boundary data. Using this we develop a recipe for locating charge points which allows us to reach error norms of typically 10^{-11} on a wide variety of analytic domains. At high frequencies of order only 3 points per wavelength are needed, which compares very favorably to boundary integral methods.
© 2008 Elsevier Inc. All rights reserved.

MSC: 65N12; 65N35; 78M25

Keywords: Helmholtz equation; Boundary value problem; Method of fundamental solutions; Analytic continuation; High frequency waves

1. Introduction

The method of fundamental solutions (MFS), also known as the charge simulation method or the method of auxiliary sources, is a well known method for solving Laplace or Helmholtz boundary value problems

* Corresponding author. Tel.: +1 603 646 3178; fax: +1 603 646 1312.
E-mail addresses: ahb@math.dartmouth.edu (A.H. Barnett), timo.betcke@manchester.ac.uk (T. Betcke).
URLs: http://www.math.dartmouth.edu/~ahb (A.H. Barnett), http://www.maths.man.ac.uk/~tbetcke (T. Betcke).

0021-9991/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2008.04.008

(BVPs). The idea is to approximate the solution by fundamental solutions of the Laplace or Helmholtz equation whose singularities lie outside the domain. Consider the boundary value problem

$$\Delta u + k^2 u = 0 \quad \text{in } \Omega, \tag{1a}$$

$$u = v \quad \text{on } \partial\Omega, \tag{1b}$$

where $\Omega \subset \mathbb{R}^2 = \mathbb{C}$ is a simply connected planar domain with analytic boundary $\partial\Omega$. This means that there exists a parameterization of $\partial\Omega$ in the form of an analytic function $Z(s)$ for $s \in [0, 2\pi]$, with $Z(0) = Z(2\pi)$. For simplicity we will often equivalently say that Ω is analytic. We call the boundary data v analytic if $v(s)$ is a 2π-periodic analytic function.

Recall that the solution is unique if and only if k^2 is not a Dirichlet eigenvalue (of the Laplacian) for the domain; physically this is a resonance effect. The idea of the MFS is to approximate u by a linear combination of fundamental solutions of the form

$$u(\mathbf{x}) \approx u^{(N)}(\mathbf{x}) = \frac{i}{4} \sum_{j=1}^{N} \alpha_j H_0^{(1)}(k|\mathbf{x} - \mathbf{y}_j|), \quad \mathbf{y}_j \in \mathbb{R}^2 \setminus \overline{\Omega}, \tag{2}$$

where $H_0^{(1)} = J_0 + iY_0$ is a Hankel function of the first kind of order zero, J_0 is a Bessel function of the first kind of order zero and Y_0 is a Bessel function of the second-kind of order zero. N is the number of approximating functions each of which is associated with a charge point $\mathbf{y}_j$. It is well known that $H_0^{(1)}$ satisfies the Helmholtz equation in $\mathbb{C} \setminus \{0\}$ with a singularity at zero. It is common to choose charge points lying on a smooth curve; we then may interpret the MFS as a discretization of the following external single layer potential representation of u. Let Γ be a closed curve enclosing $\overline{\Omega}$ such that $\operatorname{dist}(\Gamma, \partial\Omega) := \min\{|\mathbf{x} - \mathbf{y}|, \mathbf{x} \in \partial\Omega, \mathbf{y} \in \Gamma\} > 0$, then given a density $g \in L^1(\Gamma)$ we may write

$$u(\mathbf{x}) \approx \frac{i}{4} \int_\Gamma H_0^{(1)}(k|\mathbf{x} - \mathbf{s}|) g(\mathbf{s}) \mathrm{d}\mathbf{s}, \quad \mathbf{x} \in \Omega. \tag{3}$$

If $g(\mathbf{s}) = \sum_{j=1}^{N} \alpha_j \delta(\mathbf{s} - \mathbf{y}_j)$ for some point set $\{\mathbf{y}_j\} \in \Gamma$, where δ is the Dirac delta, we recover the MFS formulation (2). Note that Y_0 or $H_0^{(2)}$ may be used instead of $H_0^{(1)}$ in the MFS [10,13]; see Remark 1 below.

An overview of the history of the method and its applications, which have largely been in the engineering community, is given in [11]. The rate of convergence of the MFS for the Laplace ($k = 0$) BVP was investigated in [4,15–18]. It turns out that if the boundary data is analytic one can achieve exponential convergence for the MFS for the Laplace problem on analytic domains if the charge points $\mathbf{y}_j$ are suitably chosen. The MFS differs from the common approach of discretizing boundary integral equations (BIE) [5] in that, since Γ is separated from $\partial\Omega$, there is no jump relation nor singularity of the kernel. Thus one advantage over BIE is that the solution may be simply and accurately evaluated up to the boundary. There appear to be few performance comparisons of MFS against BIE in the literature [2]; our goal in this work is not to undertake such a task, but we will show that MFS may be competitive and hence deserves analysis.

One of the main drawbacks of MFS is that (unlike with second-kind BIE) systems of equations or linear least-squares problems have to be solved that are ill-conditioned. This feature is shared by Radial Basis Functions [9]. The effects of this ill-conditioning on the quality of the MFS solution have been investigated for the Laplace case in [20,21]. In this paper, we investigate more closely for the Helmholtz problem the conditions on the charge points $\mathbf{y}_j$ which lead to a numerically stable representation of an approximate solution. It turns out that this depends on how far into the complex plane a solution of (1a) can be analytically continued. The importance of this in the context of scattering problems has already been observed [23,24,8].

Our work also has consequences for the efficient numerical solution of related more challenging and widely applicable PDE problems. We have in mind (i) finding eigenmodes of the Laplace operator in Ω with homogeneous boundary conditions (where the MFS has been used at low [10] and very high eigenvalue [3]) and (ii) scattering of time-harmonic waves (the exterior Helmholtz boundary value problem in $\mathbb{R}^2 \setminus \overline{\Omega}$). In both these situations the boundary data is almost always analytic: in problem (i) it is zero and in (ii) a plane wave or point source.

We will study convergence of the MFS approximation in the boundary error norm

$$t = \|u^{(N)} - v\|_{L^2(\partial\Omega)}. \tag{4}$$

By applying [22, Eq. (7)], this controls the interior error of the solution as follows:

$$\|u^{(N)} - u\|_{L^2(\Omega)} \leqslant \frac{C_\Omega}{d}\|u^{(N)} - v\|_{L^2(\partial\Omega)}, \tag{5}$$

where $d := \min_j |k^2 - E_j|/E_j$, the domain's Dirichlet eigenvalues are E_j, and C_Ω is a domain-dependent constant. This shows that for any fixed nonresonant k, we may use the boundary norm.

In Section 2 we give rigorous results for the convergence and the numerical stability of the MFS for Helmholtz problems on the unit disc, with analytic boundary data, using charge points on a concentric circle. We then present a heuristic model for behavior in finite-precision arithmetic and show it explains well numerical results observed at both low and high wavenumbers. A key conclusion will be that it is the growth in norm of the coefficient vector that in practice limits the achievable error, so this norm should be kept as small as possible to retain high accuracy. The reader should take care throughout not to confuse statements about the *coefficient norm* (which depending on the choice of MFS charge points may either grow or not grow with N as the error converges to zero), with statements about the *condition number* of the problem (which always grows with N since the MFS (2) approximates a single layer operator (3) which is compact).

In Section 3.1 we move to general analytic domains, and review results for the analytic continuation of solutions u of (1a), in particular how both the boundary data and the domain shape may lead to singularities in the continuation of u. Exterior conformal maps play an important theoretical role in the analysis of the MFS for Laplace problems [17], thus in Section 3.2 we explore the use of such maps to choose charge points for Helmholtz problems in several challenging domains. We propose and provide evidence for conjectures in general domains analogous to our theorems on convergence rate and stability in the unit disc. For nonconvex domains charge points determined by exterior conformal maps can lead to bad results. This motivates a heuristic conformally related method proposed in Section 3.3, which circumvents this problem and in Section 3.4 is shown to be an efficient representation in the high frequency (large k) limit.

2. The MFS on the unit disc

In this section, we analyse the accuracy and coefficient sizes that result in the unit disc $\Omega = \{\mathbf{x} : |\mathbf{x}| < 1\}$, for the MFS using charge points $\mathbf{y}_j = Re^{i\phi_j}$, $j = 1, 2, \ldots, N$, with $\phi_j = 2\pi j/N$, that is, equally spaced on a larger circle of radius $R > 1$ (see Fig. 1).

Before we embark we need to define the Fourier series of a function $g \in L^2([0, 2\pi])$,

$$g(\theta) = \sum_{m=-\infty}^{\infty} \hat{g}(m)e^{im\theta}, \quad \hat{g}(m) = \frac{1}{2\pi}\int_0^{2\pi} g(\theta)e^{-im\theta}\,d\theta. \tag{6}$$

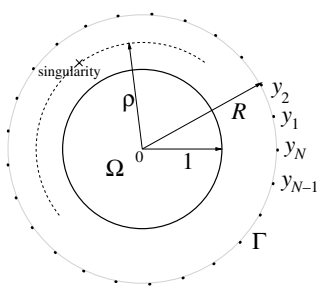


Fig. 1. Geometry for the MFS in the unit disc.

Parseval's identity is then

$$\|g\|^2_{L^2([0,2\pi])} = 2\pi \sum_{m=-\infty}^{\infty} |\hat{g}(m)|^2 =: 2\pi \|\hat{g}\|^2_{\ell^2(\mathbb{Z})}. \tag{7}$$

We also need to represent the coefficient vector $\boldsymbol{\alpha} := \{\alpha_j\}_{j=1,\ldots,N}$ in a discrete Fourier basis labeled by $-N/2 < k \leqslant N/2$ (we will always choose N even),

$$\alpha_j = \sum_{k=-N/2+1}^{N/2} \hat{\alpha}_k e^{ik\phi_j}, \quad \hat{\alpha}_k = \frac{1}{N}\sum_{j=1}^{N} \alpha_j e^{-ik\phi_j}, \tag{8}$$

where inversion follows from $\sum_{j=1}^{N} e^{2\pi ikj/N} = N\delta_{k0}^{(N)}$, with $\delta_{k0}^{(N)}$ the periodized Kronecker delta defined by

$$\delta_{kj}^{(N)} = \begin{cases} 1, & k \equiv j (\mathrm{mod}\, N) \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Parseval's identity now gives $|\boldsymbol{\alpha}|^2 = N|\hat{\boldsymbol{\alpha}}|^2$, where $|\boldsymbol{\alpha}| := (|\alpha_1|^2 + \cdots + |\alpha_n|^2)^{1/2}$ is the standard Euclidean norm.

2.1. Map from layer potential to Fourier basis on the unit circle

For simplicity we first consider the layer potential version of the problem, which can be interpreted as the $N \to \infty$ limit of the MFS. The single layer potential lying on the outer circle $\Gamma = \{\mathbf{y} : |\mathbf{y}| = R\}$ is

$$u(\mathbf{x}) = \frac{i}{4}\int_0^{2\pi} H_0^{(1)}(k|\mathbf{x} - Re^{i\phi}|)g(\phi)d\phi. \tag{10}$$

Note $g \in L^1([0, 2\pi])$ is the density with respect to angle measure $d\phi$ rather than the usual length measure $Rd\phi$. The Fourier–Bessel decomposition of a fundamental solution located at $Re^{i\phi}$, evaluated at $\mathbf{x} = Re^{i\theta}$ is, using Graf's addition formula [1, Eq. (9.1.79)],

$$\frac{i}{4}H_0^{(1)}(k|\mathbf{x} - Re^{i\phi}|) = \frac{i}{4}\sum_{m\in\mathbb{Z}} H_m^{(1)}(kR)J_m(kr)\cos m(\theta - \phi) = \frac{i}{4}\sum_{m\in\mathbb{Z}} H_m^{(1)}(kR)e^{-im\phi}\cdot J_m(kr)e^{im\theta} \tag{11}$$

where the second step involved the reflection formulae [1, Eq. (9.1.5)] $J_{-m}(z) = (-1)^m J_m(z)$ and $Y_{-m}(z) = (-1)^m Y_m(z)$. Hence the Fourier–Bessel coefficients are $\frac{i}{4}H_m^{(1)}(kR)e^{-im\phi}$. The restriction of (10) to $\mathbf{x} \in \partial\Omega$ gives a single layer operator $S : L^2([0, 2\pi]) \to L^2([0, 2\pi])$ of convolution type, which is therefore diagonal in the Fourier basis $\{e^{im\theta}\}_{m\in\mathbb{Z}}$ and entirely described by its eigenvalues. Comparing (10) and (11) and using orthogonality gives $\hat{u}(m) = \hat{s}(m)\hat{g}(m)$ where the eigenvalues of S are

$$\hat{s}(m) = \frac{i\pi}{2}H_m^{(1)}(kR)J_m(k). \tag{12}$$

Remark 1. Since the Hankel function (of real argument) is never zero, an eigenvalue can vanish only when $J_m(k) = 0$, corresponding to a Dirichlet eigenvalue (resonance) of Ω. In contrast if Y_0 were chosen as the fundamental solution in (2), $Y_m(kR)$ may accidentally be very small giving poor or spurious numerical results, although in practice this happens rarely [10]. In general one may avoid this problem by using $(Y_0 + i\eta J_0)$ for any η with Re $\eta \neq 0$ (an analogous idea is used in layer potentials [5, p. 48]).

Since its kernel is continuous S is compact, so $\lim_{|m|\to\infty}\hat{s}(m) = 0$, S^{-1} is unbounded, and we expect arbitrarily large $\|g\|$ will be needed to represent certain unit-norm boundary functions v.

In the Laplace ($k \to 0$) limit we recover the following known result (e.g. [16, Eq (3.2)]). We use the small-argument asymptotics $J_m(k) \sim (k/2)^m/\Gamma(m+1)$ and $Y_m(kR) \sim -\frac{1}{\pi}\Gamma(m)(kR/2)^{-m}$ for all integer $m > 0$, and the reflection formulae, and get

$$\hat{s}(m) \to \frac{1}{2|m|}R^{-|m|}, \quad m \in \mathbb{Z}\setminus\{0\}, \quad k \to 0. \tag{13}$$

To analyze convergence rate we will need the asymptotic behavior as $|m| \to \infty$ for fixed k. Using in (12) the large-order asymptotics (9.3.1 in [1]) $J_m(z) \sim \frac{1}{\sqrt{2\pi m}}(ez/2m)^m$ and $Y_m(z) \sim -\sqrt{\frac{2}{\pi m}}(ez/2m)^{-m}$, where z is fixed, gives the leading-order behavior

$$\hat{s}(m) \sim \frac{1}{2|m|}R^{-|m|}, \quad |m| \to \infty, \tag{14}$$

which coincides with the Laplace case (13). We therefore have the exponential uniform bounds, for some constants c_s and C_s depending only on R and k,

$$\frac{c_s}{|m|}R^{-|m|} \leqslant |\hat{s}(m)| \leqslant \frac{C_s}{|m|}R^{-|m|} \leqslant C_s R^{-|m|}, \quad m \in \mathbb{Z} \setminus \{0\}, \tag{15}$$

where C_s is chosen large enough such that $|\hat{s}(0)| \leqslant C_s$ also holds. Fig. 2a shows that for low wavenumber k the leading-order asymptotic is reached rapidly, hence the tightest possible ratio C_s/c_s is not too large (here it is about 10^3). Thus (14) becomes accurate well before the dynamic range begins to exceed machine double precision $\epsilon_{\text{mach}} \approx 10^{-16}$. However, the situation can differ radically for large k, as Fig. 2b illustrates. Here the Laplace asymptotic is not relevant for the eigenvalues within a factor ϵ_{mach} of the maximum. Worse still, the ratio C_s/c_s must be exceedingly large (several tens of orders of magnitude). We will present more useful asymptotic approximations for the eigenvalues in Section 2.4. For now we will use only (15) to prove exponential convergence rates.

2.2. Map from MFS coefficients to Fourier basis

We adapt the above to the discrete source case. Define the density

$$g(\phi) = \sum_{j=1}^{N} \alpha_j \delta(\phi - \phi_j). \tag{16}$$

It follows that

$$u^{(N)}(e^{i\phi}) = \frac{i}{4}\sum_{j=1}^{N} \alpha_j H_0^{(1)}(k|e^{i\phi} - Re^{i\phi_j}|) = (Sg)(\phi).$$

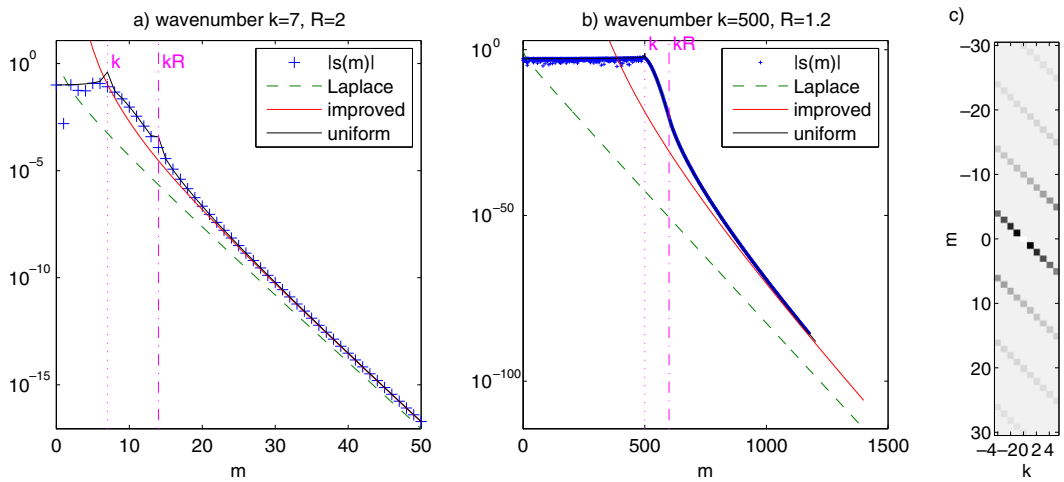


Fig. 2. Comparison for the unit disc of layer potential eigenvalue magnitudes $|\hat{s}(m)|$ given in (12) against various asymptotic expressions: 'Laplace' (14), 'improved' (39), and 'uniform' (40). (a) low wavenumber, (b) high wavenumber. Panel (c) shows density plot of matrix elements (20) of Q for $N = 10$, in the domain $|m| \leqslant 30$. In (c) we chose unrealistically small values of R and k in order to make the super- and sub-diagonals more visible.

We have

$$\|Sg\|^2_{L^2([0,2\pi])} = 2\pi\|\widehat{Sg}\|^2_{\ell^2(\mathbb{Z})} = 2\pi\sum_{m=-\infty}^{\infty}|\hat{s}(m)\hat{g}(m)|^2 = \frac{N^2}{2\pi}\sum_{m=-\infty}^{\infty}|\hat{s}(m)\hat{\alpha}_{m\bmod N}|^2,$$

where $m\bmod N$ denotes the unique integer lying in the range $-N/2+1,\ldots,N/2$ which differs from m by an integer multiple of N. The last equality follows from the Fourier series representation of (16),

$$\hat{g}(m) = \frac{1}{2\pi}\sum_{j=1}^{N}\alpha_j e^{-\mathrm{i}m\phi_j} = \frac{N}{2\pi}\hat{\alpha}_{m\bmod N}, \quad m\in\mathbb{Z}. \tag{17}$$

Applying Hölder's inequality and (15) we obtain

$$\|Sg\|^2_{L^2([0,2\pi])} \leqslant \frac{N^2}{2\pi}\max_{j=-\frac{N}{2}+1,\cdots,\frac{N}{2}}|\hat{\alpha}_j|^2\cdot\sum_{m=-\infty}^{\infty}|\hat{s}(m)|^2 \leqslant \frac{N^2}{2\pi}|\hat{\boldsymbol{\alpha}}|^2 C_s^2\frac{R^2+1}{R^2-1}. \tag{18}$$

Define the operator $Q:\mathbb{R}^N\to\ell^2(\mathbb{Z})$ by $Q\hat{\boldsymbol{\alpha}} := \widehat{Sg}$, which maps the discrete Fourier coefficient vector $\hat{\boldsymbol{\alpha}}$ to the Fourier series coefficients on the boundary $\partial\Omega$. From (18) we immediately obtain the following.

Lemma 2. *For $R>1$ the operator Q is bounded, with norm*

$$\|Q\| \leqslant C_s\frac{N}{2\pi}\sqrt{\frac{R^2+1}{R^2-1}},$$

where $\|Q\| := \max_{\hat{\boldsymbol{\alpha}}\in\mathbb{R}^N\setminus\{0\}}\frac{\|Q\hat{\boldsymbol{\alpha}}\|_{\ell^2(\mathbb{Z})}}{|\hat{\boldsymbol{\alpha}}|}$.

The action of Q is multiplication by a generalized matrix of width N but (bi-)infinite height,

$$\hat{u}(m) = \sum_{k=-N/2+1}^{N/2} q_{mk}\hat{\alpha}_k, \quad \text{for } m\in\mathbb{Z}. \tag{19}$$

From $Q\hat{\boldsymbol{\alpha}}(m) = (\widehat{Sg})(m) = \hat{s}(m)\hat{g}(m)$ and (17) it follows that the matrix elements are

$$q_{mk} = \frac{N}{2\pi}\hat{s}(m)\delta^{(N)}_{mk}. \tag{20}$$

Fig. 2c shows a greyscale picture of a piece of this matrix. It is dominated by a main diagonal proportional to the diagonal of the S operator defined in (12), but with (exponentially) smaller entries on an infinite sequence of super- and sub-diagonals. This off-diagonal part can be interpreted as aliasing 'overtones' due to discrete sampling of a continuous layer potential. In the Laplace case using (13) in (20) recovers the results of Katsurada [18, Lemma 1, case 2].

2.3. Convergence rate and coefficient sizes in the disc with analytic data

We are now in a position to express the boundary error norm (4) in terms of $\hat{\boldsymbol{\alpha}}$. Combining with (7) and (19) gives

$$t[\hat{\boldsymbol{\alpha}}] = \sqrt{2\pi}\|Q\hat{\boldsymbol{\alpha}} - \hat{\mathbf{v}}\|_{\ell^2(\mathbb{Z})}, \tag{21}$$

where $\hat{\mathbf{v}}\in\ell^2(\mathbb{Z})$ is the vector of Fourier coefficients of the boundary data v on the unit circle. Assume that v can be analytically continued to the annulus $\{z\in\mathbb{C}:\frac{1}{\rho}<|z|<\rho\}$ for some $\rho>1$, that is, the closest singularity of the analytic continuation of v has the radius ρ or $1/\rho$. We then have asymptotically exponential decay of the Fourier coefficients,

$$|\hat{v}(m)| \sim C\rho^{-|m|}, \quad |m|\to\infty \tag{22}$$

for some constant C. A simple example is boundary data arising from an nth-order pole $v(z) = Re(z-\rho)^{-n}$ for $z\in\partial\Omega$, $n=1,2,\ldots$

Minimizing (21) over $\hat{\boldsymbol{\alpha}}$ is a least-squares problem involving the generalized matrix Q. But since the columns of Q are orthogonal this separates into N independent single-variable minimizations. We may use a diagonal approximation to choose $\hat{\boldsymbol{\alpha}}$ which is sufficient for the following convergence rate bounds.

Theorem 3. *Let $R > 1$ and N be even. For analytic boundary data v obeying (22), the minimum boundary error (4) achievable with the MFS in the unit disc satisfies*

$$t \leqslant \begin{cases} C\rho^{-N/2}, & \rho < R^2, \\ C\sqrt{N}R^{-N}, & \rho = R^2, \\ CR^{-N}, & \rho > R^2, \end{cases} \tag{23}$$

where each time C means a different constant which may depend on k, R, and v, but not N. Furthermore if v is analytically continuable to an entire function, the last of the three cases holds for any $R > 1$.

Proof. We choose coefficients $\hat{\alpha}_m = \hat{v}(m)/q_{mm}$ for $-N/2 < m \leqslant N/2$. This exactly matches the Fourier coefficients in this interval, therefore errors are due only to frequencies lying outside the interval. Eqs. (21) & (20) and the triangle inequality in $\ell^2(\mathbb{Z})$ give

$$t = \left(2\pi \sum_{m \notin \left[-\frac{N}{2}+1, \frac{N}{2}\right]} |(Q\hat{\boldsymbol{\alpha}})(m) - \hat{v}(m)|^2\right)^{1/2} \leqslant \sqrt{2\pi}(E_u + E_v),$$

where

$$E_u^2 = \sum_{m \notin \left[-\frac{N}{2}+1, \frac{N}{2}\right]} |(Q\hat{\boldsymbol{\alpha}})(m)|^2 = \sum_{-\frac{N}{2} < n \leqslant \frac{N}{2}} \left|\frac{\hat{v}(n)}{\hat{s}(n)}\right|^2 \sum_{b \neq 0} |\hat{s}(bN+n)|^2 \tag{24}$$

and

$$E_v^2 = \sum_{m \notin \left[-\frac{N}{2}+1, \frac{N}{2}\right]} |\hat{v}(m)|^2. \tag{25}$$

We can bound both error terms since all terms in the sums have exponential bounds. First we note that using (15) and (22) gives

$$\begin{aligned} E_u^2 &\leqslant C_1 \sum_{\substack{-\frac{N}{2} < n \leqslant \frac{N}{2} \\ n \neq 0}} \left(\frac{R}{\rho}\right)^{2|n|} |n|^2 \sum_{b \neq 0} \frac{R^{-2|bN+n|}}{|bN+n|^2} + C_2 \left|\frac{\hat{v}(0)}{\hat{s}(0)}\right|^2 \sum_{b \neq 0} \frac{R^{-2|bN|}}{|bN|^2} \\ &\leqslant C_1 \sum_{\substack{-\frac{N}{2} < n \leqslant \frac{N}{2} \\ n \neq 0}} \left(\frac{R}{\rho}\right)^{2|n|} \sum_{b \neq 0} R^{-2|bN+n|} + C_2 \left|\frac{\hat{v}(0)}{\hat{s}(0)}\right|^2 \sum_{b \neq 0} R^{-2|bN|} \end{aligned} \tag{26}$$

for sufficiently large constants C_1 and C_2. We can bound

$$\sum_{b \neq 0} R^{-2|bN+n|} \leqslant C_3 R^{-2N+2|n|}, \quad -\frac{N}{2} < n \leqslant \frac{N}{2} \tag{27}$$

for a large enough constant C_3. Inserting (27) into (26) and absorbing $|\frac{\hat{v}(0)}{\hat{s}(0)}|^2$ into the constants gives

$$E_u^2 \leqslant CR^{-2N} \sum_{-\frac{N}{2} < n \leqslant \frac{N}{2}} \left(\frac{R^2}{\rho}\right)^{2|n|} \tag{28}$$

for a sufficiently large constant C. Similarly $E_v^2 \leqslant C\rho^{-N}$ follows from (22). We now study the sum in (28). For $\rho < R^2$, (28) can be estimated by $C\rho^{-N}$ for some constant $C > 0$. This means both E_u^2 and E_v^2 have the same exponential decay $C\rho^{-N}$. For $\rho = R^2$, the sum contains N equal terms and therefore $E_u^2 \leqslant CNR^{-2N}$, which

decays slower than the bound $C\rho^{-N}$ on E_v^2. For $\rho > R^2$, (28) can be estimated by CR^{-2N} for some $C > 0$, which means E_v is of higher negative order in N than E_u and can be dropped. For the case of v continuable to an entire function, we may take $\rho \to \infty$ and the case $\rho > R^2$ applies. □

This is a generalization of a result of Katsurada [15] from the Laplace to the Helmholtz problem. Since only the exponential bounds (15) and no other information about $\hat{s}(m)$ was used, the convergence rates are identical to those for Laplace with the same boundary data.

Remark 4. An interpretation of the two main convergence rate regimes is:

- v is 'not relatively smooth' ($R > \sqrt{\rho}$, i.e. 'distant' charge points): errors are limited by the absence of Fourier modes beyond a frequency $N/2$ in the MFS basis, hence rate is controlled by the boundary data singularity ρ.
- v is 'relatively smooth' ($R < \sqrt{\rho}$, i.e. 'close' charge points): errors are limited by aliasing errors due to the discrete representation of the single layer potential, hence rate is controlled by R.

Remark 5. By keeping track of the constants in the above proof, one can check that C is at least as big as C_s/c_s, which, as discussed, must be very large for large wavenumbers k. Thus although the convergence rates in the theorem are reached asymptotically as $N \to \infty$ in exact arithmetic, we cannot expect the bounds to be numerically useful in practice at large k.

We are interested in how the coefficient norm $|\boldsymbol{\alpha}|$ grows as we reduce the boundary error in the MFS representation (2) by increasing N. Firstly, we now show that when the MFS charge points are closer than the nearest singularity, the coefficients need not grow.

Theorem 6. *Let Ω be the unit disc, and $R < \rho$, with fixed analytic boundary data v obeying (22). Then as $N \to \infty$ there exists a sequence of coefficient vectors $\boldsymbol{\alpha}$ with bounded norm $|\boldsymbol{\alpha}|$, with corresponding boundary error norm (21) converging as in Theorem 3.*

Proof. We choose coefficients as in Proof of Theorem 7, which therefore give the desired convergence rate. Then using (22), (20) and (14),

$$\hat{\alpha}_m = \frac{\hat{v}(m)}{q_{mm}} = \frac{2\pi}{N}\frac{\hat{v}(m)}{\hat{s}(m)} \sim C\frac{|m|}{N}\left(\frac{R}{\rho}\right)^{|m|} \leqslant \frac{C}{2}\left(\frac{R}{\rho}\right)^{|m|}, \quad -\frac{N}{2} < m \leqslant \frac{N}{2}, \quad m \neq 0 \tag{29}$$

for some constant C. For $R < \rho$ this is an exponentially decaying sequence so, independent of N, $|\boldsymbol{\alpha}|$ is bounded by a constant. □

If $\rho < R$, the coefficient choice used in the above two proofs would cause $|\boldsymbol{\alpha}|$ to diverge exponentially with N. It is not immediately obvious whether there is a different choice of $\hat{\alpha}_m$ which avoids such growth. The next result, proved in Appendix A, excludes this possibility by showing that when the singularity in the analytic continuation of the boundary data is closer than the MFS source points, *any* sequence of coefficient vectors $\boldsymbol{\alpha}$ with the desired error norm convergence of Theorem 3 must blow up exponentially.

Theorem 7. *Let Ω be the unit disc, with $R > \rho$. Let the boundary data Fourier coefficients decay no faster than (22), that is, for some constant c_v,*

$$|\hat{v}(m)| \geqslant c_v \rho^{-|m|}. \tag{30}$$

For any positive even N satisfying $N > 3 + N_{\min}$, where $N_{\min} := 2\max\left[\ln\left(\sqrt{\frac{2}{\pi}}\frac{c}{c_v}\right)\Big/\ln\rho, 1\right]$, let $\boldsymbol{\alpha}$ be a coefficient vector such that the MFS representation (2) has a boundary error norm (21) satisfying

$$t \leqslant c\rho^{-N/2}, \tag{31}$$

where c is a constant independent of N. Then

$$|\boldsymbol{\alpha}| \geqslant C\sqrt{N}\left(\frac{R}{\rho}\right)^{N/2} \tag{32}$$

for some constant C which may depend on k, R, and v, but not N.

For MFS approximation in the disc, the immediate consequence is that if ρ is known for given boundary data, to prevent exponential coefficient growth one must restrict the charge point radius to $R \leqslant \rho$. We illustrate this and Theorem 3 with numerical experiments at low wavenumber using the following standard implementation. The integral in (4) is approximated using uniform quadrature with M equally spaced boundary points $\{\mathbf{x}_m\}_{m=1,\ldots,M}$. Specifically we fill a matrix $A \in \mathbb{C}^{M\times N}$ with elements

$$A_{mj} := \frac{i}{4}H_0^{(1)}(k|\mathbf{x}_m - \mathbf{y}_j|), \tag{33}$$

and a boundary value vector $\mathbf{v} \in \mathbb{C}^M$ with elements $v_m := v(\mathbf{x}_m)$, then solve the linear system $A\boldsymbol{\alpha} = \mathbf{v}$. Choosing $M > N$ this is a least-squares problem, and for M sufficiently large the quadrature of the boundary error norm gives $t \approx \sqrt{|\partial\Omega|/M}\,|A\boldsymbol{\alpha} - \mathbf{v}|$. We always choose M large enough that this value of t, and the resulting solution $u^{(N)}$, has converged with respect to M; in practice (on analytic domains) we find this usually requires $M \approx 1.5N$. For the solution of the dense least-squares problem we used the QR decomposition in double precision arithmetic via MATLAB's backslash (`mldivide`) command.

In Fig. 3 we show convergence of t using boundary data

$$v(z) = -\frac{1}{4}Y_0(k|z-\rho|), \quad z \in \partial\Omega \tag{34}$$

with real $\rho > 1$, that is, a single real valued fundamental solution. The three panels illustrate the three cases of Theorem 3. In (a) $\rho > R^2$ thus convergence is determined by R. In (b) $R^2 > \rho > R$ so we have transitioned to a convergence rate given by ρ. In both these cases the coefficient size $|\boldsymbol{\alpha}|$ is very close to constant (note by contrast that the condition number of A, not shown, grows). However in (c) $\rho < R$ so convergence rate is again determined by ρ, but now $|\boldsymbol{\alpha}|$ grows exponentially at precisely the rate indicated by Theorem 7. Fig. 4 shows the resulting approximate field $u^{(N)}$ in the case (c), and the error function $u^{(N)} - u$. Notice that the error is oscillatory at Fourier frequencies of about $N/2$ (as predicted by Remark 4; this can be seen by comparing the alternating signs in Fig. 4b to the angular spacing of source points), is concentrated on the side of $\partial\Omega$ nearest the

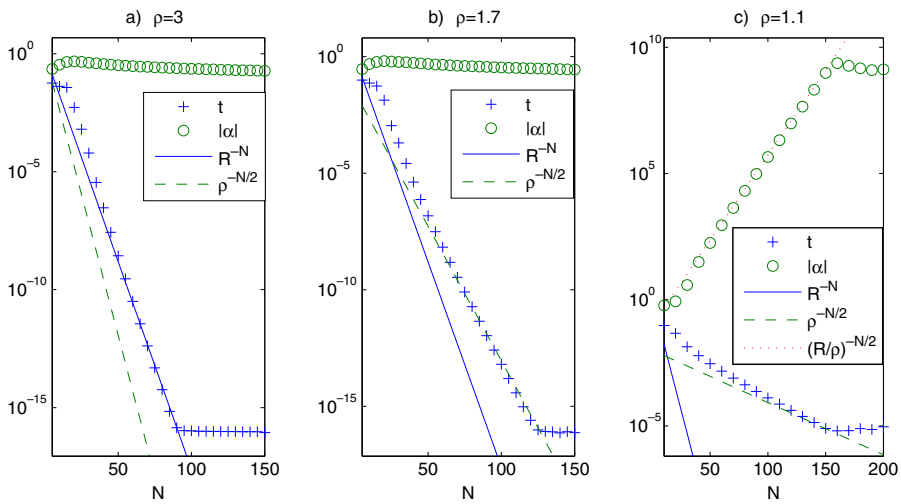


Fig. 3. Convergence and coefficient sizes as a function of N, for the MFS approximation to the interior Helmholtz BVP in the unit disc given boundary data corresponding to a single source (34) at radius ρ. The wavenumber is low ($k = 8$). The MFS sources are at $R = 1.5$. For visual comparison the relevant power laws from Theorems 3 and 7 are shown (sometimes the constants have been chosen to match the data). There were $M = 240$ quadrature points.

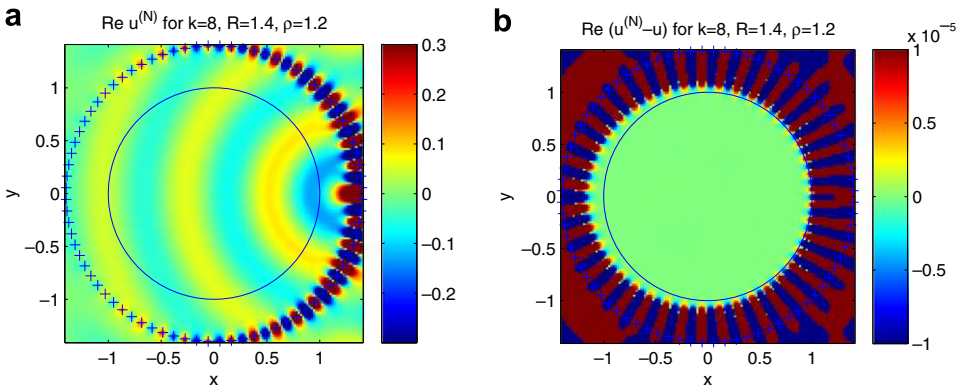


Fig. 4. MFS approximation to the interior Helmholtz BVP in the disc with given boundary data corresponding to a single source (34) outside the domain at $\rho = 1.2$, with $N = 80$ and $R = 1.4$. (a) Re $u^{(N)}$, b) residual error Re $(u^{(N)} - u)$ (note color scale is 3×10^4 times more sensitive than in a).

singularity, and decays exponentially inside the domain (it is evanescent). We have also substituted $v(z) = Re(z - \rho)^{-1}$ and find the convergence rates of Fig. 3 are very similar. We note that in each plot in this figure, the convergence eventually stops, as we now explain.

2.4. Minimum achievable error in the disc for low and high wavenumbers

So far we have proven results which hold in exact arithmetic. However machine precision limits the dynamic range of relevant eigenvalues of S: since the MFS trial functions in (2) have size $O(1)$ in Ω (for any reasonable k), each coefficient α_j will result in round-off errors of size roughly $\epsilon_{\text{mach}}\alpha_j$ in the numerical approximation $u^{(N)}$. Thus we expect convergence to stop when t reaches of order ϵ_{mach} times the coefficient norm $|\boldsymbol{\alpha}|$; this is well illustrated by Fig. 3. In panel (c) the coefficient growth thus limits achievable error norm to only about 10^{-5}. This convergence halting problem has been observed in the Laplace case in the disc [15,19,32] but not analyzed much before. We now analyse this in the Helmholtz case, and first note the following guideline, a consequence of Theorems 3 and 6.

Remark 8. For all k, for boundary data with nearest singularity radius ρ, a choice of charge point radius R in the range $\sqrt{\rho} < R < \rho$ gives optimal asymptotic convergence rate $t \sim \rho^{-N/2}$ and coefficient norm $|\boldsymbol{\alpha}| = \mathrm{O}(1)$.

We now construct a useful heuristic model which predicts, for general ρ and R in the unit disc, both the lowest achievable error norm and the basis size N required to achieve it. We first consider low k, that is, where (14) applies for the relevant eigenvalues (those exceeding $\epsilon_{\text{mach}}\max_m|\hat{s}(m)|$). We discuss both the case $R < \rho$ with $O(1)$ coefficients and the case $R > \rho$ when the coefficient norm grows. Proof of Theorem 7 suggests that, at least when $\hat{s}(m)$ is exponentially decaying, the diagonal approximation for the MFS Fourier coefficients (see (29))

$$\hat{\alpha}_m \approx \frac{2\pi}{N}\frac{\hat{v}(m)}{\hat{s}(m)} \tag{35}$$

approximates the true least-squares Fourier coefficients well, apart from an O(1) number of them lying at extreme frequencies near $m = \pm N/2$. For $R > \rho$ these Fourier coefficients grow exponentially, so dropping an O(1) factor we may consider only the largest coefficient's contribution to the l^2-norm, namely that at $m = N/2$. For $R < \rho$ there is no growth hence the norm is dominated by the coefficients of size O(1) at $m \approx 0$. Combining the two cases, we get the order-of-magnitude estimate at a given N

$$|\boldsymbol{\alpha}| \approx \max\left[\frac{1}{\sqrt{N}}\frac{|\hat{v}(N/2)|}{|\hat{s}(N/2)|}, 1\right]. \tag{36}$$

Similarly, since the boundary data coefficients die exponentially, following Remark 4 and ignoring an $O(1)$ factor we may suppose:

$$t \approx |\hat{v}(N/2)|. \tag{37}$$

We define N_0 to be the N at which convergence stops: for this we use the round-off error consideration $t/|\boldsymbol{\alpha}| \approx \epsilon_{\text{mach}}$ discussed above. In the case $R < \rho$ this implies that convergence halts when t reaches of order ϵ_{mach}, as observed in Fig. 3a and b. However for $R > \rho$, combining (36) and (37) we get an implicit equation for N_0

$$\sqrt{N_0}|\hat{s}(N_0/2)| \approx \epsilon_{\text{mach}} \quad \text{(criterion for halting of convergence, } R > \rho). \tag{38}$$

The minimum achievable boundary error is then given by (37) with the substitution $N = N_0$. As an illustration, using the (Laplace) asymptotic form (14) for the eigenvalues predicts (dropping algebraic factors) that $N_0 \approx 2\ln(1/\epsilon_{\text{mach}})/\ln R$. The parameters of Fig. 3c give $N_0 \approx 180$, then using (22) with $C = 1$ gives $t_0 \approx 10^{-4}$ which, given the heuristic nature of our model, agree well with the observed behavior.

We now briefly discuss the case of high wavenumber. In Fig. 2 we saw that (14) is useless for predicting relevant eigenvalues at high k. We may derive (Appendix B) the asymptotic

$$\hat{s}(m) \sim \frac{1}{2|m|} R^{-|m|} \mathrm{e}^{k^2(R^2-1)/4m}, \quad |m| \to \infty, \tag{39}$$

which Fig. 2a and b shows is improved, but still not useful for the relevant eigenvalues at $k = 500$ (or beyond). Therefore we use the WKBJ method to derive (see Appendix B) a uniform approximation for the eigenvalue magnitudes, defining $a^2 = m^2 - \frac{1}{4}$,

$$|\hat{s}(m)| \approx \begin{cases} [(k^2 - a^2)(k^2R^2 - a^2)]^{-1/4}, & m < k, \\ \frac{1}{2}[(a^2 - k^2)(k^2R^2 - a^2)]^{-1/4} \mathrm{e}^{I_a(k)}, & k < m < kR, \\ \frac{1}{2}[(a^2 - k^2)(a^2 - k^2R^2)]^{-1/4} \mathrm{e}^{I_a(k) - I_a(kR)}, & m > kR, \end{cases} \tag{40}$$

where

$$I_a(x) := \sqrt{a^2 - x^2} - a \ln\left[\left(a + \sqrt{a^2 - x^2}\right)/x\right]. \tag{41}$$

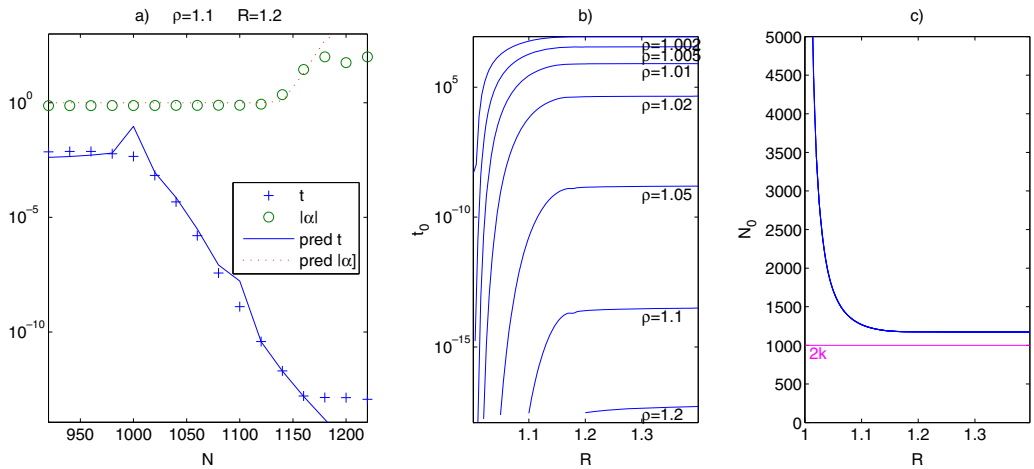


Fig. 5. (a) Convergence and coefficient sizes as a function of N, for MFS approximation of the interior Helmholtz BVP in the disc at high wavenumber $k = 500$, with boundary data (34) with singularity radius $\rho = 1.1$, and MFS sources at $R = 1.2$. The 'predicted' curves are given by (37) using (36), with (40) estimating both $\hat{s}(m)$ and $\hat{v}(m)$. (b) Minimum achievable boundary error norms t_0, and (c) corresponding basis sizes N_0, predicted for a selection of source radii ρ and MFS source point radius R, for $k = 500$, using the model in Section 2.4. Note that in (c) all the graphs for different ρ lie on top of one another.

More precisely this is an estimate of *amplitude* in the oscillatory region ($m < k$) of J_m, and of absolute value in the evanescent region (note $|H_m^{(1)}|$ is never oscillatory). Fig. 2 shows this is a highly accurate asymptotic form in all regions apart from weak algebraic singularities at the two turning points ($J_m(k)$ is at its turning point for $m \approx k$ whereas for $H_m^{(1)}(kR)$ this occurs at $m \approx kR$).

To understand MFS performance at high k in the disc we use the above low-k model (38) with (14) replaced by (40). Fig. 5a shows convergence and coefficient norm at $k = 500$ for boundary data (34), compared with the model predictions. Note that for this boundary data, $\hat{v}(m)$ is given by the right-hand side of (12) with the substitution of ρ for R, which we estimate via (40) with this same substitution. It is clear that, up to the point when convergence halts, the predictions for both error norm and coefficient norm are very close to observations (the largest deviations being spikes due to algebraic singularities discussed above, at $N = 1000$ and 1100).

A crucial feature is that no convergence happens until $N > 2k$, since $\hat{s}(m)$ remains large for $|m| < k$ (see Fig. 2b). One interpretation of this is that $2k$, corresponding to 2 degrees of freedom per wavelength on the perimeter, is the Nyquist sampling frequency for k-bandlimited functions on $\partial\Omega$; in physics this is called the 'semiclassical basis size' [3]. In Fig. 5a, $\rho = 1.1$ so the singularity is $k(\rho - 1)/2\pi \approx 8$ wavelengths from the boundary. In this case convergence is rapid, dropping ten orders of magnitude between $N = 1000$ and $N = 1150$. Convergence then halts, in agreement with (38) which predicts $N_0 \approx 1175$ and $t_0 \approx 3 \times 10^{-14}$ (we find similar agreement for other R and ρ choices). The number of quadrature points was $M = 1500$, or three points per wavelength.

How can R best be chosen to achieve the lowest boundary error for a given high wavenumber k, and ρ? We use the above model to compute N_0 and hence t_0 for a variety of ρ and R at $k = 500$, in Fig. 5b and c. Here the smallest $\rho = 1.002$ corresponds to a singularity only 0.16 wavelengths from the boundary. Two conclusions are clear: (i) If the nearest singularity in the boundary data is at least a few wavelengths away (e.g. $\rho \geqslant 1.1$), then the basis size N need only slightly exceed 2 per wavelength in order to achieve $t_0 \approx \epsilon_{\text{mach}}$. (ii) Fixing ρ, one may achieve $t_0 \approx \epsilon_{\text{mach}}$ as R tends to ρ from above, as expected from Remark 8; however, the basis size required to converge to this error diverges as $\rho \to 1^+$.

3. The MFS on analytic domains

In this section we present results for the MFS on arbitrary analytic domains. On the unit disk we have shown that the MFS coefficients start growing exponentially if the radius R of the charge points becomes larger than the distance ρ of the singularity. The story will be similar for general analytic domains, but with the following twist: singularities may now be due to both boundary data and domain shape. The MFS will be numerically investigated in detail using charge curves defined by exterior conformal maps, which are theoretically important (see Katsurada [17]) and provide a natural generalization of the unit disk case. However, for non-convex domains conformal maps are not always a good choice; we motivate a new method for placing the charge points that instead provides excellent performance. We end with an example where performance exceeds that of BIE. However, to achieve such promising results, an understanding of the singularities is needed, as we now discuss.

3.1. Analytic continuation of solutions

The question of analytic continuation is to find a domain $\tilde{\Omega} \supset \Omega$ and a function $\tilde{u}$ such that $\Delta\tilde{u} + \lambda\tilde{u} = 0$ in $\tilde{\Omega}$ and $\tilde{u}|_\Omega = u$. Since solutions of the Helmholtz equation are real analytic it follows immediately that $\tilde{u}$ is unique.

A classical result of analytic continuation is reflection on a straight arc Γ, on which u satisfies $u|_\Gamma = 0$. Without restriction let Γ be a subset of $\{iy : y \in \mathbb{R}\}$. Then u can be continued across Γ by setting $u(-x, y) := -u(x, y)$ (see also [6]). In [12], Garabedian extended these results to the case that Γ is an arbitrary analytic arc for which $u_\Gamma = 0$. More general reflection principles for linear elliptic PDEs of the type $\Delta u + a(x, y)u_x + b(x, y)u_y + c(x, y) = 0$, where $a(x, y)$, $b(x, y)$ and $c(x, y)$ are real analytic functions were treated by Lewy in [26]. He stated his results for arbitrary Dirichlet, Neumann and mixed boundary conditions but restricted Γ to be a straight line. Representations of the analytic continuation for the case that Γ is not a

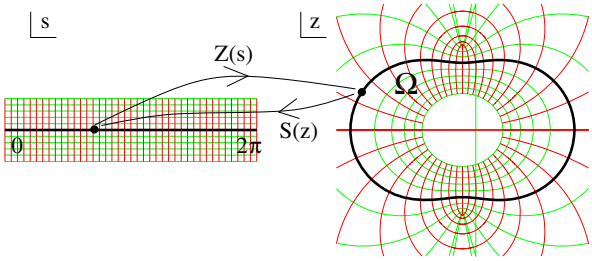


Fig. 6. Illustration of map $Z(s)$ and its inverse $S(z)$ defining a boundary curve $\partial\Omega$. The Schwarz function involves composition of $Z(s)$ and $S(z)$.

straight line were given by Millar in [28]. In [29], he discussed more in detail the analytic continuation of solutions of the Helmholtz equation.

Millar shows that there are two possible sources for singularities of the analytic continuation $\tilde{u}$ of u. The first one comes from singularities of the analytic continuation of the boundary data f. The second possible source of singularities is introduced by the shape of $\partial\Omega$. Let $Z(s) = x(s) + iy(s)$ be a parameterization of $\partial\Omega$, where $s \in [0, 2\pi]$. Assume that $x(s)$ and $y(s)$ are real analytic and that $|Z'(s)| \neq 0$ in $[0, 2\pi]$. Then there exists a complex neighborhood of $[0, 2\pi]$, in which $Z(s)$ is holomorphic and invertible. We denote its inverse by $S(z)$ and define the Schwarz function

$$G(z) := \overline{Z}(S(z)) = \overline{Z(\overline{S(z)})}.$$

Millar showed that except for special cases the singularities of $G(z)$ outside Ω are also singularities of the analytic continuation $\tilde{u}$ of u.

The Schwarz function has been studied in [7]. It is independent of the parameterization of $\partial\Omega$ and has an interpretation in terms of reflection principles on analytic arcs. Assume that z_1 is a point close to $\partial\Omega$. Then its reflection on $\partial\Omega$ can be obtained by the following steps (see Fig. 6):

(1) Compute $t_1 = S(z_1)$.
(2) Reflect t_1 on the real line to obtain the point $t_2 := \overline{t_1}$.
(3) The reflection z_2 of z_1 at $\partial\Omega$ is now obtained as

$$z_2 = Z(t_2) = Z(\overline{t_1}) = Z(\overline{S(z_1)}) = \overline{G(z_1)}. \tag{42}$$

Fig. 7 shows the singularities of the Schwarz function on three different domains, a rounded triangle, an inverted ellipse and a crescent. The domains are defined with default values for the parameters a_1 through a_4, as follows:

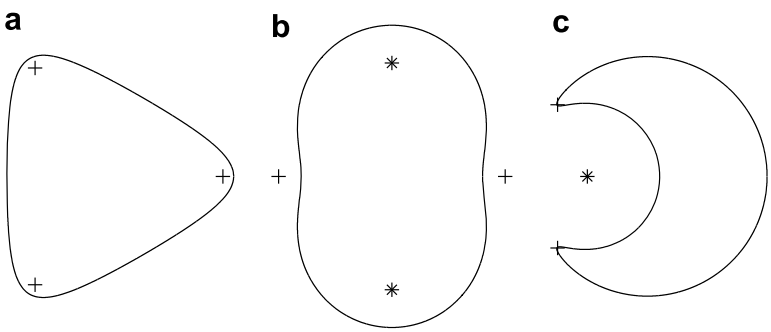


Fig. 7. Domains (left to right): (a) rounded triangle, (b) inverted ellipse, (c) crescent. Branch singularities of the Schwarz function are denoted by + and pole type singularites by ∗. Note in (c) the + signs are inside Ω but very close to the boundary.

$$\text{Rounded triangle}: \quad Z_T(s) = e^{is} + a_1 e^{-2is}, \quad a_1 = 0.2$$

$$\text{Inverted ellipse}: \quad Z_{IE}(s) = \frac{e^{is}}{1 + a_2 e^{2is}}, \quad a_2 = 0.25$$

$$\text{Crescent}: \quad Z_C(s) = e^{is} - \frac{a_3}{e^{is} + a_4}, \quad a_3 = 0.1, a_4 = 0.9$$

The branch singularities in all three domains are of square root type (see [29] for an analysis of the branch behavior of $G(z)$). The crescent has exterior singularity of pole type at $z = -1/\overline{a_4}$. For the interior Helmholtz problem only the exterior singularities of G are important since these are points where, for generic boundary data, the analytic continuation $\tilde{u}$ of u becomes singular. Conversely if we had an exterior Helmholtz problem then the interior singularities would determine the singularities of the analytic continuation.

3.2. Using exterior conformal map to place the charge points

A natural generalization of the MFS on the unit disk to general analytic domains can be defined in terms of the conformal map from the exterior of the unit disk to the exterior of the domain. This was investigated in the Laplace BVP case by Katsurada [17].

Let Ω be a simply connected domain with analytic boundary $\partial\Omega$. We can parameterize $\partial\Omega$ using the exterior conformal mapping function

$$z = \Psi(w) = cw + c_0 + \frac{c_{-1}}{w} + \frac{c_{-2}}{w^2} + \cdots, \quad c > 0,$$

which maps the exterior $D_1 := \{w : |w| > 1\}$ of the unit disk to the exterior of Ω. The quantity c is called the capacity of Ω. We denote the inverse map by $w = \Phi(z)$. Since $\partial\Omega$ is analytic $\Psi(w)$ can be analytically continued to a domain $D_r := \{w : |w| > r\}$ for some $0 < r < 1$. We denote the conformal radius of a point $z \in \mathbb{C} \setminus \Omega$ by $\rho_z := |\Phi(z)|$.

In the notation of the previous section we may write this parametrization as $Z(s) = \Psi(e^{is})$ since the unit disc is parametrized by $w = e^{is}$. Using that the reflection of a point z on the unit circle is given by $z' = \frac{1}{\bar{z}}$ the Schwarz function may now be written $G(z) = \overline{\Psi(1/\overline{\Phi(z)})}$; it follows that it is analytic in $\{z \in \mathbb{C} : 1 < \rho_z < \frac{1}{r}\}$.

In the unit disk case we placed the MFS points equally distributed on a curve with radius R. For general analytic domains we now place the points on a curve $\Gamma_R := \{z : \rho_z = R\}$ with constant conformal radius $\rho_z = R$. On this curve we distribute the points equally spaced in conformal angle, that is

$$\mathbf{y}_j := \Psi(e^{2\pi i j/N}), \quad j = 1, \ldots, N. \tag{43}$$

If Ω is the unit disk this definition coincides with that of Section 2. Replacing the disk radii R and ρ in Theorem 3 by the corresponding conformal radii we obtain the following conjecture for the rate of convergence of the MFS for Helmholtz problems on general analytic domains.

Conjecture 9. *Let t be the error of the MFS as defined in (4) by placing the MFS points equally distributed in conformal angle at a conformal distance R around Ω. Let $\rho > 1$ be the conformal radius of the closest (in the sense of conformal radius) singularity of the analytic continuation of u. Then*

$$t \leqslant \begin{cases} C\rho^{-N/2}, & \rho < R^2, \\ CR^{-N}, & \rho > R^2, \end{cases} \tag{44}$$

where C is a constant that may depend on Ω, k, R and v, but not N. Furthermore, if u continues to an entire function, the latter case holds for any $R > 1$.

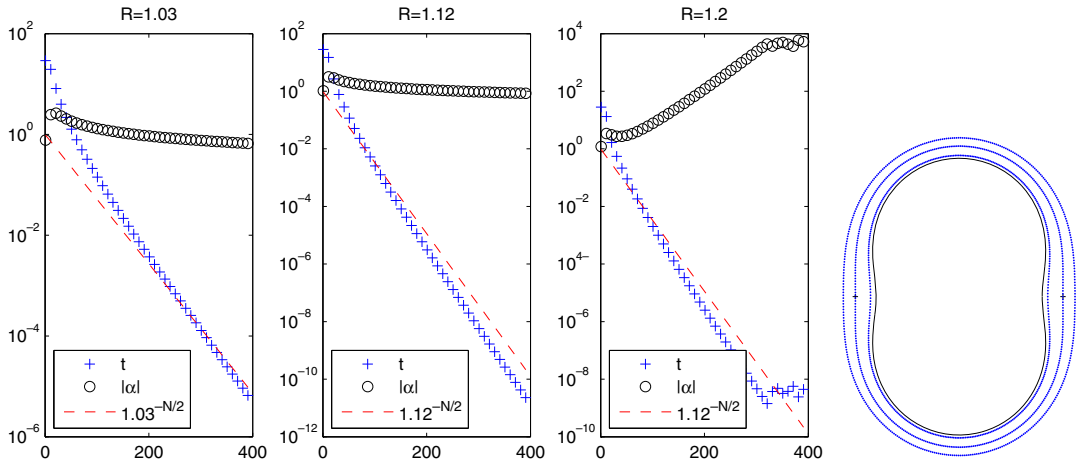


Fig. 8. Left three plots: Estimated (dashed lines) and observed ('+') rates of convergence of the MFS on the inverted ellipse for charge points with conformal radii $R = 1.03$, $R = 1.12$ and $R = 1.2$. The corresponding coefficient norms $|\boldsymbol{\alpha}|$ are denoted by 'o'. Right plot: Domain and positions of the charge points (shown as dots) for the above three conformal radii, with branch-type singularity locations (+ signs).

Remark 10. The first case ($\rho < R^2$) of this conjecture was proved in the Laplace case by Katsurada in [17] under additional restrictions on the analytic continuation of Ψ into the unit disk. In numerical studies we have observed that these conditions are not necessary to achieve the given convergence rates, so do not include them in our conjecture (compare also Remark 3.2 of [17]). We do not state a conjecture for the case $\rho = R^2$ since numerically it cannot be established if for general domains the same algebraic factor is needed as for the disk.

In Fig. 8 we plot the observed error t for the MFS on the inverted ellipse of Fig. 7, for wavenumber $k = 5$ and constant boundary condition $v \equiv 1$. The estimated rates from Conjecture 9 are denoted by dashed lines. The three plots correspond to MFS points placed at the conformal distances $R = 1.03$, $R = 1.12$ and $R = 1.2$. The conformal radius $R = 1.12$ is also the approximate conformal radius ρ of the two singularities. The corresponding MFS curves are shown on the right of Fig. 8. The estimated convergence rates are in all three cases in good agreement with the observed error t.

In the disk case we can observe exponential coefficient growth once the radius R of the charge points is larger than the radius ρ of the singularity of the analytic continuation of u (see Theorem 7). For general analytic domains we observe a similar behavior. To demonstrate this we plotted in Fig. 8 also the norm $|\boldsymbol{\alpha}|$ of the MFS coefficients. As long as the conformal radius R of the MFS points is smaller or equal to the conformal radius ρ of the singularities we do not observe any growth of $|\boldsymbol{\alpha}|$ for growing N (first two plots of Fig. 8). In the third plot we have $R > \rho$ and $|\boldsymbol{\alpha}|$ grows exponentially for growing N. It is instructive to compare this figure panel by panel against Fig. 3.

This leads to the following conjecture, which mirrors Theorems 6 and 7.

Conjecture 11. *Let Γ be any Jordan curve enclosing $\overline{\Omega}$, with* $\operatorname{dist}(\Gamma, \partial\Omega) > 0$, *on which MFS charge points are chosen asymptotically densely. Then the coefficient norm $|\boldsymbol{\alpha}|$ that minimizes t grows asymptotically exponentially as $N \to \infty$ if and only if Γ encloses a singularity of the analytic continuation of u.*

We stated this conjecture for general Jordan curves, since numerical experiments suggest that there is only coefficient growth if the curve encloses a singularity of the analytic continuation. This conjecture is analogous to (but stronger than) a result that coefficients grow if all singularities are *not* enclosed by Γ in the context of exterior scattering [8, Chapter 4, Theorem 2.4].

In Fig. 9 we show the coefficient norm $|\boldsymbol{\alpha}|$ and the approximation error t for a growing conformal distance R of the MFS source points and fixed number N in four different cases: the unit disk and the three domains from Fig. 7. In all cases we have used $k = 5$. For the disk the boundary data is given by (34) with singularity location $\rho = 1.2$, and in the other three cases by $v(z) \equiv 1$ (recall that here the Schwarz function introduces singularities in u). The

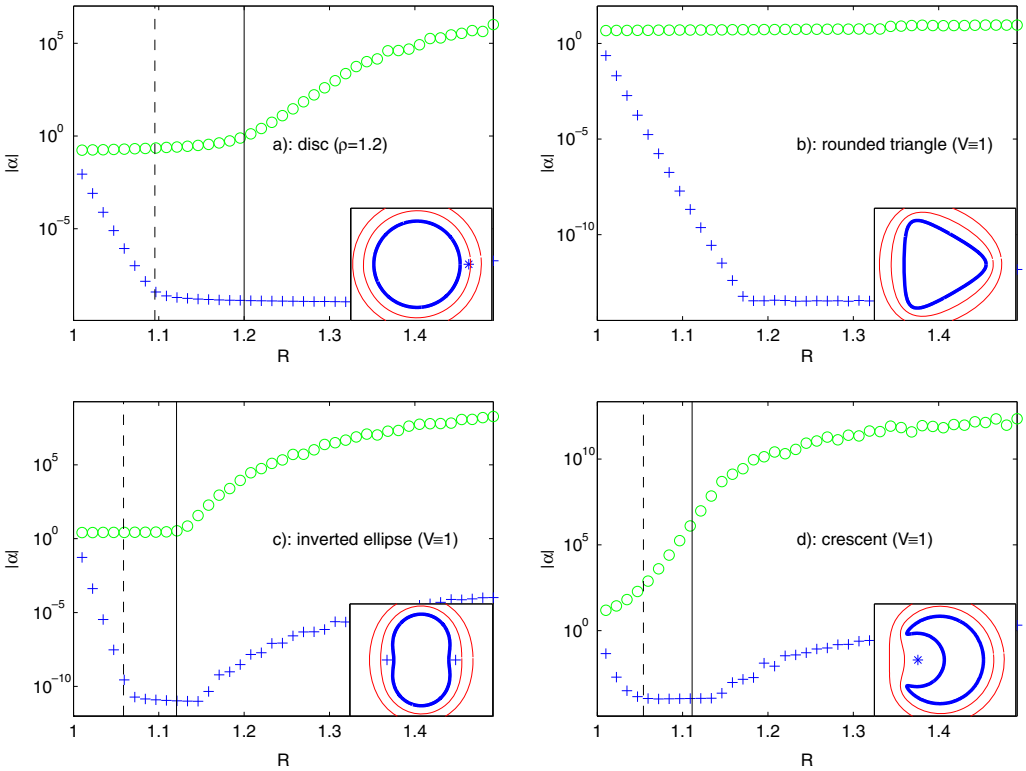


Fig. 9. Growth of coefficient norm $|\boldsymbol{\alpha}|$ (circles), and least-squares approximation error t ('+' signs), for (a) the unit disc with boundary data (34), and (b–d) the three other shapes of Fig. 7 with constant boundary data $v \equiv 1$. MFS points are chosen on curves with constant conformal radius R. The insets show domains, exterior singularities using same notation as Fig. 7, and curves with conformal radius $R = 1.24$ and 1.49. The parameters are $k = 5$. $N = 200$ for a and b, $N = 400$ for c and d. The number of quadrature points M on $\partial\Omega$ is chosen sufficiently large throughout.

vertical solid lines denote the conformal radius ρ of the singularities of the analytic continuation of u and the vertical dashed lines denote the square root $\rho^{1/2}$. Since the Schwarz function for the rounded triangle does not have any singularities in the exterior of the domain the solution u can be analytically continued to an entire function.

For the disk, the inverted ellipse and the crescent the error t does not decrease further once R passes the dashed line. This can be expected from Conjecture 9 since the upper bound on the error t does not decrease any more for fixed N and $R > \rho^{1/2}$.

For these three domains we can also observe exponential coefficient growth of $|\boldsymbol{\alpha}|$ for fixed N when $R > \rho$. For the crescent this exponential growth already starts earlier. However, this is not a contradiction to Conjecture 11. The conjecture treats the case of a fixed curve and $N \to \infty$. This does not exclude the existence of transient growth effects if the singularity lies outside the curve containing the charge points.

An explanation for these transient effects in the crescent case is that close to the pole-like singularity of the Schwarz function we need a very high number N of basis functions to sufficiently resolve a highly oscillatory Helmholtz field.

Another interesting special case is the rounded triangle. Since the analytic continuation of u is an entire function, by Conjecture 11 we do not expect any exponential growth of $|\boldsymbol{\alpha}|$. Indeed, Fig. 9b shows that $|\boldsymbol{\alpha}|$ stays virtually constant as R increases.

3.3. Using a singularity-adapted curve to place the charge points

The above exterior conformal method has the following problem: in any concave parts of Ω the exterior conformal map Ψ has a very (in fact, exponentially) large gradient. This well-known property of conformal maps is related to the so-called crowding problem. This has two consequences for concave regions: the spacing

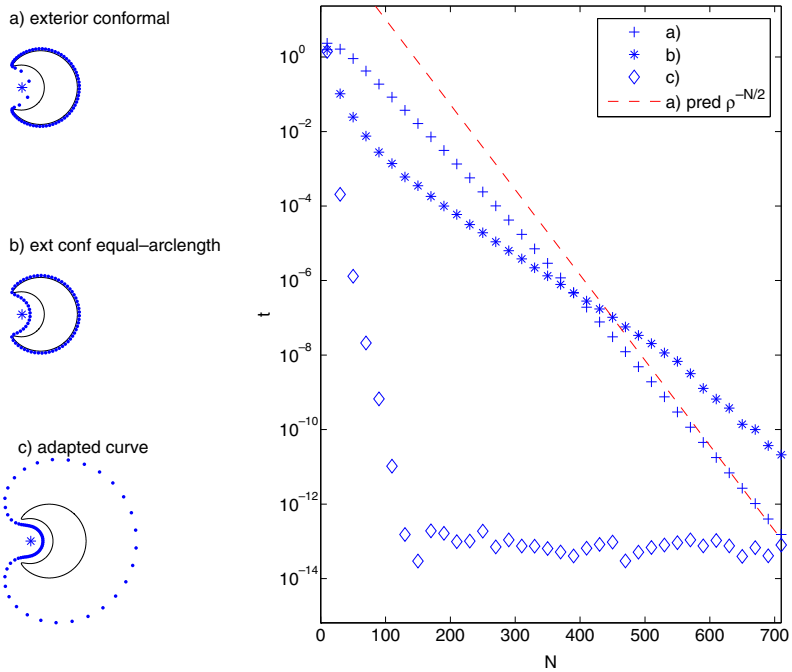


Fig. 10. Convergence rate for the crescent domain (with $a_3 = 0.1, a_4 = 0.9$) illustrating the concavity problem. The wavenumber is $k = 3$ and constant boundary data $v \equiv 1$. For (a) and (b) MFS charge points are placed on an exterior conformal curve at $R = \sqrt{\rho}$ where $\rho = 1/a_4$ is the singularity conformal distance. The point spacing is equal in (a) conformal angle, and (b) arc-length. (c) Adapted curve and spacing given by (45), see Section 3.3. The dashed line shows the predicted convergence rate for (a), a_4^N.

of charge points according to (43) becomes very large, and as R is increased from 1, the curve Γ_R moves away from $\partial\Omega$ very rapidly. Both these effects are illustrated by the MFS charge curve for the crescent in Fig. 10a.

This means that Schwarz function singularities which are a moderate distance from a concave part of $\partial\Omega$ may actually have a conformal radius extremely close to 1. The net result is that if the coefficient growth of Conjecture 11 is to be prevented, R must be very close to 1, hence by Conjecture 9 the convergence rate is necessarily very poor. For example, in Fig. 10, the conformal radius of the pole in the Schwarz function is only $\rho = 1/a_4 \approx 1.11$, and the observed rate for case a) approaches the predicted $\rho^{-N/2}$ (dashed line in the figure). One can attempt to fix the problem of the large point spacing by retaining the same MFS curve Γ_R but choosing charge points equally spaced in arc-length, as illustrated in Fig. 10b. Despite an initial improvement for small N, the asymptotic convergence rate turns out to be no better than in case a), and is believed to be the same (for errors $t < 10^{-6}$ it performs worse, we believe due to a lack of point density near the 'spiked' parts of the crescent).

In Fig. 10c we show a different 'adaptive' choice of MFS curve and point spacing which clusters charge points near the singularity but spreads them out (while taking them further away from $\partial\Omega$) away from the singularity. This gives a convergence rate over five times faster than the exterior conformal curve, for instance an error of $t \approx 10^{-13}$ is reached for $N = 140$ as opposed to $N = 730$ for case a). This simple adaptive method is based on the idea of replacing the exterior conformal map by an annular one, as follows.

From the discussion in Section 3.1, the map $Z(s)$ defines an annular conformal map which is one-to-one for s in some strip around $[0, 2\pi]$, as in Fig. 6. The external singularities control convergence rate; we label them by $\sigma = 1, \ldots, P$, where P is the number of singularities. They have s-plane locations $s_\sigma = \chi_\sigma - \mathrm{i}\tau_\sigma$ with $\tau_\sigma > 0$, Their minimum distance to the real axis is $\tau := \min_\sigma \tau_\sigma$. Katsurada et al. [19] have discussed using such an annular map to place charge points for the Laplace BVP, according to $\mathbf{y}_j := Z(2\pi j/N - \mathrm{i}\log R)$ for some $R > 1$; note this is the annular map equivalent of (43). A related annular map method has been tried in the scattering case [14]. According to Conjecture 11 in order to prevent coefficient growth in this case one would need to choose $\log R < \tau$. This may be a severe restriction: for example one may check that the crescent formula $Z_C(s)$ in Table 3.1 is identical to this domain's exterior conformal map Ψ, thus the concavity effect causes τ to be very small.

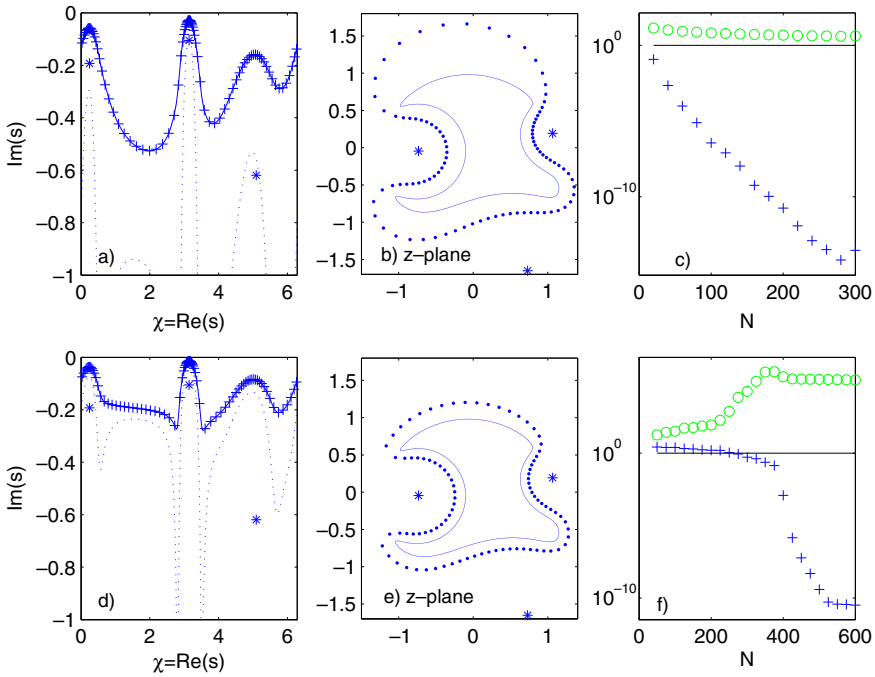


Fig. 11. Generalized crescent analytic domain parametrized by (47), showing performance of adaptive charge points of Section 3.3. The top row is at $k = 3$, the bottom row at $k = 100$. (a) and (d) show the curve $y(\chi)$ and the locations $(\chi_j, y(\chi_j))$ in the s-plane, the singularities s_σ ('$*$' symbols), and the distance-limiting function $|Z\prime(\chi)|/D_{\max}$ (dotted line). (b) and (e) show the charge locations (for clarity, $N = 90$ has been used in both cases). (c) and (f) shows error norm convergence t ('+' symbols) and coefficient norm $|\boldsymbol{\alpha}|$ ('o' symbols).

However, there are an infinite family of annular maps $Z(s)$ which parametrize the *same* boundary $\partial\Omega$, and these may have differing Schwarz singularity locations in the s-plane. Given an analytic domain Ω one is free to choose between such parametrizations. Ideally, our goal is to choose one with as large a τ as possible, to achieve a high convergence rate. However, we find it convenient to retain the given parametrization $Z(s)$, and instead build a charge curve in the s-plane which no longer has constant imaginary part, and which captures the spirit of such a reparametrization. Our curve is given by $s = \chi - \mathrm{i}y(\chi)$, where $\chi \in [0, 2\pi]$ is the real part of s, and the function y is given by

$$[y(\chi)]^{-1} = \left(\frac{D_{\max}}{|Z'(\chi)|}\right)^{-1} + \sum_{\sigma=1}^{P}\left[\gamma\tau_\sigma + \beta\frac{1-\cos(\chi-\chi_\sigma)}{\tau_\sigma}\right]^{-1}, \tag{45}$$

where parameter values performing well in most domains are $\beta = 0.7$ (interpreted as a curvature factor), $\gamma = 0.4$ (expressing the curve's fractional distance to each singularity), and

$$D_{\max} = \max[1, 25/k] \tag{46}$$

is interpreted as the maximum allowable distance of the curve from $\partial\Omega$. Eq. (45) has the effect of bringing the curve close to $\partial\Omega$ in the vicinity of each singularity (via each cos term in the sum) in the style of an equipotential curve, while allowing it to move up to $D_{\max}$ from the boundary in the absence of nearby singularities. Given the curve function $y(\chi)$ a set of N real values $\{\chi_j\} \in [0, 2\pi]$ are then chosen such that their local spacing is proportional to $y(\chi)$.[1] The MFS charge points are then given by $\mathbf{y}_j = Z(\chi_j - iy(\chi_j))$.

In Fig. 11, we illustrate the performance of this method on a generalization of the crescent domain given by

[1] In practice this can easily be done numerically by solving the ODE $g'(\chi) = 1/y(\chi)$ then spline fitting to construct an approximate inverse function g^{-1}.

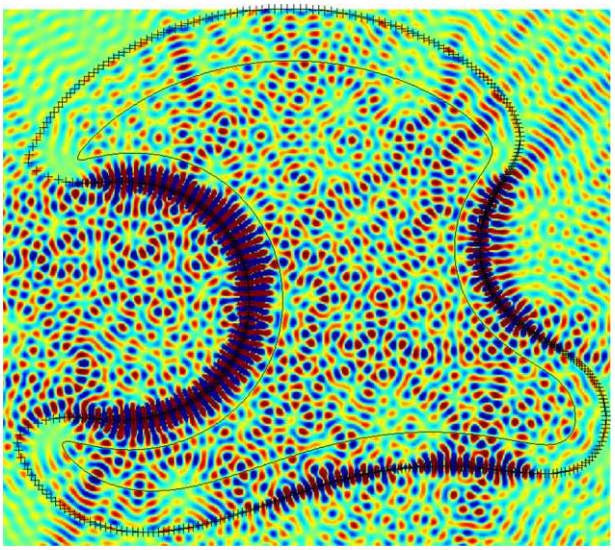


Fig. 12. Interior Helmholtz BVP solution u for the same domain as Fig. 11, at $k = 100$, with boundary data $v \equiv 1$, using as basis size of $N = 525$, and $M = 1000$ boundary points. Error norm is $t = 5 \times 10^{-11}$. The boundary (thin solid line) and charge points ('+' symbols) are shown, and the field u from (2) is shown both outside and inside Ω. CPU time was 2.9 s to compute the coefficient vector $\boldsymbol{\alpha}$, and 5.4 min to evaluate the solution u shown (2.4×10^5 points on a grid of spacing 0.005).

$$\text{Generalized crescent}: \quad Z_{GC}(s) = \mathrm{e}^{\mathrm{i}s} - \frac{0.1}{\mathrm{e}^{\mathrm{i}s} + a_5} - \frac{0.07 + 0.02\mathrm{i}}{\mathrm{e}^{\mathrm{i}s} + a_6} + \frac{0.2}{\mathrm{e}^{\mathrm{i}s} + a_7} \tag{47}$$

with $a_5 = 0.9$, $a_6 = -0.8 - 0.2\mathrm{i}$, $a_7 = -0.2 + 0.5\mathrm{i}$. The Schwarz function has three exterior pole type singularities, shown by * symbols, with s-plane locations $1/\overline{a_5}$, $1/\overline{a_6}$, and $1/\overline{a_7}$. The two rows of subfigures contrast the effect of $D_{\max}$ given by (46) at low vs high wavenumbers. For low wavenumber $D_{\max}$ is large and the contribution of the first term in (45) is small. This enables the curve to reach large negative Im s, hence large distances from $\partial\Omega$ and a large point spacing, in regions away from singularities. We observe rapid convergence, reaching $t = 10^{-14}$ by $N = 280$, and no exponential coefficient growth. As k increases, $D_{\max}$ drops and the first term in (45) starts to become significant (see dotted line in Fig. 11d), and has the effect of bringing the curve closer to the domain. As $k \to \infty$ this term dominates and the curve tends to a constant distance of about 4 wavelengths from $\partial\Omega$ everywhere on the curve. The bottom row in the figure shows the case $k = 100$. Once $N = 400$ (about 3.0 degrees of freedom per wavelength on the boundary), convergence is rapid, reaching 10^{-10} at $N = 525$. Since the coefficient norm $|\boldsymbol{\alpha}| \approx 10^5$ convergence halts here. The resulting solution u is shown in Fig. 12.[2] Values outside Ω have been included to highlight the manner in which the MFS points generate the field (very large coefficients are easily noticed due to the dark, highly oscillatory bands in these parts of Γ).

3.4. Performance of singularity-adapted MFS at high wavenumber

Finally, we test the performance of the above method for different shape at higher wavenumber, as shown in Fig. 13. This domain is challenging since it has, very close to the boundary, five exterior branch-type singularities in the Schwarz function. We also choose nonconstant boundary data which itself has a singularity (however since it is outside Γ there is no need to include its contribution in the curve formula (45)). There are about 165 wavelengths across the domain's maximum diameter. Boundary error t of order 10^{-10} is reached using $N = 1900$ corresponding to 3.3 basis functions per wavelength on the perimeter. Since $|\boldsymbol{\alpha}| \approx 10^4$, this

[2] Computation times are quoted for a single core of a 2 GHz Intel Core Duo laptop CPU, with 2 GB RAM, running GNU/Linux, coded in MATLAB. For evaluation of u the sum (2) was performed naively. The MATLAB Hankel function routine is also by no means optimal, requiring on average 2.3 μs per evaluation.

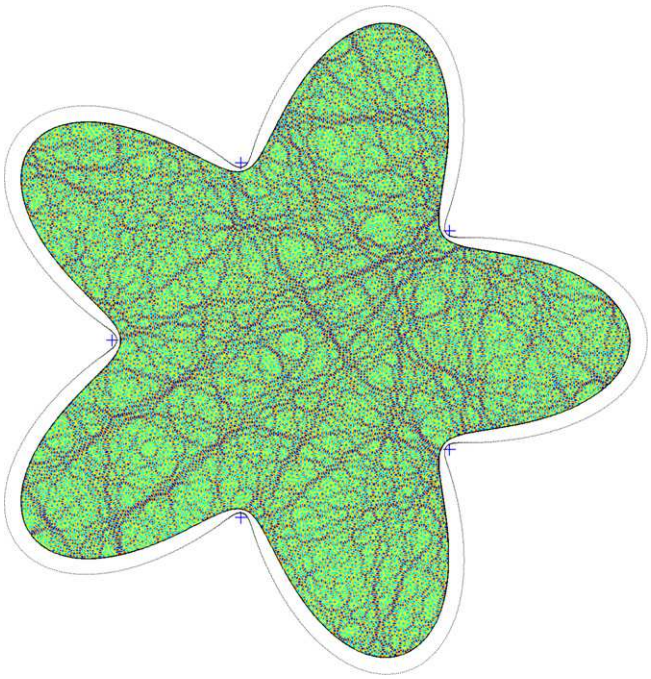


Fig. 13. Interior Helmholtz BVP solution u for the analytic star domain defined by radial function $r(\theta) = 1 + 0.3\cos(5\theta)$ at wavenumber $k = 400$, boundary data $v(z) = Re(z-\rho)^{-1}$ and $\rho = 1 + 0.5\mathrm{i}$. The color scale shows a u range $[-45, 45]$. Boundary error norm is $t = 3.6 \times 10^{-10}$, and interior values agree to 10^{-11} relative error against a BIE computation. Basis size was $N = 1900$ and $M = 2800$ quadrature points. Charge points are shown by dots and Schwarz branch singularities by + symbols. CPU time was 55 s to compute the coefficient vector $\boldsymbol{\alpha}$, and 1.0 h to evaluate the interior solution u shown (8.2×10^5 points on a grid of spacing 0.002).

value of t cannot be decreased much by increasing N. At fixed N, an adequate number M of quadrature points was found by increasing until interior values of the solution converged (seemingly spectrally) to the required number of digits.

We have verified that a pointwise error of order 10^{-12} times typical u values is achieved in the interior, by comparison against a BIE solution implemented with Martensen–Kussmaul spectral quadrature of the double layer potential, as recommended in [5, Section 3.5]. For instance, with the above parameters the MFS value of $u^{(N)}(0.045 + 0.1\mathrm{i})$ is $-50.189873472429 - 1.5 \times 10^{-11}\mathrm{i}$ compared to for BIE $-50.189873472414 + 5 \times 10^{-12}\mathrm{i}$. Note that the exact solution is real, and resonance makes its size roughly 10^2 times the $\mathrm{O}(1)$ boundary data. This MFS error is much better than the bound (5), which gives relative error of order 10^{-7} since d (defined below (5)) is very small, of order 10^{-5}. We note that to achieve the above the BIE required a larger linear system, $N = 3200$ Nyström quadrature points (these were equally spaced in θ which may not be optimal), and required 91 s of CPU time to compute the layer density coefficients, as opposed to 55 s for MFS. This serves to illustrate the potential MFS has for high frequency problems.

In domains with more distant singularities, we find barely more than 2 basis functions per wavelength are sufficient in the high k limit, similar to what was found for the disc in Section 2.4. We postpone further study of important issues such as choice of $D_{\max}$ at high wavenumber, and more detailed comparison with BIE, to future work.

4. Conclusions

The Method of Fundamental Solutions is a powerful tool for solving the Helmholtz BVP, but, as we have demonstrated, the achievable accuracy is limited by the size of the coefficient norm $|\boldsymbol{\alpha}|$ (as opposed to the condition number). Therefore we have analysed, for the first time, the growth of $|\boldsymbol{\alpha}|$ with basis size N as one converges towards Helmholtz solutions in the disc and other analytic domains. In the disc we have theorems on

convergence rate (Theorem 3) and coefficient growth (Theorems 6 and 7), and for analytic domains we have corresponding conjectures (Conjectures 9 and 11) supported by numerical experiments in many domain shapes. These show that the success (numerical stability and hence high accuracy) of the MFS relies on a choice of charge curve which does not enclose any singularities of the analytic continuation of the solution u. These singularities are associated either with the analytic continuation of the boundary data, or with the Schwarz function of the domain.

The conclusions for optimal choice of MFS charge points are as follows. For the unit disc, with concentric equally spaced charge points, a radius between $\sqrt{\rho}$ and ρ is optimal (Remark 8), ρ being the radius of the nearest singularity in boundary data. For general analytic domains, charge points placed on a curve which adapts to the singularity locations have been shown to perform very well. At high wavenumber we show that in the disc asymptotically two basis functions per wavelength on the boundary are needed, and that in more complicated nonconvex domains with nearby singularities this need only increase by less than a factor of two to achieve error norms close to machine precision. In practice this unusually small N results in rapid solution of the basis coefficients using dense linear algebra, even at high wavenumber.

Although the goal of this work is not a detailed comparison of MFS against boundary integral equations (BIE), our experiments show that MFS can be highly competitive, with a smaller basis size N, and the advantage that accurate solutions are easy to evaluate up to the boundary. MFS is useful only if condition number is not a problem, which is true for direct solvers but not iterative ones. Thus we do not expect MFS can benefit in the way BIE do from Fast Multipole (FMM) acceleration in an iterative solver. However, recent developments in fast direct solvers [27] could remove this problem and allow MFS to be efficient in large-scale ($N > 5000$) settings.

As with BIE, the CPU time to evaluate the interior solution at several grid points per wavelength is much larger than the coefficient solution time, especially using a naive implementation of the sum (2) and MATLAB's Hankel function routine. Replacing this evaluation of u with a FMM summation would be a natural next step and is expected to result in a large speedup at high wavenumbers.

We expect our findings on coefficient growth rates, and the new adaptive charge curve algorithm, to be easily extendable to the exterior Helmholtz scattering problem, for which MFS has shown promise in the engineering community [24,14,8].

Acknowledgments

AHB is supported by the National Science Foundation under Grant DMS-0507614. TB is supported by Engineering and Physical Sciences Research Council Grant EP/D079403/1. TB wishes to thank the DMV, the DFG, and the Shapiro Visitor Program, for a joint travel grant in January 2007 to visit Dartmouth. This work benefitted from important discussions with Lehel Banjai, Leslie Greengard, David Karkashadze, Fridon Shubitidze, and Nick Trefethen, and from the reviewer comments.

Appendix A

Proof of Theorem 7 For even N fix $N > N_{\min} + 3$. Using (20) in (21) implies the trivial bound

$$\sqrt{2\pi}\left|\frac{N}{2\pi}\hat{s}(m)\hat{\alpha}_{m \bmod N} - \hat{v}(m)\right| \leqslant t \quad \text{for all } m \in \mathbb{Z}. \tag{A.1}$$

Define the (positive) maximum Fourier frequency $F := \frac{N}{2} - K$, where K is the unique integer such that

$$\frac{N_{\min}}{2} \leqslant K < \frac{N_{\min}}{2} + 1.$$

Note that K is independent of N. One can verify using (31), (30) and the definition of $N_{\min}$ that

$$|\hat{v}(m)| \geqslant \sqrt{\frac{2}{\pi}}\, t \quad \text{for all } |m| \leqslant F. \tag{A.2}$$

In the frequency range $|m| \leqslant F$, it follows from (A.1) and (A.2) that $\hat{v}(m)$ is sufficiently large relative to t to bound the coefficients away from zero,

$$|\hat{\alpha}_m| \geqslant \frac{2\pi}{N}\frac{|\hat{v}(m)| - t/\sqrt{2\pi}}{|\hat{s}(m)|} \geqslant \frac{\pi}{N}\frac{|\hat{v}(m)|}{|\hat{s}(m)|} \geqslant \frac{\pi c_v |m|}{C_s N}\left(\frac{R}{\rho}\right)^{|m|}, \quad 0 < |m| \leqslant F, \tag{A.3}$$

where the last step used (15) and (30). Choosing the maximal frequency $m = F = \frac{N}{2} - K$ we obtain

$$|\hat{\alpha}_F| \geqslant \frac{\pi c_v}{C_s}\left(\frac{R}{\rho}\right)^{\frac{N}{2}-K}\left(\frac{1}{2} - \frac{K}{N}\right) > \frac{\pi c_v}{2C_s}\left(\frac{R}{\rho}\right)^{\frac{N}{2}-K}\frac{1}{N_{\min}+3}. \tag{A.4}$$

Here the latter inequality follows from

$$\frac{1}{2} - \frac{K}{N} > \frac{1}{2} - \frac{N_{\min}/2 + 1}{N_{\min}+3} = \frac{1}{2(N_{\min}+3)}.$$

Absorbing the N-independent factors of (A.4) into a constant and noticing that the Euclidean norm of a vector is at least as large as its largest component we have

$$|\boldsymbol{\alpha}| = \sqrt{N}|\hat{\boldsymbol{\alpha}}| \geqslant \sqrt{N}|\hat{\alpha}_F| \geqslant C\sqrt{N}\left(\frac{R}{\rho}\right)^{N/2} \tag{A.5}$$

for a sufficiently small constant $C > 0$, as claimed in the theorem. □

Appendix B. Bessel function asymptotics

We take the standard Taylor series [1]

$$J_m(z) = \left(\frac{z}{2}\right)^m \sum_{k=0}^{\infty} \frac{(-z^2/4)^k}{k!(m+k)!} \tag{B.1}$$

and in the large-m limit we may approximate $(m+k) \approx mm^k$, then recognize the power series for the exponential, giving

$$J_m(z) \sim \frac{1}{m!}\left(\frac{z}{2}\right)^m \mathrm{e}^{-z^2/4m}, \quad m \to \infty. \tag{B.2}$$

Similarly the standard series

$$Y_m(z) = -\frac{1}{\pi}\left(\frac{z}{2}\right)^{-m} \sum_{k=0}^{m-1} \frac{(m-k-1)!(z^2/4)^k}{k!} + \frac{2}{\pi}\ln(z/2)J_m(z) + O(z^m) \tag{B.3}$$

with $(m-k-1)! \approx m!/m^k$ gives

$$Y_m(z) \sim -\frac{m!}{\pi}\left(\frac{z}{2}\right)^{-m} \mathrm{e}^{z^2/4m}, \quad m \to \infty. \tag{B.4}$$

Neither of these asymptotic forms are given in [1], however (B.2) has been recently noted in the physics community [25]. Combining these two in (12), using the reflection formulae, and ignoring the lower-order J contribution to the Hankel function, gives (39).

Bessel's equation $u'' + u'/r + (1 - m^2/r^2)u = 0$ with the Liouville transformation $w = r^{1/2}u$ then changing variable to $x = r/a$, with $a^2 = m^2 - \frac{1}{4}$, gives the ODE

$$\frac{d^2 w}{dx^2} + a^2\left(1 - \frac{1}{x^2}\right)w = 0. \tag{B.5}$$

The WKBJ (or Liouville–Green) asymptotic approximation (Chapter 9.3 of [30]) for large parameter a is then

$$w(x) \sim \begin{cases} (x^{-2}-1)^{-1/4}\left(A\mathrm{e}^{a\int_x^1 \sqrt{x^{-2}-1}\,\mathrm{d}x} + B\mathrm{e}^{-a\int_x^1 \sqrt{x^{-2}-1}\,\mathrm{d}x}\right), & x<1 \text{ (evanescent)}, \\ (1-x^{-2})^{-1/4}\left(C\mathrm{e}^{\mathrm{i}a\int_1^x \sqrt{1-x^{-2}}\,\mathrm{d}x} + D\mathrm{e}^{-\mathrm{i}a\int_1^x \sqrt{1-x^{-2}}\,\mathrm{d}x}\right), & x>1 \text{ (oscillatory)}, \end{cases} \tag{B.6}$$

where $A, B, C, D \in \mathbb{C}$ are constants. Note that the integral in the evanescent region can be performed analytically and is $-I_1(x)$ as defined by (41); the integral in the oscillatory region is not needed since amplitude not phase is of interest. Since the solution w is continuous through the turning point $x = 1$ (even though (B.6) breaks down), there exist connection formulae relating the constants:

$$C = \mathrm{e}^{\mathrm{i}\pi/4}A + \mathrm{e}^{-\mathrm{i}\pi/4}B, \quad D = \mathrm{e}^{\mathrm{i}\pi/4}B. \tag{B.7}$$

They can be found by comparing WKBJ to large-argument asymptotics of the Airy functions Ai and Bi on either side of the turning point (e.g. comparing 10.4.59 with 10.4.60, and 10.4.63 with 10.4.64, in [1], or using 9.3.91,92 of [30], or the more rigorous presentation of the Gans–Jeffreys formulae in Chapter 11 of [31]). If $A = 0$ (w decaying as x decreases in the evanescent region) then $|C| = |D| = |B|$ giving an amplitude of $2|B|/(1-x^{-2})^{1/4}$ in the oscillatory region. When transformed back such a solution $u(r)$ corresponds to the $J_m(r)$ Bessel function. We match the asymptotic amplitude $\sqrt{2/\pi r}$ at large argument (see 9.2.1 of [1]) to fix $B = \frac{1}{2}$ for all m. Hence the Bessel function has typical size

$$|J_m(r)| \approx \begin{cases} \frac{1}{2}(a^2-r^2)^{-1/4}\mathrm{e}^{I_a(r)}, & r<a, \\ (a^2-r^2)^{-1/4}, & r>a. \end{cases} \tag{B.8}$$

Note that an amplitude is implied here in the oscillatory region $r > a$. Note also that a is defined above (B.5), and $I_a(r) < 0$ for $r < a$. Similarly matching the $H_m^{(1)}(r)$ Hankel function large-argument asymptotic gives $|C| = 1$, $D = 0$, so $|A| = 1$ (which dominates), thus typical size

$$|H_m^{(1)}(r)| \approx \begin{cases} (a^2-r^2)^{-1/4}\mathrm{e}^{-I_a(r)}, & r<a, \\ (a^2-r^2)^{-1/4}, & r>a. \end{cases} \tag{B.9}$$

These formulae have been checked against numerical evaluations of Bessel functions and accurately predict amplitudes or evanescent magnitudes everywhere apart from very close to the turning point $r = a$ where they have a weak algebraic singularity, but still provide an upper bound. Substituting (B.8) and (B.9) into (12) gives the desired (40).

References

[1] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 10th ed., Dover, New York, 1964.

[2] M.T. Ahmed, J.D. Lavers, B.E. Burke, An evaluation of the direct boundary element method and the method of fundamental solutions, IEEE Trans. Magn. 25 (1989) 3001–3006.

[3] A.H. Barnett, Asymptotic rate of quantum ergodicity in chaotic Euclidean billiards, Commun. Pure Appl. Math. 59 (2006) 1457–1488.

[4] A. Bogomolny, Fundamental solutions method for elliptic boundary value problems, SIAM J. Numer. Anal. 22 (4) (1985) 644–669.

[5] D. Colton, R. Kress, Inverse Acoustic and Electromagnetic Scattering Theory, Applied Mathematical Sciences, second ed., vol. 93, Springer-Verlag, Berlin, 1998.

[6] R. Courant, D. Hilbert, Methods of Mathematical Physics, vol. I, Interscience Publishers Inc., New York, NY, 1953.

[7] P.J. Davis, The Schwarz Function and Its Applications, The Carus Mathematical Monographs, No. 17, The Mathematical Association of America, Buffalo, NY, 1974 .

[8] A. Doicu, Y.A. Eremin, T. Wriedt, Acoustic and Electromagnetic Scattering Analysis Using Discrete Sources, Academic Press, San Diego, CA, 2000.

[9] T.A. Driscoll, B. Fornberg, Interpolation in the limit of increasingly flat radial basis functions, Comput. Math. Appl. 43 (2002) 413–422.

[10] R. Ennenbach, H. Niemeyer, The inclusion of Dirichlet eigenvalues with singularity functions, Z. Angew. Math. Mech. 76 (7) (1996) 377–383.

[11] G. Fairweather, A. Karageorghis, The method of fundamental solutions for elliptic boundary value problems, Adv. Comput. Math. 9 (1-2) (1998) 69–95, numerical treatment of boundary integral equations.

[12] P.R. Garabedian, Applications of analytic continuation to the solution of boundary value problems, J. Rational Mech. Anal. 3 (1954) 383–393.
[13] A. Karageorghis, The method of fundamental solutions for the calculation of the eigenvalues of the Helmholtz equation, Appl. Math. Lett. 14 (7) (2001) 837–842.
[14] D. Karkashadze, On status of main singularities in 3D scattering problems, in: Proceedings of VIth International Seminar/Workshop on Direct and Inverse Problems of Electromagnetic and Acoustic Wave Theory (DIPED), Lviv, Ukraine, 2001.
[15] M. Katsurada, A mathematical study of the charge simulation method. II, J. Fac. Sci. Univ. Tokyo Sect. IA Math. 36 (1) (1989) 135–162.
[16] M. Katsurada, Asymptotic error analysis of the charge simulation method in a Jordan region with an analytic boundary, J. Fac. Sci. Univ. Tokyo Sect. IA Math. 37 (3) (1990) 635–657.
[17] M. Katsurada, Charge simulation method using exterior mapping functions, Jpn. J. Ind. Appl. Math. 11 (1) (1994) 47–61.
[18] M. Katsurada, H. Okamoto, A mathematical study of the charge simulation method. I, J. Fac. Sci. Univ. Tokyo Sect. IA Math. 35 (3) (1988) 507–518.
[19] M. Katsurada, H. Okamoto, The collocation points of the fundamental solution method for the potential problem, Comput. Math. Appl. 31 (1) (1996) 123–137.
[20] T. Kitagawa, On the numerical stability of the method of fundamental solution applied to the Dirichlet problem, Jpn. J. Appl. Math. 5 (1) (1988) 123–133.
[21] T. Kitagawa, Asymptotic stability of the fundamental solution method, in: Proceedings of the International Symposium on Computational Mathematics (Matsuyama, 1990), vol. 38, 1991.
[22] J.R. Kuttler, V.G. Sigillito, Bounding eigenvalues of elliptic operators, SIAM J. Math. Anal. 9 (4) (1978) 768–778.
[23] A.G. Kyurkchan, The method of auxiliary currents and sources in wave diffraction problems, Soviet J. Commun. Tech. Electron. 30 (1985) 48–58, translated from Radiotekhn. i Èlektron. 29 (no. 11) (1984) 2129–2139 (Russian).
[24] A.G. Kyurkchan, B.Y. Sternin, V.E. Shatalov, Singularities of continuation of wave fields, Physics – Uspekhi 12 (1996) 1221–1242.
[25] B. Landry, E.J. Heller, Statistical properties of many particle eigenfunctions, J. Phys. A 40 (2007) 9259–9274.
[26] H. Lewy, On the reflection laws of second order differential equations in two independent variables, Bull. Am. Math. Soc. 65 (1959) 37–58.
[27] P.G. Martinsson, V. Rokhlin, A fast direct solver for boundary integral equations in two dimensions, J. Comput. Phys. 205 (2005) 1–23.
[28] R.F. Millar, The analytic continuation of solutions to elliptic boundary value problems in two independent variables, J. Math. Anal. Appl. 76 (2) (1980) 498–515.
[29] R.F. Millar, Singularities and the Rayleigh hypothesis for solutions to the Helmholtz equation, IMA J. Appl. Math. 37 (2) (1986) 155–171.
[30] P. Morse, H. Feshbach, Methods of Theoretical Physics, vol. 2, McGraw-Hill, 1953.
[31] F.W.J. Olver, Asymptotics and Special Functions, Academic Press, New York, 1974.
[32] Y.-S. Smyrlis, A. Karageorghis, Numerical analysis of the MFS for certain harmonic problems, M2AN Math. Model. Numer. Anal. 38 (3) (2004) 495–517.

J. Math. Biol. (2008) 57:139–159
DOI 10.1007/s00285-007-0149-8

Mathematical Biology

Analytic steady-state space use patterns and rapid computations in mechanistic home range analysis

Alex H. Barnett · Paul R. Moorcroft

Received: 30 May 2007 / Revised: 12 November 2007 / Published online: 7 December 2007
© Springer-Verlag 2007

Abstract Mechanistic home range models are important tools in modeling animal dynamics in spatially complex environments. We introduce a class of stochastic models for animal movement in a habitat of varying preference. Such models interpolate between spatially implicit resource selection analysis (RSA) and advection-diffusion models, possessing these two models as limiting cases. We find a closed-form solution for the steady-state (equilibrium) probability distribution u^* using a factorization of the redistribution operator into symmetric and diagonal parts. How space use is controlled by the habitat preference function w depends on the characteristic width of the animals' redistribution kernel: when the redistribution kernel is wide relative to variation in w, $u^* \propto w$, whereas when it is narrow relative to variation in w, $u^* \propto w^2$. In addition, we analyze the behavior at discontinuities in w which occur at habitat type boundaries, and simulate the dynamics of space use given two-dimensional prey-availability data, exploring the effect of the redistribution kernel width. Our factorization allows such numerical simulations to be done extremely fast; we expect this to aid the computationally intensive task of model parameter fitting and inverse modeling.

Keywords Mechanistic · Home range · Advection–diffusion · Resource selection analysis · Markov · Space use

Mathematics Subject Classification (2000) 92D50 · 35K15 · 65C20

A. H. Barnett (✉)
Department of Mathematics, 6188 Kemeny Hall,
Dartmouth College, Hanover, NH 03755, USA
e-mail: ahb@math.dartmouth.edu

P. R. Moorcroft
OEB Department, Harvard University, 22 Divinity Ave,
Cambridge, MA 02138, USA

1 Introduction

Due to the influences of habitats on the availability of food, shelter, mates and risk of predation, patterns of animal space use are strongly shaped by the spatial distribution of habitat types across landscapes. Revealing these effects is key for ecological studies, wildlife management, and conservation efforts. A popular technique for analyzing such relationships between habitats and patterns of animal space use has been resource selection analysis (RSA) [3,10,13], in which the intensity of space use is taken to reflect an underlying resource selection function summarizing an individual's preference for the habitat type(s) found at that location.

Recently an alternative framework has emerged in the form of mechanistic home range models [12,17,18]. In contrast to RSA, which is largely descriptive and takes into account the effects of finite movement speeds only implicitly, mechanistic models yield spatially explicit predictions for patterns of animal space use in the form of a probability density function (pdf), by modeling the process of individual movement. Mathematically, the fine-scale behavior of individuals is treated as a stochastic (Markov) process [11,19,21,23], specifying the probability of an animal at a given location moving to a subsequent location during a given time interval. From this model one can derive, in the limit of small time intervals, a continuous-time partial differential equation (PDE) for the evolution of the pdf. For example, a recent analysis of coyote home ranges in Yellowstone [18] used a "prey availability plus conspecific avoidance" (PA+CA) mechanistic home range model to account for the observed patterns of coyote home ranges within the park. In this model, individuals exhibited an avoidance response to encounters with foreign scent marks and a foraging response to prey availability (individuals decreased their mean step length in response to increasing small mammal abundance in different habitats.) In such mechanistic models, inferences about long-term space use are usually made by evolving the continuous-time PDE in order to converge to its steady-state; this can be computationally time-consuming [15,17].

In recent work [16], we developed a mechanistic home range model which reconciles these two main approaches by combining the concept of a spatial preference function $w(x)$ with a stochastic model of fine-scale movement behavior. At each time-step the movement of an individual is governed by its relative preference for the *local* habitat surrounding its current spatial position i.e., the preference function restricted to a region of size L centered on the individual's current location. The length scale L has two roles: it is the typical (jump) distance per time step, but also can be interpreted as the distance over which the animal is able to *perceive* differences in surrounding habitats. This new model, which we will call a mechanistic RSA, or MERSA model, is compared against traditional RSA in Fig. 1, and detailed in Sect. 2.

In traditional RSA, available habitat is assumed to be a fixed area such as the study area or an independent estimate of the animal's home range. Our approach has some aspects in common with more recent variants of RSA, which build on the work of Arthur et al. [1], that instead define available habitat relative to the animal's current location. Recall that RSA is generally a statistical regression to reveal the factors which determine space use relative to availability. These RSA variants restrict the definition of available habitat to a buffer region centered on the animal's current location: either

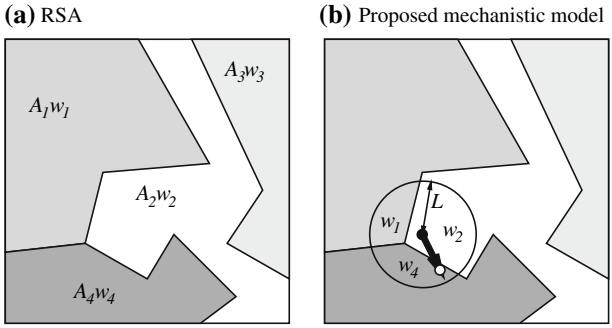


Fig. 1 Schematic comparing conventional resource selection analysis (*RSA*) against our proposed model for space use, in two dimensions. *Shaded regions* show areas of habitat in which the preference function $w(x)$ is constant. **a** Traditional RSA. The probability of finding an animal in each region is proportional to the product of preference (resource selection function) w_j and area of the region A_j. There is no account taken of the animal's current habitat type or location within the habitat. **b** Proposed MERSA model for space use. The animal responds only to the preference function within a distance scale L from its current location (*black dot*). Each time step the future location (*white dot*) is chosen randomly from a localized distribution biased by the preference function

uniformly sampled discs [1,2,4,9], or nonuniform sampling according to a known redistribution kernel [6]. Thus some movement model (i.e., upper limit or pdf on travel distances between sample times) is implicitly assumed, but rarely explicitly described, and not fitted for within the regression framework, nor explored in terms of its predictions for evolution of space use.

Although our mechnistic jump length L has similarities to this buffer size, our framework is different. It combines jump step length, direction, and local habitat preference bias into a *single* probabilistic model of fine-scale movement behavior, from which we can derive mathematically the pattern of space use that results from the underlying movement process. Our ultimate goal is then to perform inference (fitting for all parameters and comparing models) given location data, for instance extending the approach of Forester et al. [5]. In this preliminary paper we explore model predictions in various preference environments, and leave inference from real-world animal location datasets for future work.

Our main result in [16] was to show that, in one dimension (1D) with a spatially smooth habitat preference function, using the Kramers–Moyal expansion ([7,16] and Appendix A and E of [17]) one may derive an an advection–diffusion equation for the expected patterns of space use, with advection and diffusion coefficients related to the parameters of the underlying stochastic movement model. We then showed that the resulting steady-state pdf could be determined analytically, giving an intensity of space use proportional to the square of the preference function.

In this paper we expand and generalize this result to the more biologically relevant case of individuals moving in two space dimensions across landscapes, which may include discrete habitat types and the resulting *non-smooth* preference function. Specifically, in Sect. 3 we solve for the steady-state pdf of our discrete-time stochastic model of fine-scale movement behavior directly and analytically, for arbitrary preference functions, without resorting to the conventional procedure of taking the

continuous-time limit.[1] (Note that this relies on a particular algebraic feature of our model; for a general redistribution kernel such an analytic solution is not available.) By doing so, we are able rapidly to compute exact steady-state space use patterns for the model for a full range of possible length scale values L, rather than being confined to the limit of small L inherent in a PDE-based approach. We also gain an understanding of the numerically observed pdf behavior at jumps in preference function, which had remained a mystery [16]. A consequence of our solution is that the computationally intensive task of solving an inverse problem to fit multiple model parameters can become orders of magnitude faster. We discuss such numerical implications in Sect. 3.1, and give CPU timings for our numerical examples throughout.

For illustrative purposes, we choose a translationally invariant exponential distribution of jump lengths, a kernel which has proven useful for modeling coyote foraging movement [17]. However, our analytic solution holds for any redistribution kernel, also for a generalization of the model of [16] to *spatially varying* diffusion coefficient (spatially dependent length scale L) such as occurs in modeling the prey-density dependent foraging rate of wolves and coyotes [12,18,24]. In Sect. 4.1 we explore the transition from small L (where the model tends to the advection–diffusion equation derived in [16]), to large L (where the model becomes equivalent to traditional RSA). Here "large" and "small" are in comparison to the typical spatial scale on which the preference function changes. In Sect. 4.1 we explore this numerically both in 1D, and with a 2D preference function derived from discontinuous real-world small mammal abundance data appropriate for coyotes. In Sect. 4.2 we show rapid and efficient numerical simulation of the time evolution of the pdf. Furthermore in Sect. 5 we analyze the behavior at a sharp discontinuity in preference function, and show that on spatial scales larger than L this effect may be approximated by effective matching conditions in coupled advection-diffusion equations. Finally in Sect. 6 we discuss implications and draw conclusions.

2 Mechanistic spatially explicit (MERSA) model

The time-dependent pdf of an animal we will represent by $u(x,t)$, where $x \in \Omega$ is the location, $\Omega \subset \mathbb{R}^d$ represents the habitat region or study area of interest, and d is the dimensionality of space (usually 1 or 2). Thus, in a 1D setting, $u(x,t)dx$ is the probability that at time t a given individual is to be found in the interval $[x, x+dx]$. (Note in 2D we use x rather than $\boldsymbol{x}$ to represent location vector.) Its normalization is

$$\int_{\Omega} u(x,t)dx \;=\; 1 \qquad \text{for all } t \tag{1}$$

[1] One may ask whether discrete or continuous time is more appropriate for animal movement modeling. Clearly in reality animals move in continuous time, however a continuous-time diffusion equation (Brownian process) cannot be realistic on the shortest time-scales for the simple reason that this would require infinite movement speed (e.g., see [20]). Therefore models with a fixed discrete time step remain crucial for fine-scale modeling.

Our MERSA model is an example of a Markov process. At each time step of length τ the current pdf at time t is acted on by a fixed linear operator to get the pdf at time $t+\tau$. This is expressed by the master equation

$$u(x, t+\tau) = \int_{\Omega} k(x, x')u(x', t)dx' \qquad \text{for } x \in \Omega \tag{2}$$

Given an initial pdf $u(x, 0) = u_0(x)$ for all $x \in \Omega$, by iteration the pdf at arbitrarily large future times (multiples of τ) may be computed. Here the redistribution operator kernel $k(x, x')$ is defined as the conditional pdf of an individual animal's location x, a time interval τ into the future, given that its current location is x'.[2] This is an uncorrelated jump or "kangaroo" process [20] in which the animal has no memory beyond the fixed time scale τ. Since a redistribution kernel is a conditional pdf it is everywhere non-negative and ("columnwise") normalized by

$$\int_{\Omega} k(x, x')dx = 1 \qquad \text{for all } x' \in \Omega \tag{3}$$

Clearly for a given ecological situation the choice of τ determines the form of $k(x, x')$. For example, shorter τ may demand a smaller kernel width L simply because animal speed is limited. By the length scale L we mean the typical size of $|x - x'|$ for which $k(x, x')$ is significant. We will not indicate explicitly the dependence on τ of the form of k. The appropriate value of τ depends on the application; it needs to be large enough that successive animal relocations can be approximately treated as uncorrelated [17]. Real-world location data collection technology also can be a factor if fine-scale model fitting is to be done. For coyotes, a typical value of τ is 10 min [17].

We consider a spatial preference (resource selection) function $w \in L^1(\Omega)$ which controls relative preference for each location in the domain. We represent unbiased ("preference-free") diffusive animal movement with a *symmetric* redistribution kernel $\phi(x, x')$, that is,

$$\phi(x, x') = \phi(x', x) \qquad \text{for all } x, x' \in \Omega. \tag{4}$$

The kernel ϕ obeys the Markov normalization (3); from this and symmetry it follows that

$$\int_{\Omega} \phi(x, x')dx' = 1 \qquad \text{for all } x \in \Omega \tag{5}$$

which is the statement that a constant pdf is invariant under redistribution by ϕ. Since a uniform density gives no net probability mass flow, we say that (an operator with

[2] Note that in the stochastic literature the order of x and x' is often reversed [7].

kernel) ϕ is *advection-free*. Our MERSA redistribution kernel k is this advection-free jump kernel biased by the preference function, in other words,

$$k(x, x') = \frac{w(x)\phi(x, x')}{z(x')} \tag{6}$$

with normalization function (easily seen to be required to satisfy (3)),

$$z(x') := \int_{\Omega} w(x'')\phi(x'', x')dx''. \tag{7}$$

For (6) to be meaningful we must have $z > 0$ everywhere; it is sufficient that $w > 0$ everywhere for this to hold, which we will assume from now on. Note that since $\phi(x, x')$ may depend independently on x and x' (barring the symmetry constraint), it may represent a spatially varying (and also anisotropic) diffusion coefficient. The MERSA model is thus more general than that of [16], which was restricted to the translationally invariant case

$$\phi(x, x') = \tilde{\phi}(x - x'). \tag{8}$$

Here $\tilde{\phi}(\cdot)$ is a function of relative displacement alone, which limits one to a spatially invariant distribution of step-lengths in the underlying stochastic movement model.

2.1 Limiting cases of the model

We now discuss two limiting forms of MERSA. Firstly, consider the case $\phi(x, x') = 1/\mathrm{vol}(\Omega)$ for all $x, x' \in \Omega$. This corresponds to preference-free redistribution to a uniformly random location in Ω, without regard to current location. (6) then becomes

$$k(x, x') = \frac{w(x)}{\int_{\Omega} w(x'')dx''}. \tag{9}$$

Since this is independent of x', within a *single* time step (and for all future time steps) the master equation reaches its steady-state pdf $u^* := \lim_{t\to\infty} u(\cdot, t)$ given by

$$u^*(x) = C_1 w(x), \qquad \text{with normalization constant} \quad C_1^{-1} = \int_{\Omega} w(x'')dx''. \tag{10}$$

This is formally equivalent to a conventional (time-invariant) RSA model, with linear dependence on preference. This assertion is illustrated when Ω is divided into regions $j = 1 \ldots m$ each of area A_j and constant preference function w_j, for then (10) assumes the more familiar RSA form $u_j = A_j w_j / (\sum_{i=1}^{m} A_k w_k)$ where u_j is the probability of being in region j [3,10,13], see Fig. 1a. In the case of an infinite domain such as $\Omega = \mathbb{R}^d$ (in which case w alone delineates the habitat) no constant normalizable ϕ exists; however, the above result may be reproduced by considering

the limit in which the kernel ϕ becomes much wider than all spatial scales of interest in the habitat ($L \to \infty$). Then $\phi(x, x')$ tends to a constant for x and x' within the habitat, and (10) again follows. Thus for both cases above we will call this the $L \to \infty$ limit, and state that in this limit our MERSA model degenerates to traditional RSA (i.e., no restrictions on available habitat).

Secondly, consider the $L \to 0$ limit where k tends to a diagonal kernel. In order for time evolution to take place at all, we must also take the limit $\tau \to 0$. It is well known [19,20] that the correct way to balance these two limits in order to reach a well-defined diffusion coefficient is to choose the variance of the kernel k to scale as τ. In the 1D case of smooth w, and a translation-invariant kernel (8), we have derived [16] that in this limit our MERSA model gives a Fokker–Planck PDE with known advection and diffusion coefficients. From this we showed that the steady-state pdf is

$$u^*(x) = C_2 w^2(x), \qquad \text{with constant} \quad C_2^{-1} = \int_\Omega w^2(x'')dx'', \tag{11}$$

that is, quadratic in preference function. (The integral is bounded since w is smooth and in $L^1(\Omega)$). We note that in this Fokker-Planck limit, our model is equivalent to that of [8] with the "potential" function $U(x) = -2D \log w(x)$ and constant diffusion $d(x) = D$.

We will see in Sect. 4.1 how these differing $L \to \infty$ and $L \to 0$ steady-state limits are reached in practice.

3 Analytic formula for steady-state pdf

The condition that u^* be a steady-state pdf is that it be invariant under the master equation (2), in other words,

$$u^*(x) = \int_\Omega k(x, x')u^*(x')dx' \qquad \text{for all } x \in \Omega. \tag{12}$$

Our main result is the following claim.

Proposition 1 *A steady-state pdf for the model redistribution kernel* (6) *is given by*

$$u^*(x) = C w(x) z(x), \qquad \textit{with constant} \quad C^{-1} = \int_\Omega w(x')z(x')dx', \tag{13}$$

where the function z is defined by (7).

The proposition is proved by substituting (6) and (13) into the right side of (12) then noticing that z cancels, allowing the simplification

$$\begin{aligned}\int_{\Omega} k(x,x')Cw(x')z(x')dx' &= C\int_{\Omega} w(x)\phi(x,x')w(x')dx' \\ &= Cw(x)\int_{\Omega}\phi(x',x)w(x')dx' \\ &= Cw(x)z(x), \\ &= u^*(x), \end{aligned} \tag{14}$$

verifying (12). Crucially, it is the assumption that ϕ is symmetric that allows us to proceed from the first to second line. Notice that once standard assumptions about ergodicity are satisfied, the steady-state u^* is unique (e.g., see Doeblin condition in [14] p.396; in our ecological context this is satisfied since we can assume that there is always some randomness to animal motion, i.e., the kernel ϕ is always somewhat spreading at each spatial location). Since w is assumed to be everywhere positive, so is u^*.

We now derive some secondary results on the structure of k, giving intuition into the reason for existence of the simple formula (13). They may be skipped if no further mathematical insight is required. We switch to operator notation, expressing (12) as $u^* = Ku^*$ where K is a Markov operator. Recall that a Markov operator is an integral operator with non-negative kernel obeying (3), which can be expressed $K^T 1 = 1$ where 1 is the constant function and K^T the adjoint operator (with respect to uniform measure). Our model (6) expresses the factorization

$$K = W\Phi Z^{-1}, \tag{15}$$

where Φ is the (Markov) operator defined by the integral kernel $\phi(x,x')$, and W and Z are the diagonal operators which multiply by the functions w and z respectively. Thus the structure is an operator with constant invariant measure sandwiched between two diagonal operators. Note that $z = \Phi w$ since Φ is symmetric.

Then we have,

Proposition 2 *With the assumption $w \geq c$ everywhere, for some $c > 0$, our model Markov operator* (15)

1. *satisfies detailed balance, that is,*

$$k(x,x')u^*(x') = k(x',x)u^*(x) \qquad \textit{for all } x, x' \in \Omega, \tag{16}$$

2. *is self-adjoint with respect to the measure $1/u^*$, and*
3. *has all eigenvalues real.*

The proof is as follows. We define U^* to be the operator multiplying by the function u^*; it can be written $U^* = ZW$. Using (15) gives $KU^* = W\Phi W$, explicitly symmetric.

In other words

$$(KU^*)^T = KU^*, \tag{17}$$

which is equivalent to detailed balance (16). Now we use $(\cdot, \cdot)$ to indicate the $L^2(\Omega)$ (real) inner product with respect to uniform measure, and $(\cdot, \cdot)_{1/u^*}$ with respect to measure $1/u^*$. Using (17) and the boundedness of $1/u^*$ in the middle step we have

$$\begin{aligned}(a, Kb)_{1/u^*} &= \left(\frac{a}{u^*}, KU^*\frac{b}{u^*}\right) = \left(KU^*\frac{a}{u^*}, \frac{b}{u^*}\right) \\ &= (Ka, b)_{1/u^*} \qquad \text{for all } a, b \in L^2(\Omega)\end{aligned} \tag{18}$$

which proves part 2. Part 3 immediately follows by self-adjointness. We remark that the simple formula (13) for u^* can now be seen as a result of the need to symmetrize the factorization (15).

3.1 Implications for fast numerical modeling

We discuss briefly why the above result is important in practice. The analytic steady-state pdf formula (13) is special because for a general redistribution kernel no such formula exists. In particular, there is no formula for u^* for commonly used mechanistic kernels such as those expressed in polar coordinates with exponential radial jumps and a von Mises angular distribution [17, Chap. 3] In that work all steady-state distributions had to be computed in the advection–diffusion limit, often in a numerically intensive fashion; see Appendix G of [17], and [15]. In fitting model parameters to real-world location data, a large number of such steady-state pdfs must be found as part of the parameter-optimization process (e.g., Sect. 4.3 of [17]).

Numerical solution of u^* for any redistribution kernel requires discretization of the domain into N degrees of freedom. In 2D the N required for acceptable accuracy can be large (e.g., 10^4). Solving for u^* given a general kernel is then an eigenvector problem involving a (possibly dense) N-by-N matrix discretization of that kernel. The iterative solution of such large eigenproblems can be slow especially when diffusion rates are small. In constrast, the formula (13) in our MERSA model bypasses this and requires only the computational effort of the *single* matrix-vector multiplication required to evaluate the (discretized) integral (7). This is an acceleration by orders of magnitude.

There is a further numerical advantage to a special case of the MERSA model. Namely if ϕ is translationally invariant (diffusion coefficient is spatially constant) then the action of the (discretized) convolution operator Φ may be computed in time $O(N \ln N)$ via the Fast Fourier Transform (FFT) [22], which for large N is much faster than the $O(N^2)$ dense matrix-vector multiply. This further speeds up computing u^* via (7). Finally, the time-evolution $u(x, t)$ can now be computed much more efficiently. A single time-step of the master equation (2) may be performed with $O(N \ln N)$ effort by using (15) in a *split-operator method*: division by z, followed by FFT application of Φ, followed by multiplication by w.

We illustrate these numerical techniques and advantages below. All CPU timings are reported using a single core of a 2 GHz Intel Core Duo processor running MATLAB 7.0 in GNU/Linux.

4 Numerical results for an exponential jump kernel

In this section we illustrate the predictions of our MERSA model for steady-state and time evolution of space use for an idealized 1D preference function, and for a 2D preference function that is based on spatially complex, real-world measurements of prey abundance in different habitat types. Throughout we consider a translationally invariant ϕ kernel, choosing an exponential kernel with width L, which takes the form in 1D

$$\tilde{\phi}(\rho) = \frac{1}{2L} e^{-|\rho|/L} \tag{19}$$

where $\rho := x - x'$, and in 2D

$$\tilde{\phi}(\rho) = \phi(|\rho|) = \frac{1}{2\pi L} \frac{e^{-|\rho|/L}}{|\rho|} \tag{20}$$

where we remind the reader that ρ is a vector in the latter. Note that the 2D kernel is radially symmetric, and when integrated over angle it gives a distribution of jump distances $r := |\rho|$ which is exponential, $(1/L)e^{-r/L}$. The 2D kernel has been used as a model of animal movement, for instance describing the fine-scale movement of coyotes with remarkable accuracy [17] (it also seems promising for modeling observed relocation distances of elk [6, Fig. 1]).

4.1 Steady-state

Here we will assume the domain has periodic boundary conditions. It has the interpretation that the piece of habitat Ω in question is surrounded by similar (repeating) habitat, often a reasonable assumption. Mathematically this condition is achieved in the case of the unit interval $\Omega = [0, 1)$ by replacing (19) by a sum over a few nearby "image" kernels,

$$\tilde{\phi}(\rho) = \frac{1}{2L} \sum_{m=-M}^{M} e^{-|\rho - m|/L} \tag{21}$$

where M is chosen such that $\tilde{\phi}$ is periodic to some high accuracy (e.g., 10^{-6}). An analogous 2D image sum is used in the 2D case where Ω is a rectangle.

Note that our formalism can handle other boundary conditions. For instance we model a standard zero-flux boundary in Sect. 4.2. We have also checked that the periodic steady-state results are very similar to those for zero-flux, except when L is of

order the size of the whole domain (In this case for zero-flux u^* is affected by the effective boundary discontinuity in w, a complication which we will not pursue here).

4.1.1 One-dimensional case

In Fig. 2 we show u^* computed via (13) and (7) for the exponential kernel, for an ascending sequence of L values, in the unit interval. Our example preference function w has been chosen to exhibit a variety of length scales: for $x < 0.6$ it is smooth, corresponding to a gradation in habitat preference, whereas for $x > 0.6$ it is piecewise constant with discontinuous oscillations between the values 1 and 2 corresponding to isolated patches of more favorable habitat. The shortest length scale of w is 0.015, namely the size of the smallest constant patch near $x = 0.76$. We see that for very small L, u^* accurately matches w^2 for all regions apart from those with the most rapid w variations. This matches the expectation in Sect. 2.1.

A gradual transition is seen in the sequence of Fig. 2. As L becomes larger than a given feature, u^* in the vicinity of that feature starts to become locally proportional to w. Finally in plot (f), L is larger than any feature and u^* globally becomes linear

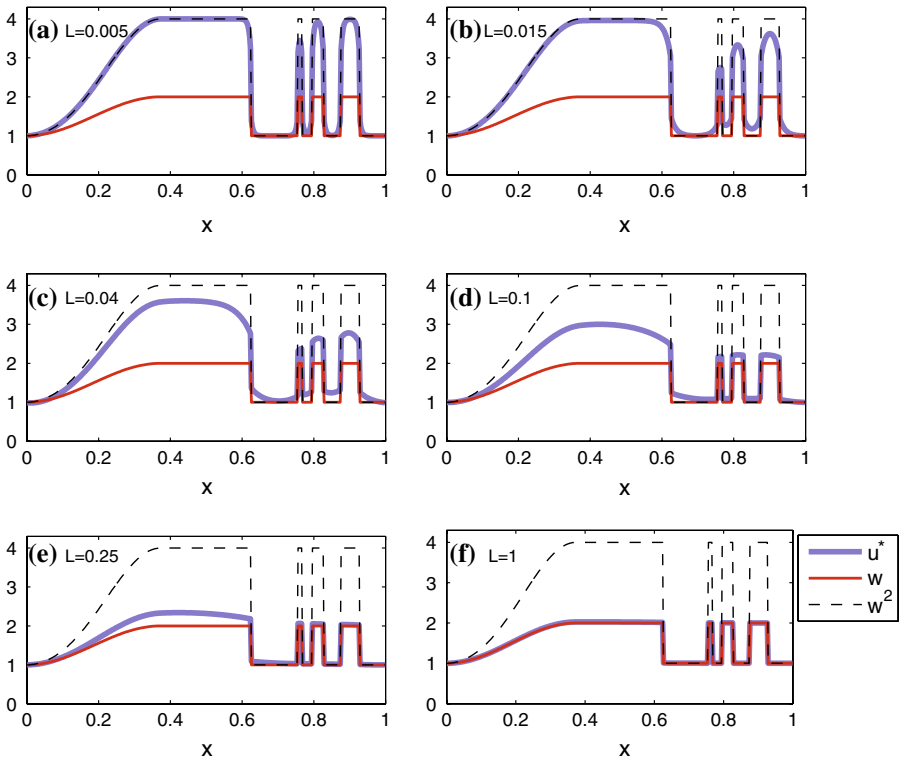


Fig. 2 Steady-state pdfs u^* for exponential jump pdf of (19) for an increasing sequence of L values. The preference function w (*thin solid line*) is chosen to be smooth on the left side of the domain, and discontinuous and oscillatory on the right side

in w. The explanation for this transition is simple and lies with (13) combined with the realization that the function z is given by the function w smoothed locally over a width of about L. Fine ($< L$) features in w will thus be smoothed away giving a locally constant z, whereas for coarse ($> L$) features $z \approx w$ and quadratic dependence in w results. This effect is common to any jump kernel with characteristic width L: fine-scale habitat features result in $u^* \propto w$, in accordance with the $L \to \infty$ limit (10), whereas coarse-scale habitat features are tracked according to $u^* \propto w^2$, in accordance with the Fokker–Planck limit (11).

Since u^* is computed analytically via (13), its numerical accuracy is limited only by quadrature of the integral. We computed z using (7) via FFT convolution (Sect. 3.1) in a few thousandths of a second on a uniform quadrature grid of $N = 400$ points. This contrasts the order 1 s needed to iteratively solve for the dominant eigenvector of the dense matrix discretization of Φ that would be required if no simplifying factorization of K or analytic solution for u^* were known.

4.1.2 Two-dimensional case with real-world habitat heterogeneity

We start with the small-mammal density data $B(x)$ shown in Fig. 3 where location x was sampled on a 0.1 km grid over a 7 km × 12 km domain Ω. This density is piecewise constant, being derived from measured prey densities (mice, ground squirrels, pocket gophers and red-backed voles) appropriate for coyotes in six different habitat types (see Chap. 7 of [17]). The main habitat feature is a strip of mesic grassland (the darkest region in the figure), which follows a valley floor and supports a high abundance of small mammals.

We use a linear relationship between prey availability and preference

$$w(x) = 1 + \alpha B(x) \qquad \text{for all } x \in \Omega \tag{22}$$

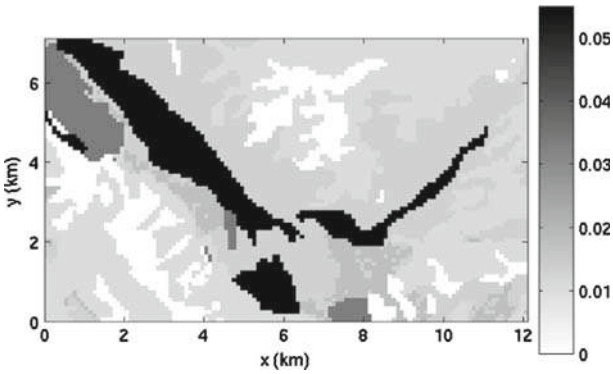


Fig. 3 Small-mammal abundance for Lamar Valley region of Yellowstone National Park collected by Crabtree et al. (unpublished); see Chap. 7 of [17]. Units are the total biomass density of small mammals in kg/ha with darker colors indicating larger values. The x- and y-axes are in kilometers. The highest abundances of small mammals are found in the mesic grassland habitats

where α (units of ha/kg) controls the strength of the preference per unit of prey biomass. The resulting steady-state pdfs u^* are shown in Fig. 4. This shows the patterns of space use for 3 different values of α, in combination with 3 different values of length scale L (20). Each column represents an α value, ranging from a weak preference (left column) to strong (right column). For comparison, the w function is shown at the top of each column.

The computational grid was moderately sized ($N = 8,591$), the same spatial resolution as the underlying estimates of prey abundance was sampled. Once $\phi(\rho)$ had been evaluated on the grid, solving for each steady-state pdf required only 0.009 s using (13) with z computed from w by 2D FFT convolution.[3] This is 100–1,000 times faster than an iterative solution for the dominant eigenvector of K if neither the factorization nor analytic formula are used (a single dense matrix application of a general K takes 0.45 s and many such iterations are required for convergence, the number depending on L and the particular w). Even if the factorization (split operator method) were used to perform each iteration, our analytic formula (13) would still be 10–100 times faster.

How is steady-state space use controlled by α and L? Comparing the $L = 6$ row (cases b, f, j) to the preference function itself (a, e, i), we see that with this large length scale u^* is very close to proportional to w, as in 1D and as as explained in Sect. 2.1. Proceeding down the figure, we see smaller L values result in a u^* with an exaggerated tendency towards space use becoming increasingly concentrated in areas of higher preference. This results in much more relative animal concentration in the mesic grassland region than would be associated with traditional RSA treating all of Ω as available habitat. This tendency has reached its limit by the bottom $L = 0.1$ row (d, h, l), where u^* is close to proportional to w^2.

Consider the right-hand column (j, k, l). Changing from $L = 6$ to $L = 0.7$ causes a substantial increase in relative space use in the western part of the mesic grassland (see the large bump to the left in k), but very little change in the eastern part of this same habitat type. The explanation is simple: the grassland strip is generally wider than $L = 0.7$ on the western side resulting in a local tendency towards $u^* \propto w^2$, but narrower than $L = 0.7$ on the eastern side giving here $u^* \propto w$. Finally in case l $L = 0.1$ is narrower than all parts of the mesic grassland and the densities equalize on west and east sides. This is analogous to the transition discussed in Sect. 4.1.1. This interesting geometric effect is unaccounted for in traditional RSA.

4.2 Evolution in time

In Fig. 5 we show the time-dependent evolution of space use $u(x, t)$ under the master equation (2) in a single space dimension for simple piecewise linear model preference function with a single discontinuity (see panel a), with exponential kernel with length scale $L = 0.01$. Zero-flux boundary conditions were created by using the non-periodic

[3] We have also tried other model preference functions on larger grids: a 200×200 grid ($N = 40,000$) requires 0.025 s to find u^*. By contrast, note that solving for u^* without the factorization of K is impractical (even representing K as a dense N-by-N matrix on this grid would require 12 GB of memory).

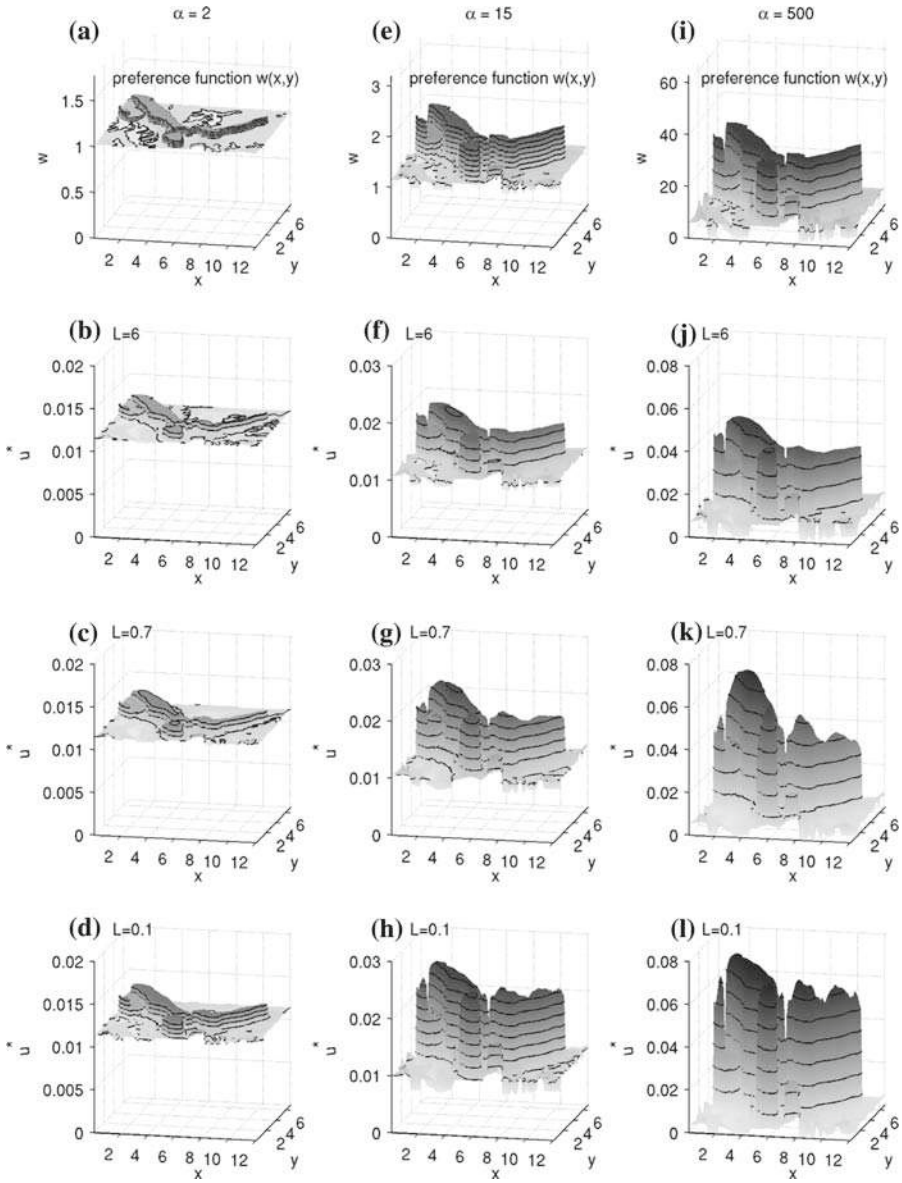


Fig. 4 Steady-state 2D pdfs u^* for the exponential jump kernel (20), applied to preference functions w derived from biomass data shown in Fig. 3. The three columns represent the choices $\alpha = 2, 15, 500$ in the preference model (22). At the top of each column is a surface plot of the preference function, followed below by surface plots of u^* for three decreasing values of L

jump kernel (19). Panels b-e show the resulting dynamics of space use, starting from an initial condition $u_0(x) = \delta(x - x_0)$ where $x_0 = 0.82$.

The computation with $N = 400$ grid points took 0.0005 s per time step; this was done with iterated multiplication by the dense K matrix since the non-periodic choice

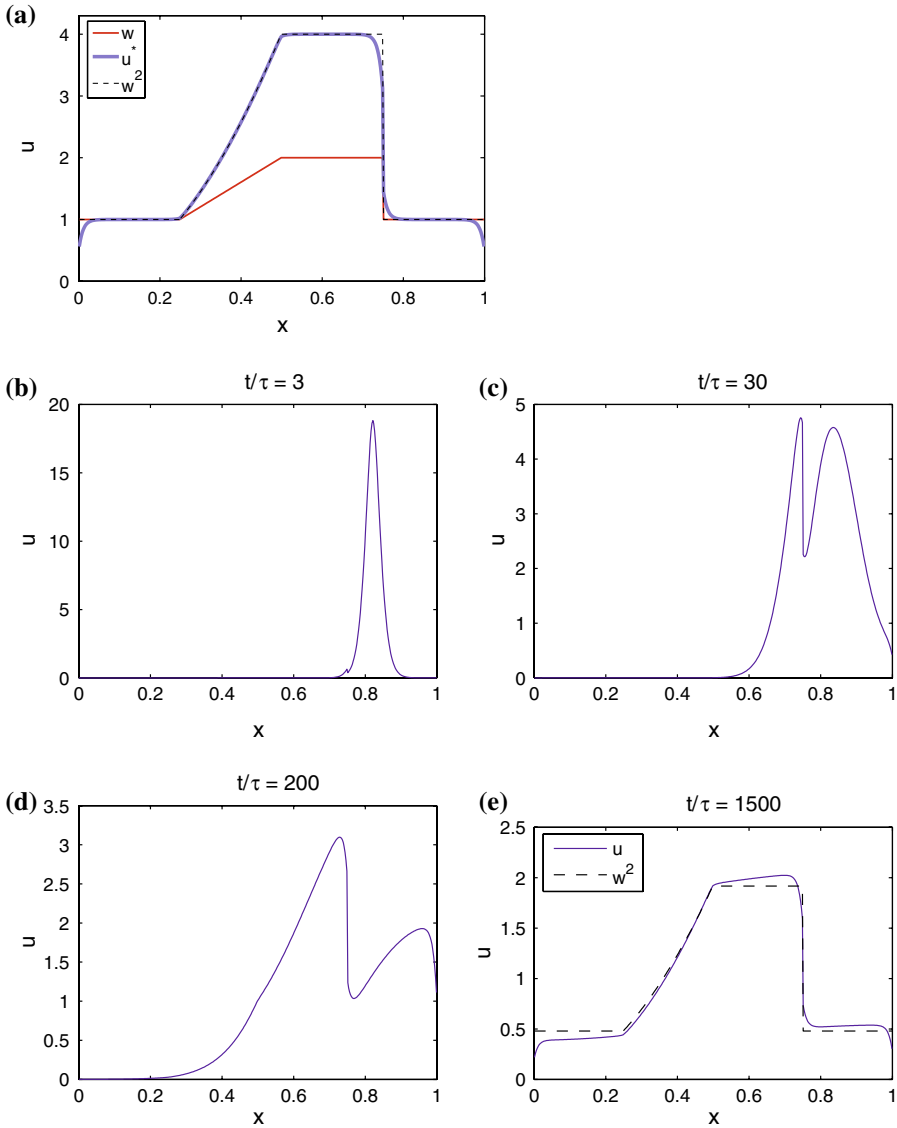


Fig. 5 **a** Piecewise linear model preference function $w(x)$, steady state pdf $u^*(x)$, and predicted limiting case of steady-state pdf (11) (*dashed*). **b–e** Snapshots of time evolution of $u(x,t)$ under master equation (2) in 1D, at four times (indicated by t/τ the number of iterations). The exponential jump pdf of (19) is used with $L = 0.01$. In **e** u is very close to steady-state; the prediction (11) is also shown (*dashed*)

of $\tilde{\phi}$ makes the Φ operator no longer a convolution. Notice the discontinuity feature in u at $x = 0.75$ establishes itself rapidly and persists throughout the full time range.

Figure 6 shows the time-dependent evolution of space use for the 2D real-world biomass preference model (22) with $\alpha = 500$ and $L = 0.7$ (same parameters as Fig. 4k), for a delta-function initial condition at $x_0 = (9, 4)$. After a single iteration

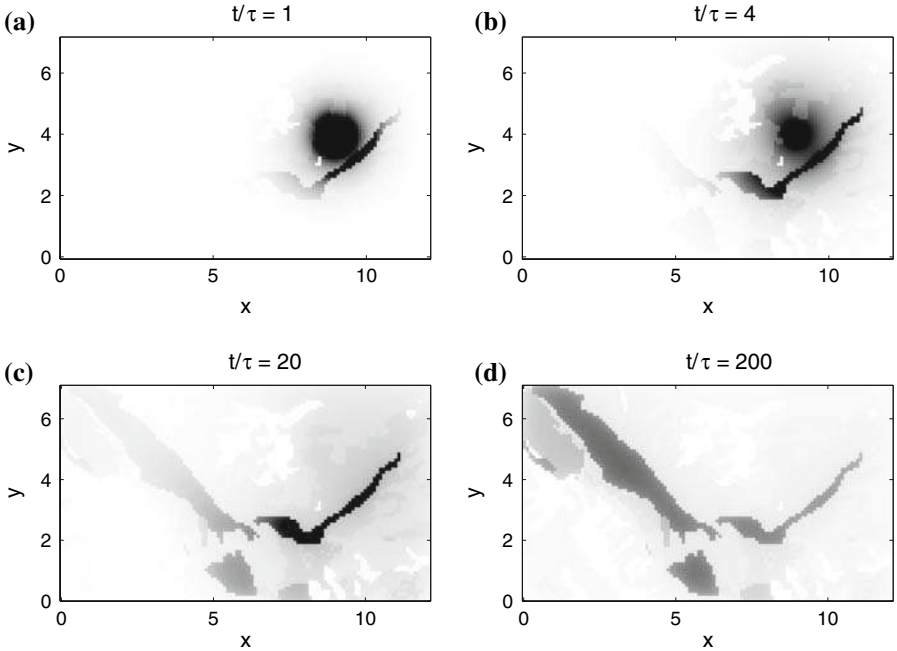


Fig. 6 Density plots snapshots of the time evolution of $u(x, t)$ under master equation (2) in 2D, at four times (indicated by t/τ the number of iterations). The exponential jump pdf of (20) is used with $L = 0.7$, and preference function derived from biomass data in Fig. 3 via Eq. (22) with $\alpha = 500$. In d) u is very close to steady-state, which is shown in Fig. 4k

(panel a), the local preference biases the individual's movements towards the narrow nearby eastern strip of mesic grassland. In panel (b) its space use is still split between an expanding radial distribution about its initial position and the nearby grassland. In panel (c), the compounded mechanistic movement steps have caused the individual's space use to become concentrated in the eastern strip of mesic grassland but the mesic grasslands in the western portion of the landscape are, at this stage, mostly unoccupied. Much later in the simulation however (panel d), the intensity of space use in the western mesic grassland areas is higher than the eastern mesic grasslands, due to the geometric effects on the steady-state u^* described in Sect. 4.1.2.

The simulation from which the snapshots in Fig. 6 were extracted is very rapid, animating smoothly in real time at 30 frames/s even though $N \approx 10^4$, allowing immediate interactive model exploration. The raw calculation (no graphical animation) takes 0.008 s per time-step, benefitting greatly from the split operator method using FFT convolution. Performing this without the aid of the factorization (15) is about 100 times slower. Furthermore, we find that the additional effort needed to extract the mean-square displacement $\langle x^2 \rangle(t) := \int_\Omega (x - x_0)^2 u(x, t) dx$, a useful measure of spreading, is negligible. We remark that in order to model zero-flux boundary conditions while still using FFT convolution, a periodized kernel was used but the w array was zero-padded with a border of width $O(L)$, at negligible extra cost.

5 Local behavior near a sharp habitat transition

We now examine in more detail what takes place at the discontinuities in preference function $w(x)$ that arise at the boundaries between different habitat types. As noted earlier, such discontinuities were not able to be treated in the derivation of the advection-diffusion limit in previous work [16].

For simplicity, we consider a single discontinuity on a 1D landscape, however we expect, and observe, similar behavior in the more biologically relevant case of multiple discontinuities arising at the edges of different habitat types on a 2D landscape. Consider a landscape in which there exists a single boundary between two habitat types located at $x = 0$, resulting in the following discontinues preference function:

$$w(x) = \begin{cases} 1, & x < 0 \\ w_0, & x > 0 \end{cases} \tag{23}$$

Given a jump kernel of width L it follows from (7) that $z(x)$ is a smoothed (mollified) step-function with transition region width L. The analytic formula (13) then tells us that the steady-state pdf jumps by a ratio of w_0 precisely at the habitat transition; however, when viewed on length scales larger than L, the pdf jumps by a ratio w_0^2. For the 1D exponential kernel (19), the analytic expression for the steady-state follows from that of z via (6), and is

$$u^*(x) = \begin{cases} 1 + \frac{W-1}{2} e^{x/L}, & x < 0 \\ W\left[W - \frac{W-1}{2} e^{-x/L}\right], & x > 0 \end{cases} \tag{24}$$

The spatial decay length is thus the same as that of the kernel ϕ. This is shown in Fig. 7a.

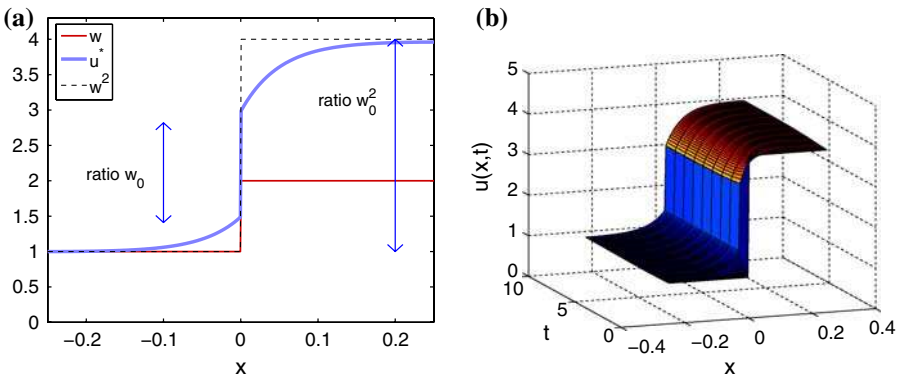


Fig. 7 **a** Detail of transition in steady-state space use occurring within distance L of a discontinuity in preference function of the form (23) with $w_0 = 2$, for the 1D exponential jump kernel of (19). **b** Time-evolution of $u(x, t)$ in this same local region for 10 time-steps of the master equation (2), starting from an initial pdf $u_0(x) = w(x)^2$

How fast is this equilibrium reached? We demonstrate in Fig. 7b that the transition region's shape reaches its approximate equilibrium in only a single time-step (a slight change also occurs over the next few time-steps). Here we chose initial conditions which already matched the global pdf ratio of w_0^2, in order to study the equilibration in this local region alone.

5.1 Effective boundary condition for advection–diffusion equation

Armed with this understanding of the local behavior at a preference function discontinuity, we can incorporate this into the Fokker–Planck PDE model for the evolution of $u(x,t)$ in the $\tau \to 0$ limit (in which case kernel width L must also go to zero at an appropriate rate). We remind the reader that the Fokker–Planck equation is

$$\frac{\partial}{\partial t}u = -\frac{\partial}{\partial x}[c(x)u] + \frac{\partial^2}{\partial x^2}[d(x)u] \tag{25}$$

where $c(x)$ and $d(x)$, representing drift and diffusion rates, take on values given by the τ-scaling of the second moment of the jump kernel ϕ [7,16,17].

We combine two observations: (i) locally the steady-state in the vicinity of a discontinuity in w enforces a multiplicative jump (the square of the w ratio) in u, and (ii) in the $\tau \to 0$ limit the Fokker–Planck equation evolves on much slower time-scales than the local equilibration in this vicinity (which happens in $O(\tau)$). Thus we expect that, for evolution on time-scales long relative to τ the transition region is in *local equilibrium*, with the effective boundary condition

$$\frac{u(x_-,t)}{(w_-)^2} = \frac{u(x_+,t)}{(w_+)^2}, \qquad \text{for all } t > 0, \tag{26}$$

where the subscripts $-$ and $+$ indicate limiting values on the left and right side of the discontinuity respectively. Similarly, by conservation of the flux $J(x) = \frac{\partial}{\partial x}[d(x)u] - c(x)u$ across the discontinuity, we must have that if $d(x)$ is continuous and $c(x) = 0$ then

$$\frac{\partial}{\partial x}u(x_-,t) = \frac{\partial}{\partial x}u(x_+,t), \qquad \text{for all } t > 0, \tag{27}$$

at a step discontinuity in w. We may now interpret (26) and (27) as *matching conditions* for coupled advection–diffusion equations on either side of the discontinuity. In this way we have a recipe to understand the diffusion limit even in the presence of discontinuous preference functions. We remark that our assumption of locally vanishing c corresponds to no gradient in w locally on either side (the case with general values of w' either side we postpone for future work). The above argument is non-rigorous, relying on reasoning based on separation of length- and time-scales. However it seems to be supported qualitatively by the evidence in Fig. 5b–e (although at early times (26) does not appear to hold accurately). We suggest that a more detailed analysis via matched asymptotic expansions should be carried out.

6 Conclusions

We have analyzed the mathematical properties of a mechanistic resource selection (MERSA) model that captures the influence of spatially localized habitat preference on the movement behavior of individuals and predicts their resulting patterns of space use. The model combines random foraging motion with a local sensitivity to habitat preference over a "perceptual radius" L. Directed movement bias is generated in a similar manner to the angle-biased (von Mises) jump kernels used in [17], and also becomes equivalent in the small-L limit to continuous-time advection–diffusion in a "confining potential" as in [8]. Our analysis shows that the model has a desirable factorization (15) which yields an unusually simple closed-form formula (13) for the steady state pdf u^*.

In this model, the effect of the compounded uncorrelated random movement decisions is to produce a space use whose scaling with the preference function w is geometry dependent: u^* is linear in w if L is large compared to local habitat features, but quadratic if L is small. This novel spatial effect has been analyzed previously only in the case of one-dimensional movement, in the context of the advection–diffusion limit in which the perceptual radius of individuals is small ($L \to 0$) [16]. The analytic results developed here extend this to any L, to the case of discontinuous preference functions (both in the steady-state and time-dependent cases), and to the more biologically relevent case of two dimensions.

This is expected to have important consequences for understanding resource selection behavior, one of which we now outline. By comparing observed animal habitat choices to those available in a buffer region, the RSA variants [1,4,6,9] discussed in the introduction implicitly assume a model for movement, similar to our proposed MERSA type. The RSA regression fits the preference function w within the buffer scale so that $u^* \propto w$. Then from our mathematical analysis we know that on larger spatial scales $u^* \propto w^2$, that is, the effects of preference are exaggerated. This indicates that it is crucial that RSA techniques take into account the correct spatial scale and distribution of animal movements if the preference function is to be meaningful: too large a buffer choice results in overestimation of strength of habitat preferences, as has been mentioned by other authors (e.g., see Step Selection Functions in [6]). We suggest that such interpretion problems could be bypassed by fitting the data with a unified model such as we propose, which includes both fine-scale movement decisions and preference bias. An example of this style of model-fitting and evidence comparison is Forester et al. [5]; it would be very interesting to combine the mechanistic aspects of our MERSA model with their state-space inference approach. We expect that by such model comparison, existing empirical movement models such as the multiscaled kernel function of [9] could be placed on a more solid foundation.

We have shown that the formula (13) and factorization property (15) of our model allows large gains in computational efficiency (when compared to a general redistribution operator). This includes the case of discontinuous 2D preference functions motivated by observations of spatially varying prey availability across discrete habitat types. We believe such efficient forward models will be important tools, as inverse modeling and fitting of multiple model parameters to data (discussed above) become more popular and numerically intensive.

Finally, although we did not construct or study them numerically in this work, we expect that the benefits demonstrated here for a model that is linear in u (with prey- or spatially dependent movement rates) will also carry over to the analysis of *non-linear* models such as those with density-dependent diffusion (e.g., Sect. 3 of [24]), and scent-mediated interactions between multiple animal packs [12,17,24]. For example, chemotaxis could be included in the preference function w, in which case our analytic formula (13) could be used to turn a coupled PDE system into coupled algebraic systems, a huge simplification. We also expect that by extending our preliminary analysis (Proposition 2), the operator spectral properties, and hence equilibration rates of animal home range space use, may be deduced.

Concise Matlab codes for computation of all figures in this work are freely available at http://math.dartmouth.edu/~ahb/moorcroft/

Acknowledgments AHB and PRM thank Robert L. Crabtree for the dataset on small mammal abundance in Yellowstone National Park, and the anonymous referees for their very useful comments and pointers to literature. AHB thanks Jonathan Goodman and Luc Rey-Bellet for useful discussions on the properties of Markov operators. AHB is partially funded by NSF-0507614.

References

1. Arthur, S.M., Manly, B.F.J., McDonald, L.L., Garner, G.W.: Assessing habitat selection when availability changes. Ecology **77**, 215–227 (1996)
2. Boyce, M.S., Mao, J.S., Merrill, E.M., Fortin, D.: Scale and heterogeneity in habitat selection by elk in Yellowstone National Park. Écoscience **10**, 421–431 (2002)
3. Boyce, M.S., McDonald, L.L.: Relating populations to habitats using resource selection functions. Trends Ecol. Evolut. **14**, 268–272 (1999)
4. Compton, B.W., Rhymer, J.M., McCollough, M.: Habitat selection by wood turtles (*clemmys insculpta*): and application of paired logistic regression. Ecology **83**, 833–843 (2002)
5. Forester, J.D., Ives, A.R., Turner, M.G., Anderson, D.P., Fortin, D., Beyer, H.L., Smith, D.W., Boyce, M.S.: State-space models link elk movement patterns to landscape characteristics in Yellowstone National Park. Ecol. Monogr. **77**, 285–299 (2007)
6. Fortin, D., Beyer, H.L., Boyce, M.S., Smith, D.W., Duchesne, T., Mao, J.: Wolves influence elk movements: behavior shapes a trophic cascade in Yellowstone National Park. Ecology **86**, 1320–1330 (2005)
7. Gardiner, C.W.: Handbook of stochastic methods for physics, chemistry, and the natural sciences, 3rd edn. Springer Series in Synergetics, vol. 13. Springer, Berlin (2004)
8. Giuggioli, L., Abramson, G., Kenkre, V.M., Parmenter, R.R., Yates, T.L.: Theory of home range estimation from displacement measurements of animal populations. J. Theor. Biol. **240**, 126–135 (2006)
9. Johnson, C.J., Parker, K.L., Heard, D.C., Gillingham, M.P.: A multiscale behavioral approach to understanding the movements of woodland caribou. Ecol. Appl. **12**, 1840–1860 (2002)
10. Johnson, D.H.: The comparison of usage and availability measurementts for evaluating resource preference. Ecology **61**, 65–71 (1980)
11. Kareiva, P.: Experimental and mathematical analyses of herbivore movement: Quantifying the influence of plant spacing and quality on foraging discrimination. Ecol. Monogr. **52**, 261–282 (1982)
12. Lewis, M.A., White, K.A.J., Murray, J.D.: Analysis of a model for wolf territories. J. Math. Biol. **35**, 749–774 (1997)
13. Manly, B.F.J., McDonald, L.L., Thomas, D.L., L., M.T., Erickson, W.P.: Resouce selection by animals: statistical design and analysis for field studies, 2nd edn. Kluwer, Dordrecht
14. Meyn, S.P., Tweedie, R.L.: Markov Chains and Stochastic Stability. Springer, London (1993)
15. Moorcroft, P.R.: Territoriality and carnivore home ranges. Ph.D. thesis, Dept. of Ecology and Evolutionary Biology, Princeton University (1997)
16. Moorcroft, P.R., Barnett, A.H.: Mechanistic home range models and resource selection analysis: a reconciliation and unification. Ecology (2007) (in press) arXiv:q-bio/0612008

17. Moorcroft, P.R., Lewis, M.A.: Mechanistic home range analysis. Monographs in population biology. Princeton University Press, Princeton (2006)
18. Moorcroft, P.R., Lewis, M.A., Crabtree, R.L.: Mechanistic home range models predict spatial patterns and dynamics of coyote territories in yellowstone. Proc. R. Soc. Ser. B **273**, 1651–1659 (2006)
19. Okubo, A.: Diffusion and ecological problems. Springer, Heidelberg (1980)
20. Othmer, H., Dunbar, S.R., Alt, W.: Models of dispersion in biological systems. J. Math. Biol. **26**, 263–298 (1988)
21. Patlak, C.: Random walk with persistence and external bias. Bull. Math. Biophys. **15**, 311–338 (1953)
22. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical recipes in C. Cambridge University Press, Cambridge (2002)
23. Turchin, P.: Translating foraging movements in heterogeneous environments into the spatial distribution of foragers. Ecology **72**, 1253–1266 (1991)
24. White, K.A.J., Lewis, M.A., Murray, J.D.: A model for wolf-pack territory formation and maintenance. J. Theor. Biol. **178**, 26–43 (1996)

Ecology, 89(4), 2008, pp. 1112–1119
© 2008 by the Ecological Society of America

MECHANISTIC HOME RANGE MODELS AND RESOURCE SELECTION ANALYSIS: A RECONCILIATION AND UNIFICATION

Paul R. Moorcroft[1,3] and Alex Barnett[2]

[1]*Department of Organismic and Evolutionary Biology, Harvard University, 22 Divinity Ave., Cambridge, Massachusetts 02138 USA*
[2]*Department of Mathematics, 6188 Kemeny Hall, Dartmouth College, Hanover, New Hampshire 03755 USA*

Abstract. In the three decades since its introduction, resource selection analysis (RSA) has become a widespread method for analyzing spatial patterns of animal relocations obtained from telemetry studies. Recently, mechanistic home range models have been proposed as an alternative framework for studying patterns of animal space-use. In contrast to RSA models, mechanistic home range models are derived from underlying mechanistic descriptions of individual movement behavior and yield spatially explicit predictions for patterns of animal space-use. In addition, their mechanistic underpinning means that, unlike RSA, mechanistic home range models can also be used to predict changes in space-use following perturbation. In this paper, we develop a formal reconciliation between these two methods of home range analysis, showing how differences in the habitat preferences of individuals give rise to spatially explicit patterns of space-use. The resulting unified framework combines the simplicity of resource selection analysis with the spatially explicit and predictive capabilities of mechanistic home range models.

Key words: home range analysis; mechanistic home range models; resource selection analysis; RSA; spatially explicit space-use.

Introduction

Since its introduction in the 1980s, resource selection analysis (RSA) has become a widespread method for identifying underlying environmental correlates of animal space-use patterns (Manly 1974, Johnson 1980). In contrast to earlier descriptive methods of home range analysis, such as the minimum convex polygon, bivariate normal, and kernel methods that simply summarize observed spatial patterns of relocations (for reviews, see Macdonald 1980, Worton 1987, Kernohan et al. 2001), conventional RSA uses a spatially implicit frequentist approach to identify habitats that are used disproportionately in relation to their occurrence.

Ratios of habitat utilization relative to habitat availability provide a simple estimate of habitat selection (Fig. 1). More commonly, however, resource selection models are specified in terms of the probability P_j of obtaining a relocation in a given habitat type j:

$$P_j = \frac{A_j w_j}{\sum_{k=1}^{n_{\text{hab}}} A_k w_k} \tag{1}$$

where w_j is the selective value of habitat j relative to other habitats ($j = 1 \ldots n_{\text{hab}}$, where n_{hab} is the number of habitat types) and A_j is the availability of habitat j on the landscape. The collection of resource selection values for all habitats within a landscape $\{w_j\}$ ($j = 1 \ldots n_{\text{hab}}$) is known as the resource selection function (Manly et al. 1993). Eq. 1 is generally preferred over simpler ratio-based estimates of habitat selectivity because of its improved statistical properties, having smaller variance and being less subject to bias (Arthur et al. 1996).

As results from numerous studies have shown, RSA can be used successfully to identify associations between animal space-use and habitat types as well as other forms of environmental heterogeneity, such as topography and resource availability, yielding insight into the underlying causes of animal space-use (for reviews, see Manly et al. 1993, Boyce and McDonald 1999, Cooper and Millspaugh 2001, Erickson et al. 2001).

More recently, mechanistic home range models have been proposed as an alternative framework for analyzing patterns of animal home ranges (Moorcroft and Lewis 2006, Moorcroft et al. 2006). In contrast to the spatially implicit, frequentist nature of RSA models such as Eq. 1, mechanistic home range models develop spatially explicit predictions for patterns of animal space-use, by, in the words of Millspaugh and Marzluff (2001), "modeling the movement process." Mathematically, this involves characterizing the fine-scale movement behavior of individuals as a stochastic movement process that is defined in terms of a redistribution kernel $k(\mathbf{x}, \mathbf{x}', \tau, t)$, where $k(\mathbf{x}, \mathbf{x}', \tau, t)\, d\mathbf{x}$ specifies the probability of an animal located at $\mathbf{x}'$ at time t moving to a location between location $\mathbf{x}$ and $\mathbf{x} + d\mathbf{x}$ in the time interval τ.

Relevant behavioral and ecological factors influencing the movement of individuals can be incorporated into

Manuscript received 31 December 2006; revised 28 June 2007; accepted 20 July 2007. Corresponding Editor: G. M. Henebry.

[3] E-mail: paul_moorcroft@harvard.edu

the redistribution kernel that defines the fine-scale stochastic movement process. For example, in a recent analysis of coyote home ranges in Yellowstone, Moorcroft et al. (2006) developed a "prey availability plus conspecific avoidance" (PA + CA) mechanistic home range model in which individuals exhibit: (1) an avoidance response to encounters with foreign scent marks (Fig. 2a, b), (2) an over-marking response to encounters with foreign scent marks, and (3) a foraging response to prey availability in which individuals decrease their mean step length in response to small-mammal abundance (Fig. 2a, c).

From this description of fine-scale movement behavior, it is then possible to derive probability density functions for the expected spatial pattern of home ranges that result from individuals moving on a landscape according to the underlying rules of movement. For example, in the case of the mechanistic home range model used by Moorcroft et al. (2006) to analyze coyote movements, the stochastic foraging response and responses to scent marks yield the following equations for the expected steady-state pattern of space-use:

$$\frac{\partial u^{(i)}(\mathbf{x},t)}{\partial t} = -\nabla\cdot[e^{-\alpha h(\mathbf{x})}\nabla u^{(i)}] + \nabla\cdot\left[e^{-\alpha h(\mathbf{x})}\beta(\mathbf{x})\vec{\mathbf{x}}_i u^{(i)}\sum_{j\neq i}^{n_{\text{pack}}} p^{(j)}\right] - \nabla\cdot[\,e^{-\alpha h(\mathbf{x})}u^{(i)}\nabla h] = 0 \qquad (2)$$

where $\nabla = (\partial/\partial x, \partial/\partial y)$, and the three bracketed terms on the right-hand side of the equation, in order, represent random motion, scent-mark avoidance, and directed movement toward areas of high resource density; and

$$\frac{dp^{(i)}(\mathbf{x},t)}{dt} = u^{(i)}\left[1 + m\sum_{j\neq i}^{n_{\text{pack}}} p^{(j)}\right] - p^{(i)} \qquad (3)$$

where the right-hand side of Eq. 3 represents the difference between scent-mark deposition and scent-mark decay. In Eqs. 2 and 3, $u^{(i)}(\mathbf{x}, t)$ is the expected space-use of individuals in pack i, $p^{(i)}(\mathbf{x}, t)$ is the scent-mark density of individuals in pack i, $h(\mathbf{x})$ is the spatial distribution of prey availability, and $\vec{\mathbf{x}}_i$ is a vector indicating the direction of the individual's home range center from its current position. The coefficients α and β of the space-use equation (Eq. 2) reflect the underlying characteristics of individual movement behavior. Specifically, $\alpha = 2\rho_1$, where ρ_1 is the mean step length of the individual, and β reflects the strength of the avoidance response to foreign scent marks (Fig. 2b, c). The parameter m of the scent-marking equation (Eq. 3) governs the strength of the over-marking response of individuals (i.e., the rate at which individuals increase their rate of scent marking following encounters with foreign scent marks relative to the background rate of marking of individuals). Fig. 3 shows the fit of the PA + CA space-use equations to the observed spatial distribution of relocations of five adjacent coyote packs in Yellowstone National Park home ranges. As the figure illustrates, the model captures the influences of resource availability and the presence of neighboring groups on the patterns of space-use within the region.

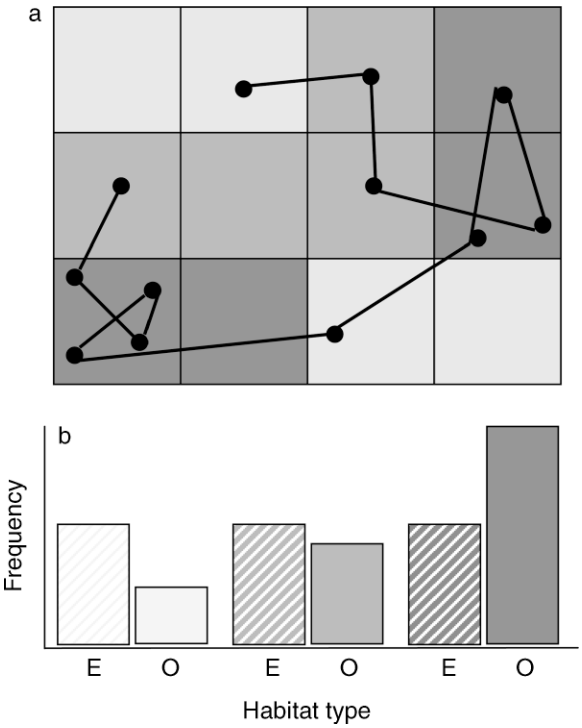


FIG. 1. Schematic illustrating the resource selection analysis (RSA) approach to analyzing patterns of animal space-use. (a) Shaded squares represent an idealized landscape comprising three equally abundant habitat types. Black lines represent the movement trajectory of an individual as it traverses the landscape, with points representing fixed-interval relocations of the individual. Because the three habitat types that comprise the landscape plotted in panel (a) are equally abundant, in the absence of preference, equal numbers of relocations would be expected to be obtained in each habitat, as indicated by the hatched bars in panel (b). The actual observed distribution of relocations, indicated by the solid bars in panel (b), shows that the individual exhibits a preference for the dark gray habitat type.

Mechanistic home range models address several limitations of RSA. First, as the fit of the PA + CA model shown in Fig. 3 illustrates, in addition to incorporating the effects of underlying landscape heterogeneities such as prey density, mechanistic home range models can also incorporate the influence of conspecifics that can significantly influence patterns of animal space-use. Second, a critical step in any RSA is to define the region within a landscape that constitutes available habitat. The spatially implicit, frequentist nature of conventional RSA models such as Eq. 1 means that all areas within the predefined region defined as "available habitat" are assumed to be equally accessible to individuals. On real landscapes, the patchy

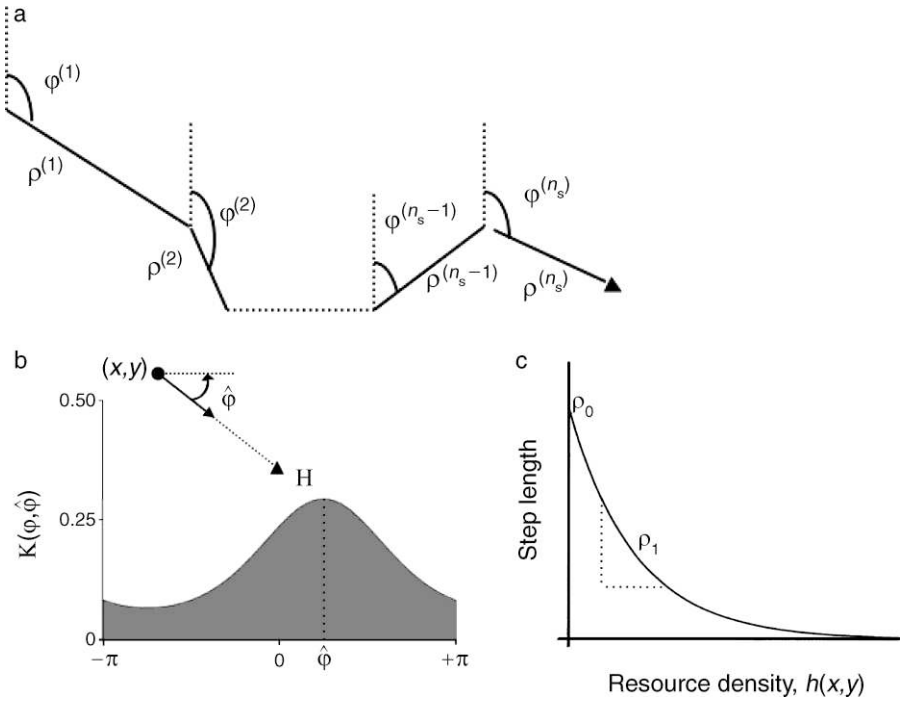


FIG. 2. (a) Schematic illustrating the model of individual movement behavior underlying mechanistic home range models. The movement trajectory of individuals is characterized as a stochastic movement process, defined in terms of sequences of movements ($i = 1 \ldots n_s$) [where n_s is the total number of steps] of distance $\rho^{(i)}$ and direction $\varphi^{(i)}$, drawn from statistical distributions of these quantities that are influenced by relevant factors affecting the movement behavior of individuals. Superscripts in parentheses indicate the step sequence. (b, c) Movement responses incorporated in the PA + CA (prey availability plus conspecific avoidance) mechanistic home range model used by Moorcroft et al. (2006) to characterize coyote home range patterns in Yellowstone. (b) Individuals display a conspecific avoidance response, in which encounters with foreign scent marks give rise to a nonuniform distribution of movement directions $K(\varphi)$, with individuals preferentially moving toward the center of their home range (H), whose direction from their current position is indicated by the angle $\hat{\phi}$. The magnitude of the avoidance response is governed by the density of foreign scent marks encountered and the value of an avoidance parameter β. (c) Consistent with observations of carnivore movement behavior, increasing prey availability causes the mean step length of individuals to decrease; ρ_0 is the mean distance between successive relocations in the absence of resources and ρ_1 governs the rate at which the mean step length of individuals decreases with increasing resource density, $h(x, y)$. The figure is from Moorcroft and Lewis (2006).

spatial distribution of habitats and resources such as that seen in Fig. 3, means that individuals are frequently required to traverse less favorable habitats in order to move between preferred areas. In conventional RSA, the time that individuals spend traversing unfavorable habitat registers as a degree of selection for the unfavorable habitat type (i.e., a nonzero probability that the individual chooses the unfavorable habitat at its next move) rather than as a constraint imposed by the spatial distribution of habitats on the landscape. In contrast, the mechanistic and spatially explicit nature of mechanistic home range models avoids the need to define available habitat, a priori, because the underlying model of individual movement behavior (Fig. 2) determines the likelihood and feasibility of an individual moving to a particular location, given its current position. Thus, by explicitly modeling the process of individual-level movement, mechanistic home range models naturally capture the influence of spatial constraints on patterns of space-use by individuals. Finally, their mechanistic nature means that mechanistic home range models can be used to predict patterns of space-use following perturbation. For example, Moorcroft et al. (2006) showed that their PA + CA model correctly captured the shifts in patterns of space-use that occurred following the loss of one of the packs in the study region. For further discussion of these issues, see Moorcroft et al. (2006) and Moorcroft and Lewis (2006).

Such considerations imply that resource selection analysis (RSA) and mechanistic home range models constitute two distinct frameworks for analyzing patterns of animal space-use. However, as we will show, some recent developments now permit a formal reconciliation and unification of these two seemingly disparate methods for analyzing patterns of animal space-use.

The first important development toward reconciling mechanistic home range models and RSA came in an analysis of polar bear relocations by Arthur et al. (1996), who argued that, rather than assuming a fixed measure of habitat availability across the entire study region, a more appropriate measure was the availability habitats

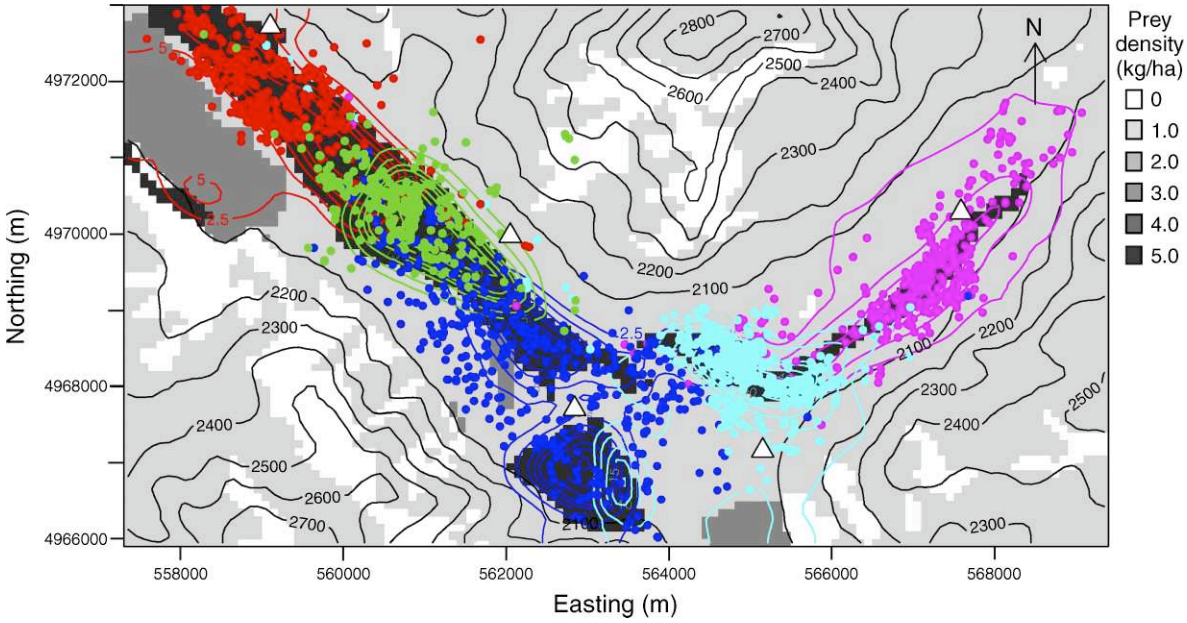


FIG. 3. Colored contour lines showing fit of the prey availability plus conspecific avoidance (PA + CA) home range model (Eqs. 2 and 3) to relocations (solid circles) obtained from five adjacent coyote packs in Lamar Valley Yellowstone National Park. The home range centers for each of the packs are also shown (open triangles), and the gray-scale background indicates small-mammal prey density in the different habitat types. Elevation contour lines (black) are in m a.s.l. The Akaike Information Criterion (AIC) score for the PA + CA home range model indicates that it gives a substantially improved goodness of fit to the observed spatial distribution of relocations compared to both an alternate model consisting of conspecific avoidance alone (ΔAIC = 1299.6) and an alternate model incorporating conspecific avoidance plus an avoidance response to steep terrain (ΔAIC = 456.9). The figure is redrawn from Moorcroft et al. (2006).

within a circle centered on the individual's current location, with radius R corresponding to the maximum distance that the individual was likely to travel in the time between successive relocations. They incorporated this into an RSA model by using a modified version of Eq. 1, in which habitat availability varies between relocations:

$$P_{ij} = \frac{A_{ij} w_j}{\sum_{k=1}^{n_{\text{hab}}} A_{ik} w_k} \tag{4}$$

where P_{ij} is the probability of choosing habitat j for the ith move, w_j is the habitat selection parameter for habitat j, and A_{ij} is the proportional availability of habitat j associated with relocation i, calculated as the fraction of the area within distance R of location i that is of type j.

The second development came in a paper by Rhodes et al. (2005), who proposed an extension of the approach of Arthur et al. (1996) to defining available habitat. Eq. 4 incorporates spatial variation in habitat availability; however, like Eq. 1, it is written in terms of the probability of observing a relocation of a given habitat type. Rhodes et al. (2005) recast the model of Arthur et al. (Eq. 4) in terms of the probability of an individual moving from location, given location a, to a subsequent location b:

$$P(a \to b) = \frac{\phi(a,b) \sum_{j=1}^{n_{\text{hab}}} w_j I(b,j)}{\sum_{j=1}^{n_{\text{hab}}} w_j \int_{I(c,j)=1} \phi(a,c)\, dc} \tag{5}$$

where $I(b, j)$ is an indicator function that takes the value 1 when location b is of type j and 0 otherwise, and $\phi(a, b)$ is given by

$$\phi(a,b) = \begin{cases} \dfrac{1}{\pi R^2} & \text{if } r_{ab} \le R \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where r_{ab} is the distance between locations a and b, and R is the maximum distance an individual is likely to travel between successive relocations, which defines the area of available habitat at location a.

The motivation of Rhodes et al. (2005) for casting the model of Arthur et al. (1996) into the form of Eq. 5 was twofold. First, they argued for a different functional form for $\phi(a, b)$ than Eq. 6, namely an exponential distribution $\phi(a, b) = \lambda \exp(-\lambda r_{ab})/2\pi r_{ab}$. Second, they introduced distance from the center of the individual's home range as a spatial covariate of the resource selection parameter; that is, w_j becomes a function $w_j(x)$. Rhodes et al. (2005) termed this a "spatially explicit habitat selection model" to reflect the spatially

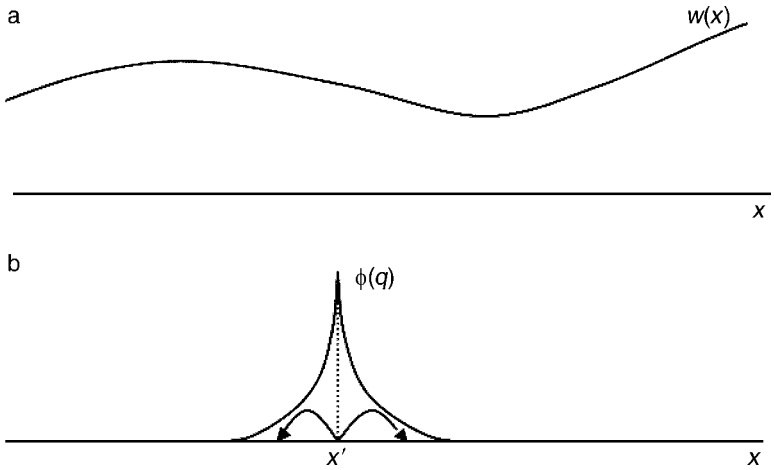


FIG. 4. Schematic illustrating a simple resource-selection-based mechanistic home range model for an animal moving in a single space dimension. Panel (a) shows the spatially dependent resource selection function $w(x)$. Panel (b) shows the individual's distribution of movement distances $\phi(q)$, where q is the displacement between the individual's current location x' and its subsequent location x.

varying resource selection coefficient. Note, however, that, unlike mechanistic home range models, resource selection models such as Eq. 5 do not directly yield predictions for actual patterns of space-use by individuals.

Switching from a model defined in terms of the probability of observing a relocation in a given habitat type to a model defined in terms of the probability of an individual moving between its current location and its subsequent location has, however, a third important consequence: Eq. 5, unlike Eq. 4, constitutes a redistribution kernel for the fine-scale movement behavior of individuals. In other words, RSA models of the form of Eq. 5 are, in effect, "modeling the movement process." This observation implies that it should be possible to establish a formal connection between RSA models of the form of Eq. 5 and corresponding mechanistic home range models. We will consider a simple pedagogical example demonstrating that this is indeed the case.

ANALYSIS

Consider an individual living on a one-dimensional landscape whose relative preference for different habitats can be expressed by a resource selection function $w(x)$ (Fig. 4a). Suppose further that, in the absence of habitat preference (i.e., $w(x)$ is constant), the individual moves to the right or left of its current position during time interval τ, with a distribution of displacements $\phi_\tau(q)$, where $q = x - x'$ is the displacement between the individual's current location x' and its subsequent location x (Fig. 4b). Note that, in contrast to Moorcroft et al. (2006), where the fine-scale movement behavior in two dimensions is described in terms of a sequence of movements of length ρ_τ and direction ϕ_τ ($i = 1 \ldots n_s$) (where n_s is the number [sequence] of steps; see Fig. 2), here we describe the fine-scale movement behavior of an individual moving in a single space dimension in terms of a sequence of displacements q_τ ($i = 1 \ldots n_s$) that have both a magnitude and a sign. Because $\phi_\tau(q)$ is a probability density function, $\int_{-\infty}^{\infty} \phi_\tau(q)\, dq = 1$.

The probability density of the individual moving to location x from its initial location x', during time interval τ, in this landscape with varying preference is then given by our model redistribution kernel:

$$P(x' \to x) = k_\tau(x, x') = \frac{\phi(x - x')w(x)}{\int_{-\infty}^{\infty} \phi(x'' - x')w(x'')\, dx''}. \quad (7)$$

Note that Eq. 7 has the same form as Eq. 5, with

$$w(x) = \sum_{j=1}^{n_{\text{hab}}} w_j I(x, j).$$

Note that the probability of moving from x' to x in the absence of habitat preference is determined only by the difference between x and x' and that the preference function w is evaluated at the location to which the individual moves, rather than its current location. In this example, we assume that an individual's redistribution kernel does not vary in time, and thus the dependency on time t can be dropped.

Defining $u(x, t)\, dx$ as the probability that the individual is located between x and $x + dx$ at time t, we can write an equation that summarizes all the possible ways that an individual located at x' can arrive within the interval $(x, x + dx)$ at time $t + \tau$:

$$u(x, t + \tau) = \int_{-\infty}^{\infty} k_\tau(x, x')u(x', t)\, dx'. \quad (8)$$

Eq. 8 is converted into an equation for the expected pattern of space-use by the individual by expanding the right-hand side using a Taylor series and then considering the limit as $\tau \to 0$, yielding the following

advection–diffusion equation:

$$\frac{\partial u(x,t)}{\partial t} = -\frac{\partial}{\partial x}[c(x)u(x,t)] + \frac{\partial^2}{\partial x^2}[d(x)u(x,t)] \quad (9)$$

where the advection and diffusion coefficients, $c(x)$ and $d(x)$, respectively, are given by

$$c(x) = \lim_{\tau\to 0}\frac{1}{\tau}\int_{-\infty}^{\infty}(x-x')k_\tau(x,x')\,dx' \quad (10a)$$

and

$$d(x) = \lim_{\tau\to 0}\frac{1}{2\tau}\int_{-\infty}^{\infty}(x-x')^2k_\tau(x,x')\,dx'. \quad (10b)$$

Details of the derivation can be found in the Appendix. Because $u(x, t)$ is a probability density function, the normalization

$$\int_\Omega u(x,t)\,dx = 1 \quad (11)$$

where Ω is the region over which the individual is able to move, is preserved for all future times t.

Inserting Eq. 7 into Eqs. 10a, b (see Appendix) yields the following equations for the coefficients d and c:

$$c(x) = \lim_{\tau\to 0}\frac{M_2(\tau)}{\tau}\frac{w_x(x)}{w(x)} \quad (12a)$$

$$d(x) = \lim_{\tau\to 0}\frac{M_2(\tau)}{2\tau} \quad (12b)$$

where the second moment is $M_2(\tau) = \int \rho^2\phi_\tau(x)\,dx$ and $w_x = dw/dx$.

Thus we see that a simple, spatially explicit resource-selection model yields an advection–diffusion equation (Eq. 9) for the expected location of an individual. Note that while the advection term (Eq. 12a) varies in space, the magnitude of the diffusion coefficient (Eq. 12b) is constant.

Inserting Eqs. 12a, b into Eq. 9, we can derive an approximation for the expected steady-state pattern of space-use $u^*(x) = \lim_{t\to\infty} u(x, t)$. The result for the case of a smooth continuous preference function $w(x)$ is

$$u^*(x) = \frac{1}{W_0}w(x)^2 \quad (13)$$

with the normalization constant $W_0 = \int_\Omega w(x)^2\,dx$. The details of the derivation are given in the Appendix. In other words, the steady-state pattern of space-use by an individual is given by the normalized square of its resource selection function $w(x)$.

Fig. 5a shows a plot of Eq. 13 and a numerical solution of Eqs. 7 and 8 for the case of an individual that, in the absence of habitat preference, moves with an exponential distribution of step lengths and with an equal probability of moving in either direction. Mathematically, this is equivalent to an individual with a Laplace distribution of displacement distances q.

As can be seen in Fig. 5a, Eq. 13 captures the pattern of space-use arising from the underlying habitat preferences of the individual.

Eq. 13 is an approximation that technically holds only when variation in $w(x)$ is at spatial scales that are large relative to the characteristic width of the individual's distribution of displacements $\phi_\tau(q)$. Fig. 5b shows a case in which the individual's resource selection function is discontinuous. In this case, Eq. 13 does not accurately capture the pattern of space-use in the region of the discontinuity; however, the errors are localized, and thus Eq. 13 still reasonably describes the overall pattern of space-use.

Discussion

The analysis presented here demonstrates that resource selection models of the form of Eq. 5 proposed by Rhodes et al. (2005) constitute an underlying stochastic movement process, and thus can be used to formulate corresponding mechanistic home range models that predict the expected patterns of space-use resulting from the underlying habitat preferences. As we would expect, increasing preference for a given habitat type (higher $w(x)$) gives rise to increasing space-use in the preferred habitats relative to the less preferred habitats (Fig. 5).

What is surprising, however, is that the relative intensity of space-use by an individual at a given location is governed by the square of its preference function for that location. Although Eq. 13 is an approximation that technically holds only when the preference function varies on spatial scales larger than the individual's redistribution kernel, as seen in Fig. 5b, it still captures the overall pattern space of space-use even when an individual's resource selection function changes rapidly in space. The mathematical explanation for this nonintuitive result is that the equilibrium pattern of space-use $u^*(x)$ is governed by the relative strength of the advection term relative to magnitude of the diffusion term, the squared term arising because of the factor of 2 in the denominator of the diffusion coefficient Eq. 12b (see Appendix for further details).

A somewhat more biological explanation is as follows. When viewed in terms of individual movement behavior, resource selection models of the form of Eq. 5 reflect a specific assumption about how environmental factors such as habitat type influence an animal's movement: the influence of habitat preference occurs by generating a differential probability of an individual moving in a given movement direction. The resulting bias in the individual's distribution of movement directions gives rise to the positive advection term (Eq. 12a), whose sign and magnitude is governed by the relative gradient in the animal's resource selection function, with preferential movement toward preferred habitat types. It is this preferential movement up gradients of habitat preference relative to the random component of the individual's motion (represented by the diffusion term) that

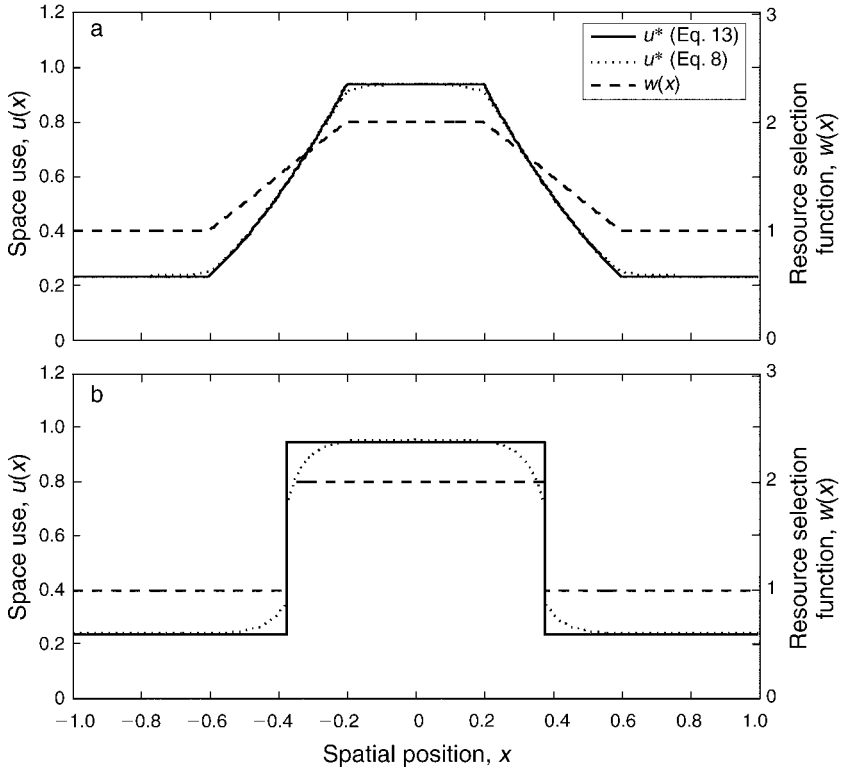


FIG. 5. Patterns of space-use arising for an individual with a resource selection function that is (a) piecewise continuous or (b) discontinuous, and who moves to the left or right with equal probability with an exponential distribution of step lengths in the absence of habitat preference [i.e., $\phi_\tau(q) = (1/2l)\exp(-|q|/l)$, with $l = 0.05$ for the example shown]. In both (a) and (b), the resource selection function $w(x)$ (dashed line), the predicted steady-state pattern of space-use $u^*(x)$ given by Eq. 13 (solid line), and $u^*(x)$ calculated by a numerical steady-state solution of Eq. 8 (dotted line) are plotted. When $w(x)$ is piecewise continuous (panel a), the solution of Eq. 13 closely approximates the steady-state pattern of space-use. When $w(x)$ is discontinuous (panel b), the solution of Eq. 13 fails to capture the pattern of space-use in the neighborhood of the discontinuities but still gives a reasonable match to the overall pattern of space-use.

determines the individual's equilibrium pattern of space-use $u^*(x)$.

In mechanistic home range models, environmental and biological factors affecting an individual's movement behavior can influence not just its distribution of movement directions, but also its distribution of movement distances. For example, in the PA + CA model used by Moorcroft et al. (2006) to analyze coyote home range patterns in Yellowstone (see *Introduction*), encounters with foreign scent marks alter an individual's distribution of movement directions (Fig. 2b). However, consistent with observations of coyote foraging responses (Laundre and Keller 1981), prey availability does not influence an individual's distribution of movement directions, but instead influences its distribution of movement distances. Specifically, the mean step length of individuals declines as an exponential function of the prey availability encountered (Fig. 2c). These two qualitatively different forms of movement response give rise to different terms in the equations for expected patterns of space-use: conspecific avoidance response gives rise to an advective term, directed movement toward the home range center, whereas the foraging response gives rise to a spatially varying diffusion term (see Eq. 2). The influence of spatial variation in step length on patterns of space-use has been explored by a number of authors, including Okubo (1980), Kareiva and Odell (1987), and Turchin (1991, 1998).

The above discussion implies that current (RSA) models embody one aspect of how animals can respond to factors affecting their movement, namely, by changing their distribution of movement directions. Changes in an animal's distribution of movement distances, such as the foraging response in the PA + CA model of Moorcroft et al. (2006), constitute a second mechanism by which animals respond to their environment, one that is not represented in resource selection models such as Eq. 5. These two qualitatively different forms of movement response (preferential movement in particular directions and spatial variation in mean-squared displacement) together determine the relative intensity of space-use in different areas.

Finally, although the example considered here of an individual moving on a one-dimensional landscape is

clearly idealized, previous analyses (Moorcroft and Lewis 2006) suggest that the reconciliation between a model of resource selection and a corresponding mechanistic home range model demonstrated here will also apply in the more biologically relevant case of individuals moving on two-dimensional landscapes consisting of multiple habitat types. In the long term, this ability to translate resource-selection-based analyses of patterns of animal relocations into corresponding mechanistic home range models offers the promise of a unified framework for analyzing patterns of animal space-use.

Acknowledgments

The authors thank Mark Lewis, Robert Crabtree and two anonymous referees for their comments and suggestions on this manuscript.

Literature Cited

Arthur, S. M., B. F. J. Manly, L. L. McDonald, and G. W. Garner. 1996. Assessing habitat selection when availability changes. Ecology 77:215–227.

Boyce, M. S., and L. L. McDonald. 1999. Relating populations to habitats using resource selection functions. Trends in Ecology and Evolution 14:268–272.

Cooper, A. B., and J. J. Millspaugh. 2001. Accounting for variation in resource availability and animal behavior in resource selection studies. Pages 246–273 *in* J. Millspaugh and J. M. Marzluff, editors. Radio tracking and animal populations. Academic Press, San Diego, California, USA.

Erickson, W. P., T. L. McDonald, K. G. Gerow, S. Howlin, and J. W. Kern. 2001. Statistical issues in resource selection studies with radio marked animals. Pages 211–242 *in* J. Millspaugh and J. M. Marzluff, editors. Radio tracking and animal populations. Academic Press, San Diego, California, USA.

Johnson, D. 1980. The comparison of usage and availability measurements for evaluating resource preference. Ecology 61:65–71.

Kareiva, P., and G. Odell. 1987. Swarms of predators exhibit "preytaxis" if individuals exhibit area-restricted search. American Naturalist 130:233–270.

Kernohan, B. J., R. A. Gitzen, and J. J. Millspaugh. 2001. Analysis of animal space-use and movements. Pages 125–166 *in* J. Millspaugh and J. M. Marzluff, editors. Radio tracking and animal populations. Academic Press, San Diego, California, USA.

Laundre, J., and B. Keller. 1981. Home range use by coyotes in Idaho. Animal Behavior 29:449–461.

Macdonald, D. W. 1980. Patterns of scent marking with urine and faeces amongst carnivore populations. Symposium of the Zoological Society of London 45:107–139.

Manly, B. 1974. A model for certain types of selection experiments. Biometrics 30:281–294.

Manly, B., L. McDonald, and D. Thomas. 1993. Resource selection by animals: statistical design and analysis for field studies. Chapman and Hall, New York, New York, USA.

Millspaugh, J. J., and J. M. Marzluff. 2001. Radio tracking and animal populations. Academic Press, San Diego, California, USA.

Moorcroft, P. R., and M. A. Lewis. 2006. Mechanistic home range analysis. Princeton University Press, Princeton, New Jersey, USA.

Moorcroft, P. R., M. A. Lewis, and R. L. Crabtree. 2006. Mechanistic home range models capture spatial patterns and dynamics of coyote territories in Yellowstone. Proceedings of the Royal Society B 273:1651–1659.

Okubo, A. 1980. Diffusion and ecological problems: mathematical models. Springer-Verlag, Berlin, Germany.

Rhodes, J. R., C. A. McAlpine, D. Lunney, and H. P. Possingham. 2005. A spatially explicit habitat selection model incorporating home range behavior. Ecology 86: 1199–1205.

Turchin, P. 1991. Translating foraging movements in heterogeneous environments in the spatial distribution of foragers. Ecology 73:1253–1266.

Turchin, P. 1998. Quantitative analysis of movement: measuring and modeling population redistribution in animals and plants. Sinauer, Sunderland, Massachusetts, USA.

Worton, B. 1987. A review of models of home range for animal movement. Ecological Modelling 38:277–298.

APPENDIX

Derivation of the Fokker-Planck equation for space-use (*Ecological Archives* E089-067-A1).

Appendix

Derivation of the Fokker-Planck equation for space-use

Here we derive equation Eqs. (9)-(10) from Eq. (8), for the case of general redistribution kernel $k_\tau(x, x')$[1]. First we note that in order for $u(x,t)$ to remain a probability density function, the kernel must satisfy

$$\int k_\tau(x, x')dx = 1, \qquad \text{for all } x'. \tag{A1}$$

Recalling the definition $q := x - x'$, we now change the variables used to describe the kernel from (x, x') to (q, x'), by defining

$$\kappa_\tau(q, x') := k_\tau(x' + q, x'). \tag{A2}$$

The purpose of this is to enable us to hold q constant while expanding a Taylor series in x'. Writing Eq. (8) in terms of this new kernel, then changing integration variable from x' to q gives

$$\begin{aligned} u(x, t+\tau) &= \int \kappa_\tau(x - x', x')u(x', t)dx' \\ &= \int \kappa_\tau(q, x - q)u(x - q, t)dq \\ &= \int \left[\kappa_\tau(q, x)u(x, t) - q\frac{\partial}{\partial x}[\kappa_\tau(q, x)u(x, t)] + \frac{q^2}{2!}\frac{\partial^2}{\partial x^2}[\kappa_\tau(q, x)u(x, t)]\cdots\right]dq. \end{aligned} \tag{A3}$$

The final step is achieved by considering the integrand as a function of two variables, q and $(x - q)$, and Taylor expanding this function with respect to the second variable about the value x, while holding the first constant.

Dividing (A3) by τ, making use of $\int \kappa_\tau(q, x)dq = 1$ which is a re-statement of Eq. (A1), and switching the order of differentiation and integration we get

$$\frac{u(x, t+\tau) - u(x, t)}{\tau} = -\frac{1}{\tau}\frac{\partial}{\partial x}\int q\kappa_\tau(q, x)dq \;\; u(x, t) + \frac{1}{2\tau}\frac{\partial^2}{\partial x^2}\int q^2\kappa_\tau(q, x)dq \;\; u(x, t)\cdots \tag{A4}$$

Taking the limit of small time interval τ gives

$$\begin{aligned} \frac{\partial u(x, t)}{\partial t} &= -\frac{\partial}{\partial x}\left[\lim_{\tau\to 0}\frac{1}{\tau}\left(\int_{-\infty}^{\infty} q\kappa_\tau(q, x)dq\right) \; u(x, t)\right] \\ &+ \frac{\partial^2}{\partial x^2}\left[\lim_{\tau\to 0}\frac{1}{2\tau}\left(\int_{-\infty}^{\infty} q^2\kappa_\tau(q, x)dq\right) \; u(x, t)\right] - \ldots \end{aligned} \tag{A5}$$

[1]This is essentially the derivation of a Fokker-Planck equation from a Chapman-Kolmogorov master equation via the Kramers-Moyal expansion – see Gardiner 1983.

The two integrals in the first and second terms are, respectively, the first and second moments of the kernel κ_τ. Considering the limit of small time-steps ($\tau \to 0$), and making the conventional assumptions that the first two moments scale with order τ and that higher order terms can be discarded, we arrive at Eq (9),

$$\frac{\partial u(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left[c(x)u(x,t)\right] + \frac{\partial^2}{\partial x^2}\left[d(x)u(x,t)\right], \tag{A6}$$

where

$$\begin{aligned} c(x) &= \lim_{\tau\to 0}\frac{1}{\tau}\int_{-\infty}^{\infty} q\kappa_\tau(q,x)dq, \quad \text{and} \\ d(x) &= \lim_{\tau\to 0}\frac{1}{2\tau}\int_{-\infty}^{\infty} q^2\kappa_\tau(q,x)dq. \end{aligned} \tag{A7}$$

By substituting for q in the above expressions we obtain Eq. (10).

Coefficients of the space-use equation arising from a spatially-dependent resource selection model

The redistribution kernel for the spatially-explicit resource selection model (Eq.7) is given by

$$k_\tau(x,x') = \frac{\phi_\tau(x-x')w(x)}{\int \phi_\tau(x''-x')w(x'')dx''} \tag{A8}$$

where $\phi_\tau(x-x')dx'dx$ the probability that an individual located between x' and $x'+dx$ will move to a location between x and $x+dx$ away from x, and $w(x)$ is a resource selection function. The denominator is a normalizing factor that ensures that Eq. (A1) holds.

We consider the case of continuous, sufficiently smooth positive preference function $w(x)$, and a bounded symmetric distribution of displacement distances $\phi_\tau(q)$ (Figure 3). Recalling the definition (A2), we can rewrite (A8), as

$$\kappa_\tau(q,x) = \frac{\phi_\tau(q)w(x+q)}{\int \phi_\tau(q')w(x+q')dq'} \tag{A9}$$

The pth moment of the distribution of displacement distances is

$$M_p(\tau) := \int q^p\phi_\tau(q)dq. \tag{A10}$$

Since ϕ_τ is a probability density function $M_0(\tau) = 1$, and since it is symmetric $M_1(\tau) = 0$ and all higher odd moments are zero. $M_2(\tau)$ is the variance, and we make the conventional assumption that in the limit $\tau \to 0$ the higher even moments can be neglected (*i.e.* they vanish faster than linearly in τ).

Under the above assumptions, Taylor expanding w yields the following expression for the denominator of (A9),

$$\int \phi_\tau(q')w(x+q')dq' = w(x) + \frac{w_{xx}(x)}{2!}M_2(\tau) + \cdots \tag{A11}$$

where $w_{xx} = \frac{d^2w}{dx^2}$. In the limit $\tau \to 0$, only the first term of Eq. A11 remains. Using this expression for the denominator and inserting (A9) into Eqs. (A7) yields

$$\begin{aligned} c(x) &= \lim_{\tau\to 0} \frac{1}{\tau} \frac{w_x(x)M_2(\tau) + w_{xxx}(x)M_4(\tau)/3! + \cdots}{w(x) + w_{xx}(x)M_2(\tau)/2! + \cdots} \\ &= \lim_{\tau\to 0} \frac{M_2(\tau)}{\tau} \cdot \frac{w_x(x)}{w(x)} \end{aligned} \tag{A12}$$

where $w_x = \frac{dw}{dx}$, etc. Similarly,

$$\begin{aligned} d(x) &= \lim_{\tau\to 0} \frac{1}{2\tau} \frac{w(x)M_2(\tau) + w_{xx}(x)M_4(\tau)/2! \cdots}{w(x) + w_{xx}(x)M_2(\tau)/2 + \cdots} \\ &= \lim_{\tau\to 0} \frac{M_2(\tau)}{2\tau}. \end{aligned} \tag{A13}$$

This is Eq (12a,b).

As two examples, we consider two special cases of the distribution of displacement distances. Both contain a length-scale L which we choose to scale as $L = \tau^{1/2}$. The distributions are (i) the fixed step length $\phi_\tau(q) = (\delta(q-L) + \delta(q+L))/2$ for which $\lim_{\tau\to 0} \frac{M_2(\tau)}{\tau} = 1$, and, (ii) the Laplace distribution (Figure 3) $\phi_\tau(q) = (1/2L)e^{-|q|/L}$, which gives $\lim_{\tau\to 0} \frac{M_2(\tau)}{\tau} = 2$. As is clear from (A12) and (A13), it is only the value of the second moment of ϕ_τ that determines the magnitude of the advection and diffusion coefficients.

Steady-state distribution of the resource selection model

The steady state distribution u^* for Equation (9) can be derived by first expressing the equation in conservation law form

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left[(du)_x - cu\right] = 0. \tag{A14}$$

where the quantity in square brackets is the flux. Setting the time derivative to zero implies that the steady-state flux is a constant independent of spatial position. Since by definition the steady-state pattern of space-use u^* does not change with time, this constant is zero. Thus in the steady state $(du)_x = cu$, which may be integrated to give the steady state pattern of space use

$$u^*(x) = \frac{C}{d(x)} \exp \int^x \frac{c(x')}{d(x')} dx' \tag{A15}$$

with C chosen to so that $u^*(x)$ integrates to 1. Substituting (A12) and (A13) into A15 gives Eq. (13).

Note that if the individual is moving in a finite, bounded region, then zero-flux boundary conditions apply to (A6) at the edges of the domain, *i.e.*

$$\frac{\partial}{\partial x}\left[(d(x)u(x,t)\right] - c(x)u(x,t) = 0, \qquad \text{at} \quad \partial\Omega \tag{A16}$$

In this case a nonzero steady-state solution Eq. (A15) is guaranteed to exist for any continuous $w(x)$.

Quantum mushroom billiards

Alex H. Barnett[a)]
Department of Mathematics, Dartmouth College, Hanover, New Hampshire 03755, USA

Timo Betcke[b)]
School of Mathematics, The University of Manchester, Manchester, M13 9PL, United Kingdom

(Received 28 September 2006; accepted 1 November 2007; published online 12 December 2007)

We report the first large-scale statistical study of very high-lying eigenmodes (quantum states) of the mushroom billiard proposed by L. A. Bunimovich [Chaos **11**, 802 (2001)]. The phase space of this mixed system is unusual in that it has a single regular region and a single chaotic region, and no KAM hierarchy. We verify Percival's conjecture to high accuracy (1.7%). We propose a model for dynamical tunneling and show that it predicts well the chaotic components of predominantly regular modes. Our model explains our observed density of such superpositions dying as $E^{-1/3}$ (E is the eigenvalue). We compare eigenvalue spacing distributions against Random Matrix Theory expectations, using 16 000 odd modes (an order of magnitude more than any existing study). We outline new variants of mesh-free boundary collocation methods which enable us to achieve high accuracy and high mode numbers ($\sim 10^5$) orders of magnitude faster than with competing methods. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2816946]

Quantum chaos is the study of the quantum (wave) properties of Hamiltonian systems whose classical (ray) dynamics is chaotic. Billiards are some of the simplest and most studied examples; physically their wave analogs are vibrating membranes, quantum, electromagnetic, or acoustic cavities. They continue to provide a wealth of theoretical challenges. In particular "mixed" systems, where ray phase space has both regular and chaotic regions (the generic case), are difficult to analyze. Six years ago Bunimovich described[1] a mushroom billiard with simple mixed dynamics free of the usual island hierarchies of Kolmogorov-Arnold-Moser (KAM). He concluded by anticipating the *growth of "quantum mushrooms*;" it is this gardening task that we achieve here, by developing advanced numerical methods to collect an unprecedented large number n of eigenmodes (much higher than competing numerics[2] or microwave studies[3]). Since uncertainties scale as $n^{-1/2}$, a large n is vital for accurate spectral statistics and for studying the semiclassical (high eigenvalue) limit. We address three main issues: (i) The conjecture of Percival[4] that semiclassically modes live exclusively in invariant (regular or chaotic) regions, and occur in proportion to the phase space volumes. (ii) The mechanism for dynamical tunneling, or quantum coupling between classically isolated phase space regions. (iii) The distribution of spacings of nearest-neighbor eigenvalues, about which recent questions have been raised.[3] We show many pictures of modes, including the boundary phase space (the so-called Husimi function).

I. INTRODUCTION

The nature of eigenfunctions of linear partial differential operators in the short wavelength, or semiclassical limit remains a key open problem which continues to engage mathematicians and physicists alike. When the operator is the quantization of a classical Hamiltonian dynamical system, the behavior of eigenfunctions depends on the class of dynamics. In particular, hyperbolic dynamics (exponential sensitivity to initial conditions, or chaos) leads to irregular eigenfunctions, the study of which forms the heart of a field known as "quantum chaos"[5] or "quantum ergodicity."[6,7] The planar billiard, or particle undergoing elastic reflection in a cavity $\Omega \subset \mathbb{R}^2$, is one of the simplest examples. Billiards exhibit a menagerie of dynamical classes[8] ranging from complete integrability (ellipses and rectangles) to complete ergodicity (e.g., Sinai billiard[9]). Bunimovich introduced the "mushroom" billiard[1,10] with the novelty of a well-understood divided phase-space comprising a single integrable (KAM) region and a single ergodic region. (We note that a related Penrose-Lifshits mushroom construction[51] continues to find use in isospectral problems.[52]) As seen in Fig. 1(a), the mushroom is the union of a half-disk (the "hat") and a rectangle (the "foot"); only trajectories reaching the foot are chaotic. The simplicity of its phase space has allowed analysis of phenomena such as "stickiness" (power-law decay of correlations) in the ergodic region.[11,12]

The quantum-mechanical analog of billiards is the spectral problem of the Laplacian in Ω with homogeneous boundary conditions (BCs). Choosing Dirichlet BCs (and units such that $\hbar = 2m = 1$) we have

$$-\Delta \phi_j = E_j \phi_j \quad \text{in } \Omega, \tag{1}$$

$$\phi_j = 0 \quad \text{on } \partial\Omega. \tag{2}$$

This "drum problem" has a wealth of applications throughout physics and engineering.[13] Eigenfunctions (or eigenmodes, modes) ϕ_j may be chosen to be real-valued and orthonormalized, $\langle \phi_i, \phi_j \rangle := \int_\Omega \phi_i(\mathbf{r}) \phi_j(\mathbf{r}) d\mathbf{r} = \delta_{ij}$, where

a)Electronic mail: ahb@math.dartmouth.edu.
b)Electronic mail: timo.betcke@manchester.ac.uk.

 © 2007 American Institute of Physics

Downloaded 16 Nov 2010 to 129.170.28.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions

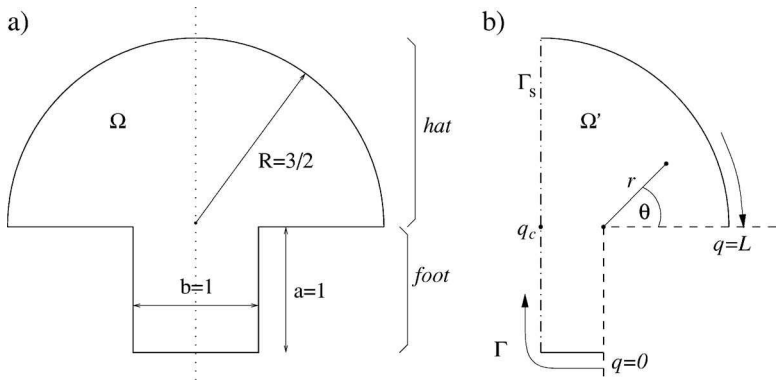


FIG. 1. (a) Mushroom billiard Ω used in this work. The dotted line shows the reflection symmetry. (b) Desymmetrized half-mushroom Ω′ used for mode calculation, and polar coordinates. Dashed lines meeting at this corner are zeros enforced by basis functions. The remaining part of $\partial\Omega'$ is Γ, comprised of two pieces: Dirichlet boundary conditions on the parts shown as solid, while boundary conditions vary (see text) on the dashed-dotted vertical line Γ_s. Boundary coordinate $q \in [0,L]$ parametrizes Γ.

$d\mathbf{r} := dxdy$ is the usual area element. "Energy" eigenvalues $E_1 < E_2 \leq E_3 \leq \cdots \to \infty$ may be written $E_j = k_j^2$, where the (eigen) wavenumber k_j is 2π divided by the wavelength.

Traditional numerical methods to compute eigenvalues and modes employ finite differences or finite elements (FEM). They handle geometric complexity well but have two major flaws: (i) it is very cumbersome to achieve high convergence rates and high accuracy, and (ii) since several nodes are needed per wavelength they scale poorly as the eigenvalue E grows, requiring of order E degrees of freedom (e.g., for the mushroom deMenezes *et al.*[2] appear limited to $j < 400$). The numerical difficulty is highlighted by the fact that *analog* computation using microwave cavities is still popular in awkward geometries.[3,14]

In contrast, we use boundary-based methods, as explained in Sec. II. These (i) achieve spectral accuracy, allowing eigenvalue computations approaching machine precision as exhibited for low-lying modes in Sec. III, and (ii) require only of order $E^{1/2}$ degrees of freedom (with prefactor smaller than boundary integral methods[15]). Furthermore at high E we use an accelerated variant, the scaling method,[16–18] which results in another factor of order $E^{1/2}$ in efficiency. These improvements allow us to find large numbers of modes up to $j \sim 10^5$; in Sec. IV we show such modes along with their Husimi (microlocal) representations on the boundary. Visualization of modes can be an important tool, e.g., in the discovery of scars.[19]

We are motivated by a growing interest in quantum ergodicity.[7,20] For purely ergodic billiards, the Quantum Ergodicity Theorem[21–24] (QET) states that in the $E \to \infty$ limit almost all modes become equidistributed (in coordinate space, and on the boundary phase space[25,26]). However no such theorem exists for mixed billiards, thus numerical studies are vital. It is a long-standing conjecture of Percival[4] that for mixed systems, modes tend to localize to one or another invariant region of phase space, with the occurrence in proportion to the phase space volumes, and that those in ergodic regions are equidistributed. (This has been tested in a smooth billiard,[27] and recently proved for certain piecewise linear quantum maps.[28]) We test the conjecture via a matrix element (10) sensitive to the boundary (for numerical efficiency); we then can categorize (almost all) modes as regular or ergodic. We address two issues which have also been raised by recent microwave experiments in the mushroom.[3] (i) The mechanism for dynamical tunneling[29] is unknown (although it has been studied in KAM mixed billiards[30]). In Sec. V we propose and test a simple model for coupling strength (related to Ref. 31) which predicts observed features of matrix element distributions. (ii) The level-spacing distribution, conjectured to be a universal feature,[5,32] is studied in Sec. VI, where we also examine spacing distributions for regular and ergodic subsets of modes. Note that we use an order of magnitude more modes than any existing experiment or study. Finally we draw conclusions in Sec. VII.

II. NUMERICAL METHODS

In this section we outline the numerical methods that make our investigation possible; the reader purely interested in results may skip to Sec. III.

A. The method of particular solutions

Our set of basis functions, or *particular solutions*, $\{\xi_n(\mathbf{r})\}_{n=1\ldots N}$ satisfy $-\Delta\xi_n = E\xi_n$ at some trial eigenvalue parameter E, but do not individually satisfy Eq. (2). The goal is now to find values of E such that there exists nontrivial linear combinations $x_1\xi_1 + x_2\xi_2 + \cdots + x_N\xi_N$, which are small on the boundary. These are then hopefully good approximations for an eigenfunction.

Let us make this precise. We define the space $\mathcal{H}(E)$ of trial functions at a given parameter E as

$$\mathcal{H}(E) = \mathrm{Span}\{\xi_1, \ldots, \xi_N\}.$$

If we denote by $\|u\|_{\partial\Omega}$ and $\|u\|_{\Omega}$ the standard L^2-norm of a trial function $u \in \mathcal{H}(E)$ on the boundary $\partial\Omega$ and in the interior Ω, we can define the normalized boundary error (also called the tension) as

$$t[u] := \frac{\|u\|_{\partial\Omega}}{\|u\|_{\Omega}}. \tag{3}$$

It is immediately clear that $t[u]=0$ for $u \in \mathcal{H}(E)$ if and only if u is an eigenfunction and E is the corresponding eigenvalue on the domain Ω. However, in practice we will rarely achieve exactly $t[u]=0$. We therefore define the smallest achievable error as $t_m(E) := \min_{u \in \mathcal{H}(E)} t[u]$. This value gives us directly a measure for the error of an eigenvalue approximation E, namely, there exists an eigenvalue E_j such that

Downloaded 16 Nov 2010 to 129.170.28.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions

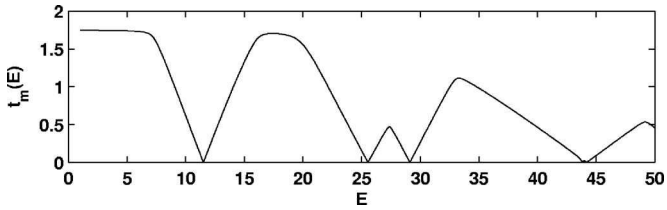


FIG. 2. The tension $t_m(E)$ plotted as a function of trial eigenvalue parameter E, for the half-mushroom with Dirichlet boundary conditions. The minima indicate the eigenvalues of this domain. Close to $E=44$ there is a cluster of two eigenvalues.

$$\frac{|E-E_j|}{E_j} \leq Ct_m(E), \tag{4}$$

where C is an $O(1)$ constant that only depends on the domain Ω. This result is a consequence of error bounds of Moler and Payne.[33,34] Hence, by searching in E for minima of $t_m(E)$ we find approximate eigenvalues with relative error given by a constant times $t_m(E)$. Figure 2 shows such a plot of $t_m(E)$ for our mushroom domain.

The implementation of this Method of Particular Solutions (MPS) depends on (i) basis set choice, and (ii) how to evaluate $t_m(E)$. The former we address in the next section. The latter requires a set of quadrature points $\{\mathbf{y}_i\}_{i=1\ldots M}$ on which to approximate the boundary integral $\|u\|_{\partial\Omega}$. One must take into account that Helmholtz basis sets tend to be ill-conditioned, that is, the $M\times N$ matrix A with entries $A_{in}:=\xi_n(\mathbf{y}_i)$ becomes numerically rank-deficient for desirable choices of N. The tension $t_m(E)$ can then be given by the square-root of the lowest generalized eigenvalue of the matrix pair (A^TA, B^TB), or by the lowest generalized singular value of the pair (A,B), where B is identical to A except with the replacement of $\{\mathbf{y}_i\}$ by interior points.[13,17,35] These different approaches are discussed in Ref. 36. Here, we use the generalized singular value implementation from Ref. 36, which is highly accurate and numerically stable. We note that these methods are related to, but improve upon, the plane wave method of Heller.[37]

B. Choice of basis functions

In order to obtain accurate eigenvalue and eigenfunction approximations from the MPS it is necessary to choose the right set of basis functions. In this section we propose a basis set that leads to exponential convergence, i.e., errors which scale as e^{-cN} for some $c>0$, as N the number of basis functions grows.

To achieve this rate we first desymmetrize the problem. The mushroom shape Ω is symmetric about a straight line going vertically through the center of the domain (see Fig. 1). All eigenmodes are either odd or even symmetric with respect to this axis. Hence, it is sufficient to consider only the right half, Ω'. The odd modes are obtained by imposing zero Dirichlet boundary conditions everywhere on the boundary $\partial\Omega'$ of the half mushroom. The even modes are obtained by imposing zero Neumann conditions on the symmetry axis Γ_s and zero Dirichlet conditions on the rest of $\partial\Omega'$.

Eigenfunctions of the Laplacian are analytic everywhere inside a domain except possibly at the boundary.[38] Eigenfunctions can be analytically extended by reflection at corners whose interior angle is an integer fraction of π.[13] The only singularity appears at the re-entrant corner with angle $3\pi/2$ [where dashed lines meet in Fig. 1(b)]. Close to this corner any eigenfunction ϕ_j can be expanded into a convergent series of Fourier-Bessel functions of the form

$$\phi_j(r,\theta) = \sum_{n=1}^{\infty} a_k J_{2n/3}(k_j r)\sin\frac{2n}{3}\theta, \tag{5}$$

where the polar coordinates (r,θ) are chosen as in Fig. 1(b). The function J_α is the Bessel function of the first kind of order α.

The expansion (5) suggests that the basis set $\xi_n := J_{2n/3}(kr)\sin 2n/3\,\theta$, where $k^2=E$, might be a good choice since these functions capture the singularity at the re-entrant corner and automatically satisfy the zero boundary conditions on the segments adjacent to this corner [dashed lines in Fig. 1(b)]. Hence, we only need to minimize the error on the remaining boundary Γ which excludes these segments. The boundary coordinate $q\in[0,L]$ parametrizes Γ; its arc length is $L=3(1+\pi/4)$. This Fourier-Bessel basis originates with Fox, Henrici, and Moler[39] for the L-shaped domain; we believe it is new in quantum physics. In Ref. 40, the convergence properties of this basis set are investigated and it is shown that for modes with at most one corner singularity the rate of convergence is exponential. Indeed, in practice we find $t_m(E_1)=O(e^{-cN})$ for some $c>0$ as the number N of basis functions grows. Hence, for the minimum $\hat{E}$ of $t_m(E)$ in an interval containing E_1, it follows from Eq. (4) that

$$\frac{|\hat{E}-E_1|}{E_1} \leq Ct_m(\hat{E}) \leq Ct_m(E_1) = O(e^{-cN}),$$

which shows the exponential convergence of the eigenvalue approximations $\hat{E}$ to E_1 for growing N.

C. Scaling method at high eigenvalue

For all odd modes apart from the lowest few we used an accelerated MPS variant, the *scaling method*,[16–18] using the same basis as above (to our knowledge the scaling method has not been combined with a re-entrant corner-adapted basis before now). Given a center wavenumber k_0 and interval half-width Δk, the scaling method finds all modes ϕ_j with $k_j\in[k_0-\Delta k, k_0+\Delta k]$. This is carried out by solving a single indefinite generalized eigenvalue problem involving a pair of matrices of the type A^TA discussed above. The "scaling" requires a choice of origin; for technical reasons we are forced to choose the singular corner. Approximations to eigenvalues lying in the interval are related to the matrix generalized eigenvalues, and the modes to the eigenvectors. The errors grow[17] as $|k_j-k_0|^3$, thus the interval width is determined by the accuracy desired; we used $\Delta k=0.1$ which ensured that $t_m(E)$ errors associated with the modes rarely exceeded 3×10^{-4}. Since the search for minima required by the MPS has been avoided, and on average $O(k)$ modes live in each

Downloaded 16 Nov 2010 to 129.170.28.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions

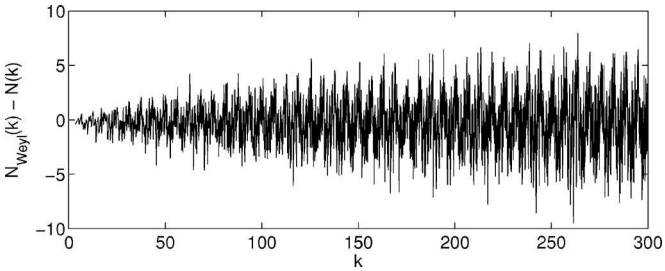


FIG. 3. Difference between the mode counting function $N(k)$ and the two-term Weyl's prediction $N_{\mathrm{Weyl}}(k)$ defined by Eq. (6), for the 16 061 odd-symmetric eigenvalues with $k_j < 300$.

interval, efficiency per mode is thus $O(k)=O(E^{1/2})$ greater than the MPS. By choosing a sequence of center wavenumbers k_0 separated by $2\Delta k$, all modes in a large interval may be computed. Rather than determine the basis size N by a convergence criterion as in Sec. II B, for $E > 10^3$ we use the Bessel function asymptotics; for large order $J_\alpha(x)$ becomes exponentially small for $x/\alpha < 1$ (the turning point is $x=\alpha$). Equating the largest argument kR (with $R=3/2$) with the largest order $2N/3$ gives our semiclassical basis size $N \approx 9k/4 = O(E^{1/2})$.

We are confident that the scaling method finds all odd modes in a desired eigenvalue window. For instance we compute all 16 061 odd symmetry modes with $k_j < 300$, using 1500 applications of the scaling method (at $k_0 = 0.1, 0.3, \ldots, 299.9$). This computation takes roughly 2 days of CPU time (all calculation times are reported for one core of a 2.4 GHz Opteron running C++ or MATLAB under linux/GNU). We verify in Fig. 3 that there is zero mean fluctuation in the difference between the (odd) level-counting function $N(k) := \#\{j : k_j \leq k\}$ and the first two terms of Weyl's law,[5]

$$N_{\mathrm{Weyl}}(k) = \frac{\mathrm{vol}(\Omega')}{4\pi}k^2 - \frac{|\partial\Omega'|}{4\pi}k, \tag{6}$$

where $|\partial\Omega'|$ is the full perimeter of the half mushroom domain. Note that there is no known variant of the scaling method that can handle Neumann or mixed BCs, hence we are restricted to odd modes. It is interesting that the method is still not completely understood from the numerical analysis standpoint.[16–18]

In applying the scaling method to the mushroom, the vast majority of computation time involves evaluating Bessel functions $J_\alpha(x)$ for large nonintegral α and large x. This is especially true for producing 2D spatial plots of modes as in Fig. 4, for which of order 10^9 evaluations are needed (1 h CPU time). We currently use independent calls to the GSL library[41] for each $J_\alpha(x)$ evaluation. This is quite slow, taking between 0.2 and 50 μs per call, with the slowest being in the region $\alpha < 50$, $10^2 < x < 10^3$. However, we note that Steed's method,[42,43] which is what GSL uses in this slow region, is especially fast at evaluating sequences $J_\alpha(x)$, $J_{\alpha-1}(x)$, $J_{\alpha-2}(x), \ldots$, and that since α is a multiple of a rational with denominator 3, only 3 such sequences would be needed to evaluate all basis functions $\{\xi_m(\mathbf{r})\}_{m=1\cdots M}$ at a given location $\mathbf{r}$. We anticipate at least an order of magnitude speed gain could be achieved this way.

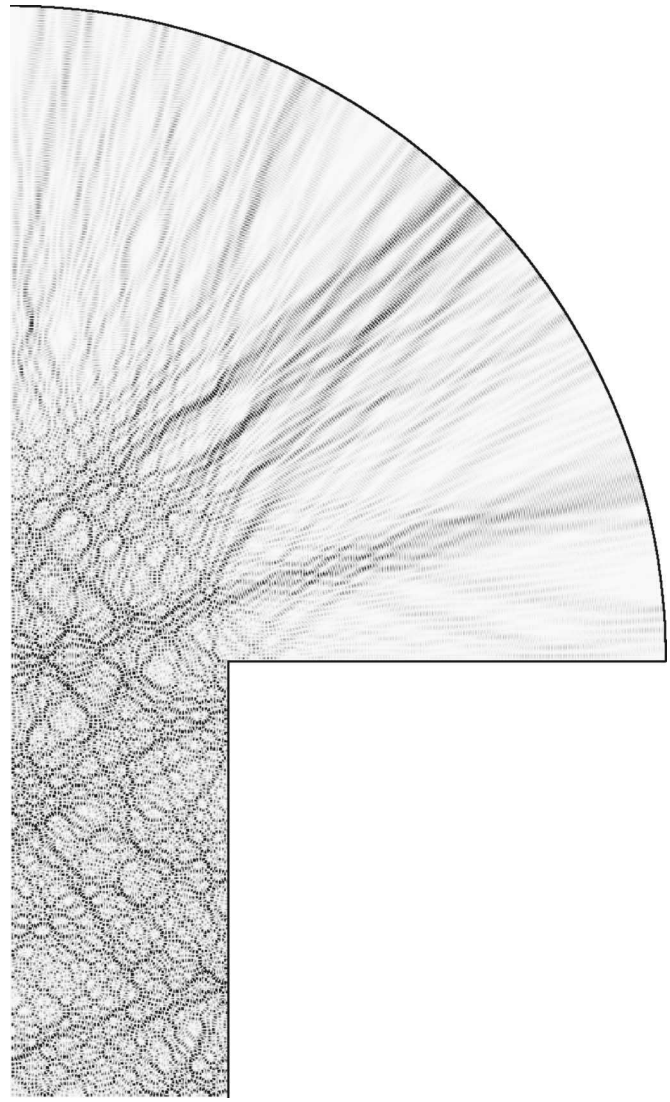


FIG. 4. High-energy eigenmode with $k_j = 499.856\cdots$, at around odd mode number $j \approx 45\,000$. This mode appears to live in the ergodic region.

III. LOW EIGENVALUE MODES

In this section we present highly accurate results for the first few even and odd modes. Odd modes are obtained by solving the eigenvalue problem with zero Dirichlet boundary conditions on the half mushroom from Fig. 1(b), using the MPS, by locating minima in the tension function of Fig. 2. In Table I(a) the eigenvalues are listed to at least 10 significant digits, and in Fig. 5(a) the corresponding modes are plotted. We emphasize that it is the exponential convergence of our basis that makes such high accuracies a simple task.

For even modes we impose Neumann BCs on Γ_s and Dirichlet BCs on the remaining part of Γ. This was achieved in the MPS by modifying the tension function (3) to read

Downloaded 16 Nov 2010 to 129.170.28.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions

TABLE I. Tables of (a) lowest 10 odd and (b) lowest 10 even eigenvalues of the mushroom. All digits shown are believed to be correct.

	j	E_j		j	E_j
	1	11.507 908 98		1	5.497 868 889
a)	2	25.550 152 54	b)	2	13.363 962 53
	3	29.124 676 10		3	18.067 786 79
	4	43.856 983 00		4	20.805 793 68
	5	44.208 992 53		5	32.589 926 04
	6	53.052 597 77		6	34.194 889 64
	7	55.200 116 30		7	41.911 982 64
	8	66.423 329 21		8	47.375 671 40
	9	69.225 768 22		9	54.624 970 98
	10	82.010 937 12		10	65.187 132 35

$$t[u] := \frac{(\|\partial_n u\|^2_{\Gamma_s} + \|u\|^2_{\Gamma\setminus\Gamma_s})^{1/2}}{\|u\|_{\Omega'}}, \tag{7}$$

where the normal derivative operator on the boundary is $\partial_n := \mathbf{n}\cdot\nabla$, the unit normal vector being $\mathbf{n}$. Table I(b), gives the smallest 10 even modes on the mushroom billiard, and the corresponding modes are plotted in Fig. 5(b).

Although we are far below the semiclassical regime we already see properties of the underlying classical dynamical system. For example, the eighth odd and the sixth even mode live along a caustic and therefore show features of the classically integrable phase space while the seventh odd and tenth even mode already shows features of the classically ergodic phase space. For comparison, in Fig. 6 we show some odd modes with intermediate eigenvalues of order 10^4 (odd mode number of order 10^3), a similar quantum number to that measured in a microwave cavity by Dietz *et al.*[3] As these authors noted, modes at this energy usually live in either the integrable or to the ergodic regions of phase space; we pursue this in detail in Sec. V.

IV. BOUNDARY AND HUSIMI FUNCTIONS

We choose a Poincaré surface of section (PSOS) (Ref. 5) defined by Birkhoff coordinates $(q,p) \in \Gamma \times [-1,1] =: Z$, where q is the boundary location as before [see Fig. 1(b)] and p is the tangential velocity component, in the clockwise sense, for a unit speed particle. (If the incident angle from the normal is θ, then $p = \sin\theta$.) The structure of this PSOS phase space is shown in Fig. 7. Our choice (which differs from that of Porter *et al.*[10]) is numerically convenient since it involves only the part of the boundary on which matching is done (Sec. II). Despite the fact that it does not cover the

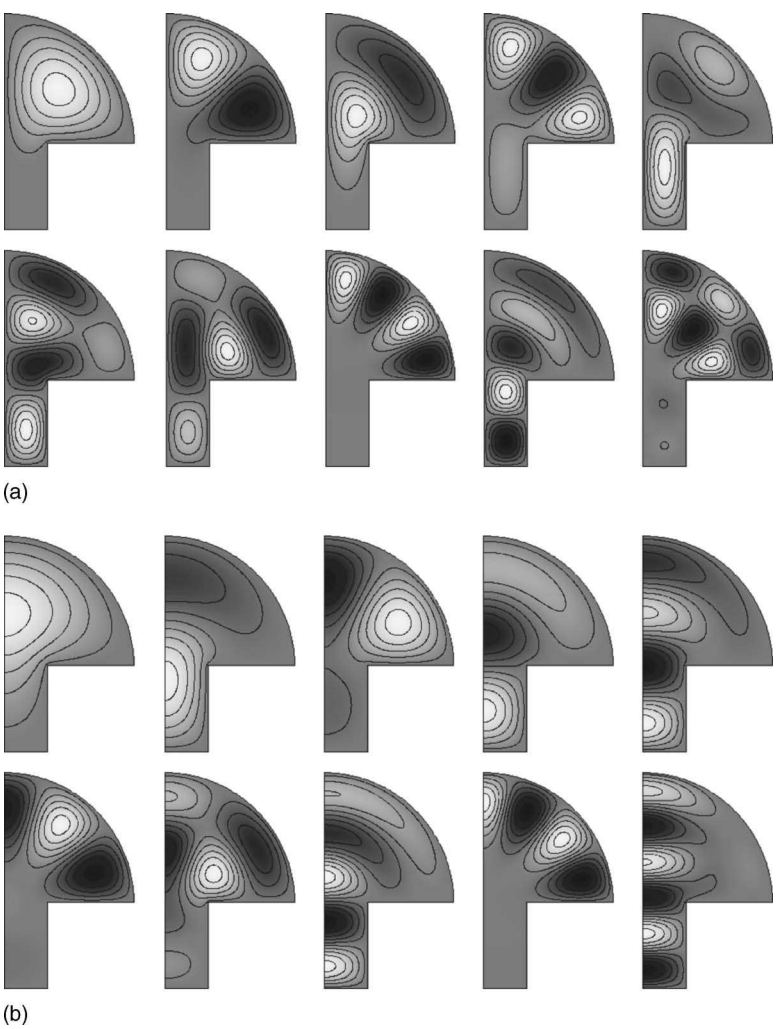


FIG. 5. The first 10 odd (a) and even (b) modes of the mushroom shape, shown as density plots. Eigenvalue increases rightwards from the top left. White corresponds to positive and black to negative values.

Downloaded 16 Nov 2010 to 129.170.28.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions

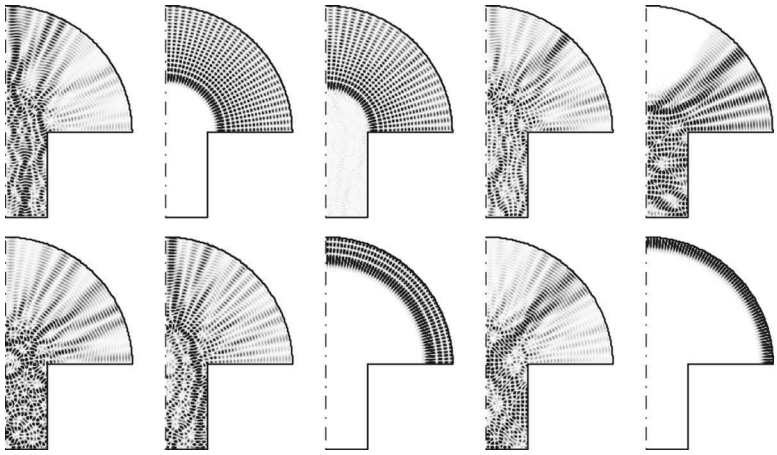


FIG. 6. The 10 odd modes of the mushroom whose eigenwavenumbers lie in the range $90<k_j<90.35$, at mode number at about $j\approx 1430$. Intensity $|\phi_j|^2$ is shown with zero white and larger values darker.

whole boundary $\partial\Omega'$, it is a valid PSOS since all trajectories must hit Γ within bounded time.

Integrable phase space consists of precisely the orbits which, for all time, remain in the hat[1] but which never come within a distance $b/2$ from the center point q_c. [This requirement is needed to exclude the zero-measure set of marginally unstable periodic orbits (MUPOs) in the ergodic region which nevertheless remain in the hat for all time.[11,12]] Simple geometry shows that the curved boundary between ergodic and integrable regions consists of points (q,p) satisfying

$$q - q_c = \frac{b/2}{\sqrt{1-p^2}}, \quad \text{for } p^2 \le p_0^2 := 1 - \frac{b^2}{4R^2}. \tag{8}$$

For our shape, $q_c=a+b/2=3/2$, $p_0^2=8/9$. In the domain $q\in[q_c+R,L]$ the boundary occurs at the lines $p=\pm b/2R=\pm 1/3$. Successive bounces that occur on Γ are described by the PSOS billiard map $f:Z\to Z$. Any such Poincaré map is symplectic and therefore area-preserving.[5]

The quantum boundary functions $\partial_n\phi_j(q)$ for $q\in[0,L]$ are convenient and natural representations of the modes. Note that they are not $L^2(\partial\Omega)$ normalized; rather they are normalized according to a geometrically weighted L^2 boundary norm via the Rellich formula (see Refs. 18 and 44)

$$\int_{\partial\Omega}(\mathbf{r}\cdot\mathbf{n})|\partial_n\phi_j|^2 dq = 2E_j, \tag{9}$$

where $\mathbf{r}(q)$ is the location of the boundary point q relative to an arbitrary fixed origin. Figure 8 shows the intensities of the first 600 odd boundary functions. Features include an absence of intensity near the corners (over a region whose size scales as the wavelength). The region $3<q<L$, in which phase space is predominantly integrable, has a more uniform intensity than $0<q<2$, which is exclusively ergodic. The region $2<q<3$ is almost exclusively integrable, but is dominated by classical turning-points corresponding to caustics; these appear as dark Airy-type spots. In $1/2<q<3/2$ there are horizontal dark streaks corresponding to horizontal "bouncing-ball" (BB) modes in the foot. Finally, a series of slanted dark streaks is visible for $3/2<q<2$; these interesting fringes move as a function of wavenumber and we postpone analysis to a future publication.

In Fig. 9 we show a sequence of 20 much higher modes with consecutive eigenvalues near wavenumber $k=500$ (eigenvalue $E=2.5\times 10^5$). These modes are a subset of the modes produced via a single generalized matrix eigenvalue problem (of size $N\approx 1200$) using the scaling method at $k_0=500$. The full set of 77 modes (evaluating boundary functions) took only 20 min CPU time. Typical tension $t_m(E)$ values were below 10^{-3}. Naively applying Eq. (4), we would conclude only about 3 relative digits of accuracy on eigenvalues. However, it is possible to rigorously improve this bound by factor $O(E_j^{1/2})$,[35] giving about 6 digits.

Figure 4 shows the 14th in the sequence in more detail. The corresponding boundary function is shown in Fig. 10(a), along with the intensity, and its Husimi distribution. The Husimi distribution is a coherent-state projection of the mode onto the PSOS phase space (see Appendix A). The

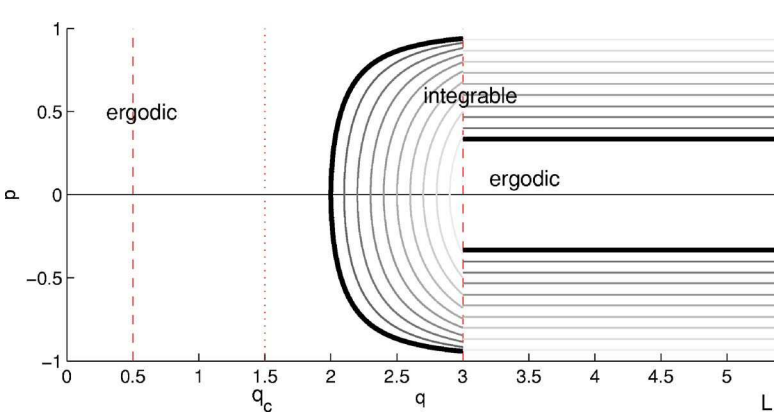


FIG. 7. (Color online) Poincaré surface of section (PSOS), that is, the classical phase space in boundary coordinates q [as shown in Fig. 1(b)] and p (sin of incidence angle). Vertical dashed lines show location of corners. The vertical dotted line shows location of q_c the focal point of the hat. The dark line shows the border of integrable phase space; note that $q=2$ corresponds to the smallest possible caustic for integrable phase space. Families of orbits defined by constant angular momentum are shown by lines in the integrable region. Note that they exchange vertical ordering at the corner, as indicated by their gray scale color labeling.

Downloaded 16 Nov 2010 to 129.170.28.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions

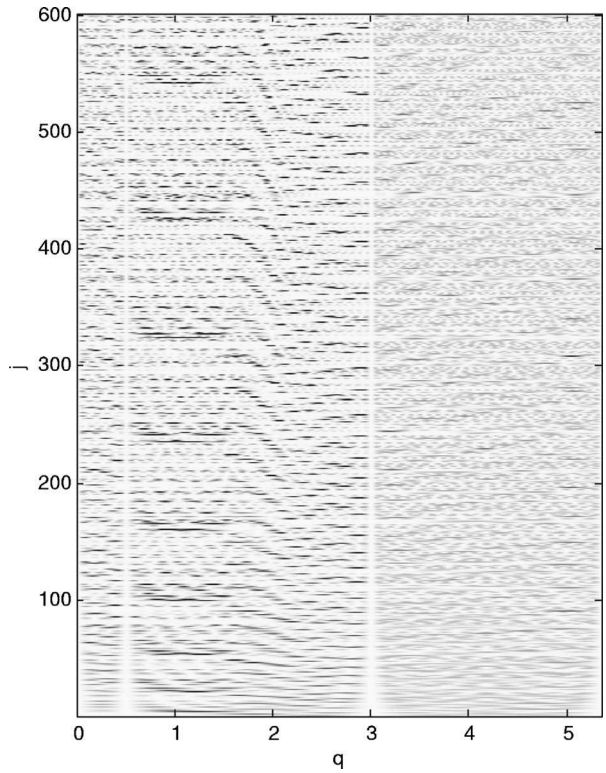


FIG. 8. Intensity of boundary normal-derivative functions $|\partial_n \phi_j(q)/k_j|^2$, plotted vs boundary coordinate q on the horizontal axis and odd mode number $j \in [1,600]$ on the vertical. The density plot shows white as zero, and larger values are darker.

choice of the aspect ratio σ is somewhat arbitrary but it is expected[19] that phase space structures have spatial scale $O(k^{-1/2})$, so we chose a scaling similar to this; with $k=500$ we used $\sigma=0.076$. By comparing the phase space (Fig. 7) we see localization to the ergodic region. The only part of the ergodic phase space that is not well covered contains BB modes in the foot (the white "box"). A scar is also visible as the 9 darkest spots: 4 pairs of spots surrounding the white box correspond to 4 bounces in the foot, and a single spot at $q \approx 5$ corresponds to a normal-incidence bounce off the circular arc. By contrast, Fig. 10(b) shows the boundary function of a mode living in the regular region (the 15th in Fig. 4); the energy-shell localization is clear. The full set of 20 Husimi functions is shown in Fig. 11. We remind the reader that in purely ergodic systems boundary functions obey the QET (Refs. 25 and 26) with almost every $\partial_n \phi_j / k_j$ tending to an invariant Husimi density of the form $C\sqrt{1-p^2}$. We might expect a similar result for the ergodic subset of modes in the ergodic phase space of the mushroom. However, Fig. 11 highlights that, despite being at a high mode number of roughly 45 000, we are still a long way from reaching any invariant density; the 7 ergodic modes have highly nonuniform distributions.

V. PERCIVAL'S CONJECTURE AND DYNAMICAL TUNNELING

In the small set of 20 high-lying modes discussed above, Percival's conjecture holds: modes are either regular or cha-

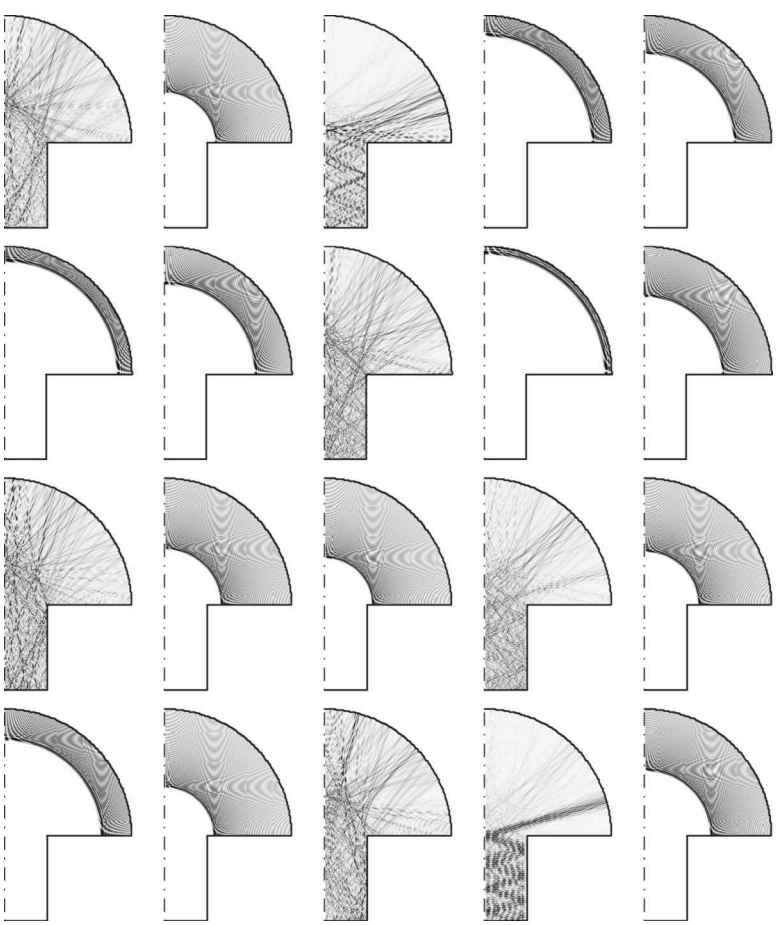


FIG. 9. Twenty high-eigenvalue consecutive modes, covering the range $k_j \in [499.800, 499.869]$, with mode number $j \approx 45\,000$. Mode number increases horizontally from the top left. $|\phi_j|^2$ is shown with zero white and larger values darker.

Downloaded 16 Nov 2010 to 129.170.28.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions

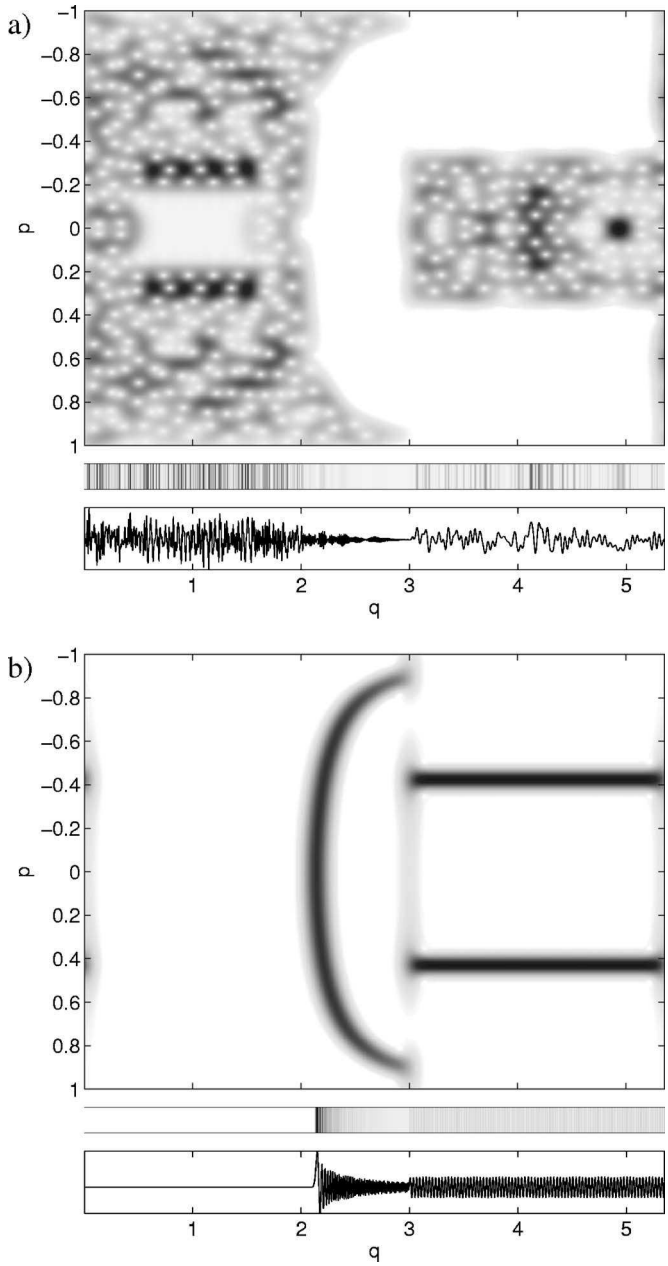


FIG. 10. (a) The mode of Fig. 4 Husimi distribution $H_{\partial_n\phi_j,\sigma}(q,p)$ defined by Eq. (A4) (top), density plot of $|\partial_n\phi_j|^2$ (middle), and graph of $\partial_n\phi_j$ (bottom). Note the q coordinate is common to the three plots. (b) Similar representation of the next highest mode at $k_j=499.858$, the 15th in the sequence of Fig. 9, which lives in the regular region.

otic but not a mixture of both. We will now study this statistically with a much larger set, the first $n=16\,061$ odd modes corresponding to $0<k_j<300$. Since the PSOS phase space in $0<q<3/2$ is ergodic for all p, the following "footsensing" quadratic form, or diagonal matrix element, is a good indicator of an ergodic component,

$$f_j := \frac{1}{2E_j}\int_0^{3/2} (\mathbf{r}\cdot\mathbf{n})|\partial_n\phi_j|^2 dq, \tag{10}$$

where, as Fig. 1(b) shows, $\mathbf{r}\cdot\mathbf{n}$ takes the value 1 for $0<q<1/2$, and 1/2 for $1/2<q<3/2$. [The weighting by $\mathbf{r}\cdot\mathbf{n}$ is chosen to mirror Eq. (9); scaling by E_j is necessary for a well-defined semiclassical limit.[25,45]]

The observed distribution of f_j is shown in Fig. 12(a). The main feature is a cluster around $O(1)$ (we associate with ergodic modes) and a wider distribution of smaller values (predominantly regular modes). We have tested that the apparent cluster lying roughly from 10^{-14} to 10^{-9} is merely an artifact reflecting the size of numerical errors in $\partial_n\phi_j$; the key point is that there is a *continuum* of values [see error bars in Fig. 12(a)] which extends from $O(1)$ down to exponentially small values. Roughly 0.75% of the total number of modes fall within each decade from 10^{-2} to 10^{-8}. We believe that in the absence of numerical errors a similar distribution would extend down many tens of orders of magnitude.

Percival's conjecture would imply that the sequence $\{f_j\}_{j=1\ldots\infty}$ has (for all but a set of vanishing measure) two limit points: zero (for regular modes), and some positive constant (for ergodic modes). Even though most mode numbers are large ($\sim 10^4$) the upper cluster still has a wide standard deviation of 0.1 (its mean is 0.39); this is in line with our recent work confirming the slow algebraic semiclassical convergence of matrix elements.[18]

We would like to test whether the relative mode frequencies of regular versus ergodic modes are in proportion to the corresponding classical phase-space volumes. We categorize modes by defining them as "regular" if $f_j<0.1$. This choice of cutoff value is necessarily a compromise between lying below the whole ergodic peak yet capturing the full dynamic range of regular modes. This gives a fraction $\alpha_{\mathrm{reg}} := n_{\mathrm{reg}}/n = 7178/16\,061 = 0.4469\cdots$ of regular modes, which is only 1.7% less than the integrable phase space fraction $\mu_{\mathrm{reg}}=0.4549\cdots$ (computed in Appendix B). Assuming that each regular mode counted arose randomly and independently due to some underlying rate (fraction of level density), we may associate a standard error of $\sqrt{n_{\mathrm{reg}}(n-n_{\mathrm{reg}})/n^3}=0.004$ with the measured fraction. Thus the discrepancy is only 2 sigma, not inconsistent with the (null) hypothesis that $\alpha_{\mathrm{reg}}=\mu_{\mathrm{reg}}$. To check whether this result persists semiclassically we computed a smaller set of $n=615$ high-lying modes sampled from the range $500<k_j<750$, up to mode number $j\approx 10^5$, and found $\alpha_{\mathrm{reg}} = 0.441\pm 0.015$, again consistent with Percival's conjecture.

A. Results and model for dynamical tunneling

The continuum of matrix element values in Fig. 12(a) is a manifestation of dynamical tunneling,[29] quantum coupling between regular and ergodic invariant phase space regions. This has recently been seen in mushroom microwave cavity modes,[3] and these authors raised the question as to the mechanism for tunneling in this shape. We address this by proposing and numerically testing a simple such model. First we notice that the density of $\ln f_j$ is roughly constant (in the range $f_h>10^{-8}$ where numerical errors are negligible). This suggests a coupling strength which is the *exponential* of some uniformly distributed quantity. We may ask whether this density is dependent on eigenvalue magnitude (energy); Fig. 12(b) shows that the density appears to die as $E^{-1/3}$,

Downloaded 16 Nov 2010 to 129.170.28.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions

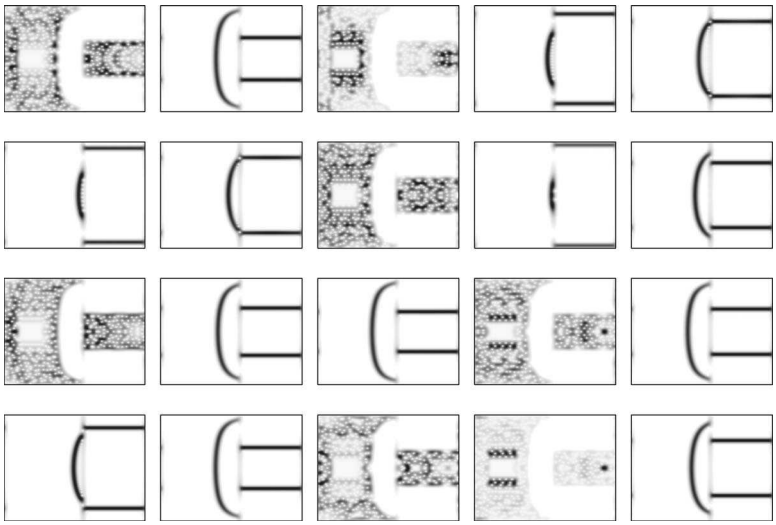


FIG. 11. Husimi distributions $H_{\partial_n \phi_j,\sigma}(q,p)$ of the 20 high-eigenvalue modes shown in Fig. 9, and in the same order. The q and p axes are as in Fig. 10.

consistent with the expectation that all f_j values for regular modes vanish in the semiclassical limit.

Our model is to assume that f_j values are controlled by a matrix element γ_j giving the *rate* of dynamical tunneling from the regular to the ergodic region. Each regular mode closely approximates an (n,m)-mode of the quarter disk, which are the product of angular function $(2/\sqrt{\pi})\sin m\theta$ and radial function

$$\psi_{mn}(r) = \frac{\sqrt{2}}{RJ'_m(k_{mn}R)} J_m(k_{mn}r), \tag{11}$$

where $n=1,2,\ldots$ is the radial mode number and $m=2,4,\ldots$ is the angular mode number, and $k_{mn}R$ is the argument of the nth zero of the J_m Bessel function. Quarter-disk eigenwavenumbers are k_{mn}. The normalization is $\int_0^R |\psi_{mn}(r)|^2 r dr = 1$. A wavepacket initially launched from such a disk mode will, in the mushroom, leak into the ergodic region due to the openness of the connection into the foot. We take the rate proportional to the probability mass of ψ_{mn} "colliding" with the foot,

$$\gamma_j := \int_0^{b/2} |\psi_{mn}(r)|^2 r dr$$
$$= \frac{4}{[k_{mn}RJ'_m(k_{mn}R)]^2} \sum_{l=0}^{\infty} (m+1+2l) \left| J_{m+1+2l}\left(\frac{b}{2}k_{mn}\right)\right|^2 \tag{12}$$

where we used [Ref. 46, Eq. (11.3.2)] to rewrite the integral. This model is similar to that proposed recently by Bäcker *et al.*[31] (in our case the "fictitious integrable system" is the quarter-disk). γ_j is exponentially small only when the Bessel function turning point lies at a radius greater than $b/2$; at eigenvalue E this occurs for $b\sqrt{E}/(2m) < 1$.

We compare in Figs. 13(a) and 13(b) f_j values for regular against γ_j values computed using all relevant (m,n) quantum numbers for the quarter-disk. It is clear that although the densities are similar, f_j is irregularly distributed, whereas γ_j values fall on a regular lattice. However, upon closer examination there is a strong correlation. We attempted to match each disk mode (m,n) seen in panel (b) with its corresponding mushroom mode j as seen in panel (a); in most of the 1051 cases there was a very clear match, with relative eigenvalue difference $|E_j - k_{mn}^2|/E_j < 10^{-4}$ in 90% of the cases, and $|E_j - k_{mn}^2|/E_j < 3\times 10^{-6}$ in 74% of the cases. (Note that, although it is not needed for our study, it would likely be possible to improve the fraction matched using data from $\partial_n \phi_j$.) As shown in panel (c), f_j values are quite correlated with the γ_j values of their matched mode. Note that an overall prefactor of $c=15$ was included to improve the fit. The resulting ratio f_j/γ_j is shown in panel (d), and has a spread of typically a factor 10^2. Since this is much less than the spread of 10^8 in the original matrix elements, this indicates that the above model is strongly predictive of dynamical tunneling strength, mode for mode. We suggest the remaining variation, and the value of c, might be explained by varying eigenvalue gaps (resonant tunneling) between quarter-disk

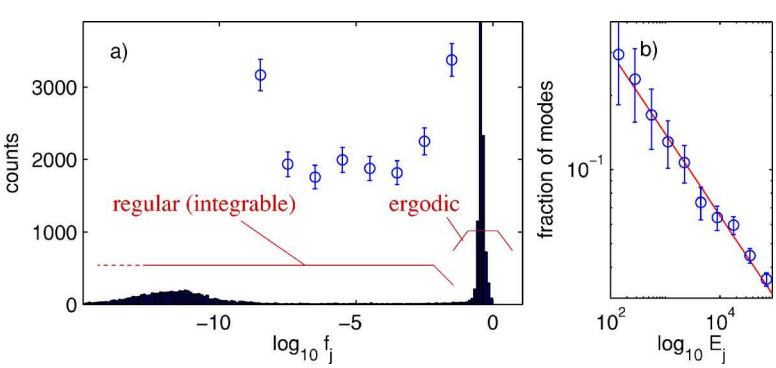


FIG. 12. (Color online) (a) Histogram of the logarithm of f_j, the "foot-sensing" matrix element (10), for the first 16 061 odd modes. Error bars show counts of f_j lying in each decade (errors assuming independent counts), on a vertical scale magnified by a factor 15. (b) Fraction of modes with $10^{-8} < f_j < 10^{-2}$ lying in logarithmically spaced E_j intervals (error bars), compared to power law $E^{-1/3}$ (solid line).

Downloaded 16 Nov 2010 to 129.170.28.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions

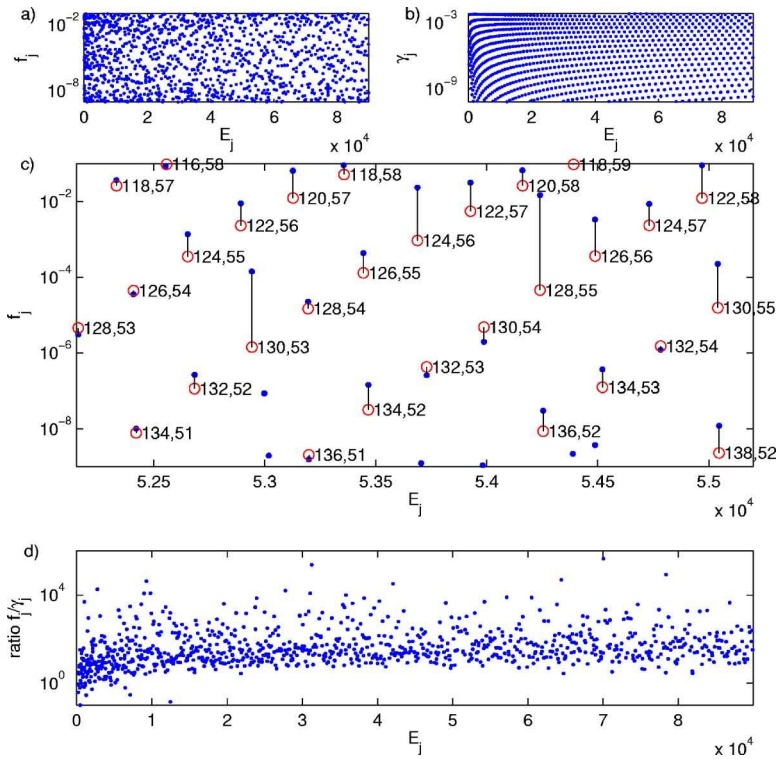


FIG. 13. (Color online) (a) Distribution of matrix elements f_j for regular modes, as a function of eigenvalue E_j. (b) Corresponding rates γ_j predicted by the model in the text. (c) Zoom of (a), showing f_j (dots) connected by vertical lines to best-matching values of $c\gamma_j$ (circles labeled by disk quantum numbers m,n), using constant $c=15$. (d) Ratio f_j/γ_j for all matched pairs.

and ergodic modes (such variation is discussed in Ref. 30), although this is an open question. Also in this simple model it is clear from the E-dependence in panel (d) that there are algebraic prefactors that should be included in a more detailed model.

Using the model we may predict the decay $\sim E^{-1/3}$ in the density of f_j values reported above, by returning to the sum in Eq. (12). For regular modes where $\gamma_j \ll 1$, the Bessel functions in the $l \geq 1$ terms have turning points successively further away from $b/2$, thus the sum may be approximated by the $l=0$ term (this has been checked numerically). We make the approximation that the turning point is close to $b/2$, that is $\varepsilon \ll 1$, where

$$\varepsilon := 1 - \frac{bk_{mn}}{2m}. \tag{13}$$

We focus on the exponentially small behavior of γ_j and drop algebraic prefactors. In Eq. (12) using Debye's asymptotics for the Bessel function [Ref. 46, Eq. (9.3.7)] and keeping leading terms for small $\varepsilon > 0$ gives

$$g := -\ln \gamma_j \approx (\text{const}) + \frac{1}{2}\ln \varepsilon + \ln k_{mn} + \frac{4\sqrt{2}}{3} m\varepsilon^{3/2}. \tag{14}$$

(This can be interpreted as the tail of the Airy approximation to the Bessel.) For fixed $\varepsilon \ll 1$ we need keep only the last term as $m \to \infty$. Fixing m while increasing n by 1 causes a small wavenumber change $k_{m,n+1} - k_{mn} \approx \pi/(p_0 R)$, causing via Eq. (13) a change $\Delta\varepsilon \approx -\pi b/(2mp_0 R)$, which in turn causes via Eq. (14) a change

$$\Delta g \approx -\pi\sqrt{2\varepsilon}/(p_0 R) \approx -\frac{\pi}{p_0 R}\left(\frac{3g}{2m}\right)^{1/3}, \tag{15}$$

where in the last step we expressed ε in terms of the asymptotic for g. Realizing that, for $\varepsilon \ll 1$ we have $m \approx bk_{mn}/2 = b\sqrt{E}/2$, and that adjacent curves of constant m in the (E,g)-plane are separated in E by $\Delta E \approx 8\sqrt{E}/b$, gives our result, the density of points in the (E,g)-plane,

$$d(E,g) = \frac{1}{|\Delta g \Delta E|} \approx \frac{p_0 R}{8\pi}\left(\frac{b^4}{3gE}\right)^{1/3}. \tag{16}$$

Recall that Figs. 13(a)–13(c) illustrate the (E,g)-plane. In Fig. 12(a) small dynamic range and counting statistics prevents this weak dependence of density on g from being detected. However the main conclusion from Eq. (16) is that the density of γ_j (and hence f_j) values lying in any fixed interval scales asymptotically as $E^{-1/3}$, in agreement with Fig. 12(b).

VI. LEVEL SPACING DISTRIBUTION AND LEVEL DENSITY FLUCTUATION

We show the nearest-neighbor spacing distribution (NNDS) of the complete set of the first $n=16\,061$ eigenvalues of odd-symmetric modes E_j with $k_j < 300$, in Fig. 14(a). Spacings were unfolded in the standard way,[32] thus a histogram of $s_j := (E_{j+1} - E_j)/\bar{E}$, where $\bar{E}$ is the mean level spacing, was collected. This is compared in the figure against the Berry-Robnik prediction[47] for a mixed system with a single regular component (of phase-space fraction $\mu_{\text{reg}} = 0.4549\cdots$) and single ergodic component. The agreement is excellent, with deviations consistent with the standard error for each

Downloaded 16 Nov 2010 to 129.170.28.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions

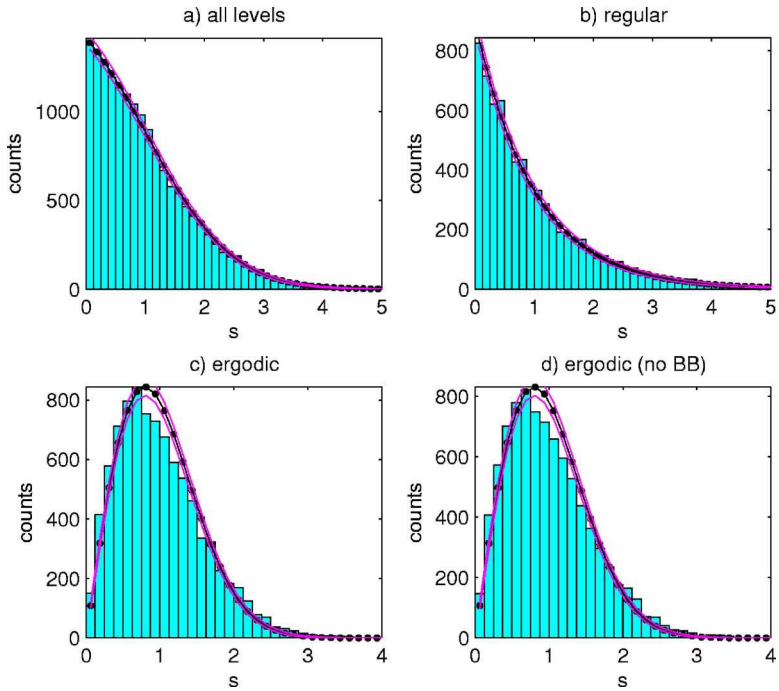


FIG. 14. (Color online) Nearest-neighbor spacing distributions (NNDS) $p(s)$ for the 16 061 modes with $k_j < 300$, estimated via a histogram with bins of width $\Delta s = 0.125$. Data (in counts per bin) are shown by bars. Predictions are shown by dots, with ±1 standard error (solid lines above and below). (a) All modes vs Berry-Robnik formula; (b) regular modes vs Poissonian formula e^{-s}; (c) ergodic modes including BB modes vs Wigner's approximate GOE formula $(\pi/2)se^{-\pi s^2/4}$, and (d) ergodic modes with BB modes removed using the categorization in Sec. VI vs the same.

bin count. In their recent work Dietz *et al.*[3] claim that there is a dip in the NNDS around $s=0.7$ associated with the super-shell structure in the hat (two periodic orbits of close lengths). Their choice of mushroom shape differs from ours only in the foot. Our results, computed using over 16 times their number of levels, show no such dip. This suggests that their observed dip is a statistical anomaly, or that it does not carry over to the rectangular-foot mushroom and therefore is not associated with the hat.

In order to study this further we computed the partial NNDS associated with regular or ergodic modes, categorized using the method of Sec. V. Regular modes [Fig. 14(b)] fit the Poisson level spacing distribution well. Ergodic modes [Fig. 14(c)] fit Wigner's standard approximate form for the GOE distribution reasonably well, however there are visible deviations; the data systematically favors small spacings $s<0.75$ while disfavoring intermediate spacings $0.75<s<1.6$. This can be quantified by comparing 0.392, the fraction of spacings with $s<0.75$, to 0.357, the corresponding fraction predicted using the Wigner distribution. Using the normal approximation to the binomial distribution, this discrepancy is nearly 7σ and is thus statistically very significant (similar conclusions are reached by the standard Kolmogorov-Smirnov test for comparing distributions). We conjecture that, as with mode intensities discussed above, the discrepancy is another manifestation of slow convergence to the semiclassical limit.

One difference between our mushroom and that of Dietz *et al.* is that our foot supports BB orbits and theirs does not. Therefore to eliminate this as a cause of difference, in Fig. 14(d) we show the ergodic NNDS with BB modes removed. Here BB modes were identified as those with $f_j>0.7$ but small integral on the base of the foot, namely, $\int_0^{1/2}(\mathbf{n}\cdot\mathbf{r})|\partial_n\phi_j|^2 dq<0.1$; the BB subset comprises only 0.8% of the total. The difference between panels (c) and (d) is barely perceptible, indicating that BB modes are not a significant contribution in our setting.

Finally, in Fig. 15 we show the amplitude spectrum $\tilde{\rho}(l):=\Sigma_{j=1}^{n}e^{ik_jl}$ of the density of states, which highlights contributions from periodic orbits of length l. Panel (a) shows all levels, while (b) and (c) shows the contribution only of levels categorized as either regular or ergodic, according to the above method. The periodic peaks at the integers in panel (a) [and absent in (b)] are due to the BB mode in the foot. As expected, (b) contains only the regular clusters of peaks associated with hat orbits which unfold to polygons in the disk. Each cluster of peaks has an upper limit point at multiples of $\pi R=3\pi/2$ corresponding to whispering-gallery rays. It is interesting that (c) contains contributions not only from UPOs but from all the peaks of (b) too.

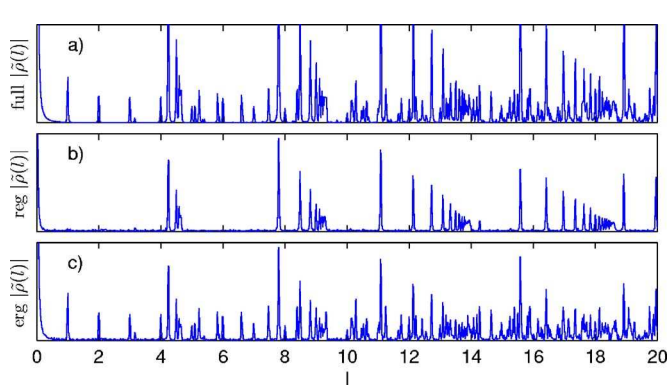


FIG. 15. (Color online) Absolute value of the Fourier transform $\tilde{\rho}(l)$ of the density of states $\rho(k):=\Sigma_{j=1}^{\infty}\delta(k-k_j)$, vs orbit length l, for (a) all eigenvalues lying below $k_j<300$, (b) regular modes only, (c) ergodic modes only.

Downloaded 16 Nov 2010 to 129.170.28.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions

VII. CONCLUSION

We have presented the first known high-lying eigenmode calculations of Bunimovich's mushroom, which has unusually simple divided phase space without KAM hierarchy. Using a basis set adapted to the re-entrant corner, the Method of Particular Solutions achieves very high accuracy for low modes, and the scaling method enables us to find high modes orders of magnitude more efficiently than any other known numerical approach, allowing the lowest $n=16\,061$ odd modes to be computed in reasonable time. Since statistical estimation errors scale as $1/\sqrt{n}$, we are therefore able to reach the 1% level for many quantities.

Chaotic modes and Husimi functions have been shown to be nonuniform and scarred even at mode number $\approx 45\,000$, evidence that the semiclassical limit is reached very slowly. Using a separation into regular versus chaotic modes, Percival's conjecture has been verified to within 2%. A new model for dynamical tunneling (similar to that of Bäcker *et al.*[31]) has been described, and shown to predict the chaotic component of predominantly regular modes to within a factor of roughly an order of magnitude (over a range of 10^8). Its prediction (via Bessel asymptotics) that the density of occurrence of modes which are regular-chaotic superpositions dies asymptotically like $E^{-1/3}$ agrees well with the first known measurement of this density.

Our study of nearest-neighbor eigenvalue spacing finds good agreement with the Berry-Robnik distribution, and for the regular subset, good agreement with the Poisson distribution. The ergodic subset shows statistically significant deviations from Wigner's GOE approximation, favoring small spacings. However we find no evidence for the dip reported at $s=0.7$ by Dietz *et al.*;[3] recall we study over 16 times their number of modes.

This study is preliminary, and raises many interesting questions: Can our model for dynamical tunneling be refined to give agreement at the impressive level found in quantum maps?[31] Does the ergodic level-spacing distribution eventually tend to the GOE expectation? Finally, can spectral manifestations of stickiness[11,12] be detected?

ACKNOWLEDGMENTS

We thank Mason Porter, Eric Heller, and Nick Trefethen for stimulating discussions and helpful advice. A.H.B. is partially funded by NSF Grant No. DMS-0507614. T.B. is supported by the Engineering and Physical Sciences Research Council Grant No. EP/D079403/1.

APPENDIX A: HUSIMI TRANSFORM

We define the Husimi transform[48] of functions on $\mathbb{R}$, for convenience reviewing the coherent state formalism in dimensionless ($\hbar$-free) units. Given a width parameter (phase space aspect-ratio) $\sigma>0$, it is easy to show that the annihilation operator

$$a := \frac{1}{\sqrt{2}}\left(\frac{q}{\sigma}+\sigma\partial_q\right) \tag{A1}$$

has a kernel spanned by the L^2-normalized Gaussian $\psi_0(q) := (\pi\sigma^2)^{-1/4}e^{-q^2/2\sigma^2}$. We work in $L^2(\mathbb{R})$, in which the Hermitian adjoint of a is $a^\dagger=(q/\sigma-\sigma\partial_q)/\sqrt{2}$. From the commutator $[a,a^\dagger]=1$ it follows, $\forall z\in\mathbb{C}$, that the coherent state

$$\psi_z := e^{-|z|^2/2}e^{za^\dagger}\psi_0 \tag{A2}$$

is an eigenfunction of a with eigenvalue z. The fact that it is L^2-normalized requires the Hermite-Gauss normalization $\|(a^\dagger)^n\psi_0\|_2^2=n!$, $\forall n\in\mathbb{N}$, which can be proved by induction. The Bargmann representation[49,50] of a function $v:\mathbb{R}\to\mathbb{C}$ is then $\langle\psi_z,v\rangle$; the Husimi representation is its squared magnitude $H_{v,\sigma}(z):=|\langle\psi_z,v\rangle|^2$. We need a more explicit form than Eq. (A2). $\psi_z=e^{za^\dagger-z^*a}\psi_0$ follows by the Baker-Campbell-Hausdorff formula $e^{A+B}=e^{-[A,B]/2}e^Ae^B$ for $[[A,B],A]=[[A,B],B]=0$. Applying this formula again and writing $z:=(q_0/\sigma+i\sigma k_0)/\sqrt{2}$ where $q_0,k_0\in\mathbb{R}$ gives

$$\psi_z(q) = e^{ik_0q_0/2}e^{ik_0q}\psi_0(q-q_0). \tag{A3}$$

This shows that the coherent state is localized in position (around q_0) and wavenumber (around k_0), thus the Husimi is a microlocal (phase space) representation,

$$H_{v,\sigma}(q_0,k_0) := \left|\int_{-\infty}^{\infty}v(q)e^{ik_0q}\psi_0(q-q_0)dq\right|^2. \tag{A4}$$

This is also known as the Gabor transform or spectrogram (windowed Fourier transform), and it can be proven equal to the Wigner transform convolved by the smoothing function ψ_0^2. Given a normal-derivative function $\partial_n\phi_j$ we periodize it in order to apply the above. We also scale the wavenumber by k_j, thus the Birkhoff momentum coordinate is $p=k_0/k_j$.

APPENDIX B: INTEGRABLE PHASE-SPACE FRACTION

The total phase space (restricting to the unit-speed momentum shell) has volume $V_{\text{tot}}=\text{vol}(\Omega'\times S^1)=2\pi\,\text{vol}\,\Omega'=2\pi(ab/2+\pi R^2/4)$. Define the function $\alpha(\mathbf{r}):=2\pi-4\sin^{-1}[b/2d(\mathbf{r})]$, where $d(\mathbf{r})$ is the distance from $\mathbf{r}$ to the center point q_c. When $\mathbf{r}$ is in the hat and $d(\mathbf{r})\in[b/2,R]$, $\alpha(\mathbf{r})$ gives the measure of the set of angles in S^1 for which orbits launched from $\mathbf{r}$ are integrable (i.e., never leave the annulus $d(\mathbf{r})\in[b/2,R]$). The regular phase space volume is found by integrating $\alpha(\mathbf{r})$ over the quarter-annulus using polar coordinates (ρ,ϕ),

$$\begin{aligned} V_{\text{reg}} &= \int_0^{\pi/2}d\phi\int_{b/2}^{R}\alpha(\rho)\rho d\rho \\ &= \frac{\pi^2}{2}\left(R^2-\frac{b^2}{4}\right)-2\pi\int_{b/2}^{R}\rho\sin^{-1}\frac{b}{2\rho}d\rho \\ &= \pi R^2\left(\cos^{-1}\frac{b}{2R}-\frac{b}{2R}p_0\right). \end{aligned}$$

The same result is given without calculus using the space of oriented lines in a full annulus, that is, $4V_{\text{reg}}=2\pi$ times the area of the segments $[(x,y):x^2+y^2<R^2,|y|>b/2]$. For our parameters we get $\mu_{\text{reg}}:=V_{\text{reg}}/V_{\text{tot}}=0.4549\cdots$.

Downloaded 16 Nov 2010 to 129.170.28.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions

[1]L. A. Bunimovich, Chaos **11**, 802 (2001).
[2]D. D. de Menezes, M. Jar E. Silva, and F. M. de Aguiar, Chaos **17**, 023116 (2007).
[3]B. Dietz, T. Friedrich, M. Miski-Oglu, A. Richter, and S. Schäfer, Phys. Rev. E **75**, 035203 (2007).
[4]I. C. Percival, J. Phys. B **6**, L229 (1973).
[5]M. C. Gutzwiller, "Chaos in classical and quantum mechanics," Vol. 1 in *Interdisciplinary Applied Mathematics* (Springer-Verlag, New York, 1990).
[6]S. De Bièvre, "Second Summer School in Analysis and Mathematical Physics (Cuernavaca, 2000)," Vol. 289 in *Contemporary Mathematics* (Amer. Math. Soc., Providence, 2001), pp. 161–218.
[7]S. Zelditch, in *Elsevier Encyclopedia of Mathematical Physics* (Academic, New York, 2006); arXiv:math-ph/0503026.
[8]Y. G. Sinaĭ, Not. Am. Math. Soc. **51**, 412 (2004).
[9]Y. G. Sinaĭ, Usp. Mat. Nauk **25**, 141 (1970).
[10]M. A. Porter and S. Lansel, Not. Am. Math. Soc. **53**, 334 (2006).
[11]E. G. Altmann, A. E. Motter, and H. Kantz, Chaos **15**, 033105 (2005).
[12]E. G. Altmann, A. E. Motter, and H. Kantz, Phys. Rev. E **73**, 026207 (2006).
[13]J. R. Kuttler and V. G. Sigillito, SIAM Rev. **26**, 163 (1984).
[14]A. Dhar, D. Madhusudhana Rao, U. Shankar N., and S. Sridhar, Phys. Rev. E **68**, 026208 (2003).
[15]A. Bäcker, "The mathematical aspects of quantum maps" Vol. 618 in *Lecture Notes in Physics* (Springer, Berlin, 2003), pp. 91–144.
[16]E. Vergini and M. Saraceno, Phys. Rev. E **52**, 2204 (1995).
[17]A. H. Barnett, Ph.D. thesis, Harvard University (2000), available at http://www.math.dartmouth.edu/~ahb/thesis_html/.
[18]A. H. Barnett, Commun. Math. Phys. **59**, 1457 (2006).
[19]E. J. Heller, Phys. Rev. Lett. **53**, 1515 (1984).
[20]F. Faure, S. Nonnenmacher, and S. De Bièvre, Commun. Math. Phys. **239**, 449 (2003).
[21]A. I. Schnirelman, Usp. Mat. Nauk **29**, 181 (1974).
[22]Y. Colin de Verdière, Commun. Math. Phys. **102**, 497 (1985).
[23]S. Zelditch, Duke Math. J. **55**, 919 (1987).
[24]S. Zelditch and M. Zworski, Commun. Math. Phys. **175**, 673 (1996).
[25]A. Hassell and S. Zelditch, Commun. Math. Phys. **248**, 119 (2004).
[26]A. Bäcker, S. Fürstberger, and R. Schubert, Phys. Rev. E **70**, 036204 (2004).
[27]G. Carlo, E. Vergini, and A. J. Fendrik, Phys. Rev. E **57**, 5397 (1998).
[28]J. Marklof and S. O'Keefe, Nonlinearity **18**, 277 (2005).
[29]M. J. Davis and E. J. Heller, J. Chem. Phys. **75**, 246 (1981).
[30]S. D. Frischat and E. Doron, Phys. Rev. E **57**, 1421 (1998).
[31]A. Bäcker, R. Ketzmerick, S. Löck, and L. Schilling, preprint (2007), arXiv:0707.0217 [nlin.CD].
[32]O. Bohigas, *Chaos et Physique Quantique (Les Houches, 1989)* (North-Holland, Amsterdam, 1991), pp. 87–199.
[33]C. B. Moler and L. E. Payne, SIAM (Soc. Ind. Appl. Math.) J. Numer. Anal. **5**, 64 (1968).
[34]J. R. Kuttler and V. G. Sigillito, SIAM J. Math. Anal. **9**, 768 (1978).
[35]A. H. Barnett, "Improved inclusion of high-lying Dirichlet eigenvalues with the method of particular solutions" (unpublished).
[36]T. Betcke, "The generalized singular value decomposition and the method of particular solutions" (2006), preprint, submitted to SIAM J. Sci. Comp.
[37]E. J. Heller, in *Chaos et Physique Quantique (Les Houches, 1989)* (North-Holland, Amsterdam, 1991), pp. 547–664.
[38]P. R. Garabedian, *Partial Differential Equations* (Wiley, New York, 1964).
[39]L. Fox, P. Henrici, and C. Moler, SIAM (Soc. Ind. Appl. Math.) J. Numer. Anal. **4**, 89 (1967).
[40]T. Betcke and L. N. Trefethen, SIAM Rev. **47**, 469 (2005).
[41]M. Galassi *et al.*, *GNU Scientific Library Reference Manual*, http://www.gnu.org/software/gsl/.
[42]A. R. Barnett, D. H. Feng, J. W. Steed, and L. J. B. Goldfarb, Comput. Phys. Commun. **8**, 377 (1974).
[43]W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C* (Cambridge University Press, Cambridge, 2002).
[44]F. Rellich, Math. Z. **46**, 635 (1940).
[45]A. Bäcker, S. Fürstberger, R. Schubert, and F. Steiner, J. Phys. A **35**, 10293 (2002).
[46]M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 10th ed. (Dover, New York, 1964).
[47]M. V. Berry and M. Robnik, J. Phys. A **17**, 2413 (1984).
[48]J.-M. Tualle and A. Voros, Chaos, Solitons Fractals **5**, 1085 (1995).
[49]V. Bargmann, Commun. Pure Appl. Math. **14**, 187 (1961).
[50]V. Bargmann, Commun. Pure Appl. Math. **20**, 1 (1967).
[51]J. Rauch, Am. Math. Monthly **85**, 359 (1978).
[52]S. A. Fulling and P. Kuchment, Inverse Probl. **21**, 1391 (2005).

Downloaded 16 Nov 2010 to 129.170.28.95. Redistribution subject to AIP license or copyright; see http://chaos.aip.org/about/rights_and_permissions

Asymptotic Rate of Quantum Ergodicity in Chaotic Euclidean Billiards

A. H. BARNETT
Courant Institute

Abstract

The quantum unique ergodicity (QUE) conjecture of Rudnick and Sarnak is that every eigenfunction ϕ_n of the Laplacian on a manifold with uniformly hyperbolic geodesic flow becomes equidistributed in the semiclassical limit (eigenvalue $E_n \to \infty$); that is, "strong scars" are absent. We study numerically the *rate* of equidistribution for a uniformly hyperbolic, Sinai-type, planar Euclidean billiard with Dirichlet boundary condition (the "drum problem") at unprecedented high E and statistical accuracy, via the matrix elements $\langle \phi_n, \hat{A}\phi_m \rangle$ of a piecewise-constant test function A. By collecting 30 000 diagonal elements (up to level $n \approx 7 \times 10^5$) we find that their variance decays with eigenvalue as a power 0.48 ± 0.01, close to the estimate $\frac{1}{2}$ of Feingold and Peres. This contrasts with the results of existing studies, which have been limited to E_n a factor 10^2 smaller. We find strong evidence for QUE in this system. We also compare off-diagonal variance as a function of distance from the diagonal, against Feingold-Peres at the highest accuracy (0.7%) thus far in any chaotic system. We outline the efficient scaling method used to calculate eigenfunctions. © 2006 Wiley Periodicals, Inc.

1 Introduction

The nature of the quantum (wave) mechanics of Hamiltonian systems whose classical counterparts are chaotic has been of longstanding interest, dating back to Einstein in 1917 (see [51] for a historical account). The field now called "quantum chaos" is the study of such quantized systems in the short-wavelength (semiclassical, $\hbar \to 0$, or high energy) limit and has become a fruitful area of enquiry for both physicists (for reviews, see [29, 31]) and mathematicians (see [45, 46, 65]) in recent decades. In contrast to those of integrable classical systems, eigenfunctions are irregular. A central issue, and the topic of this numerical study, is their behavior in the semiclassical limit.

We consider billiards, a paradigm problem in this field. A point particle is trapped inside a bounded planar domain $\Omega \subset \mathbb{R}^2$ and bounces elastically off the boundary $\Gamma = \partial\Omega$. Its phase-space coordinate is $(\mathbf{r}, \theta) \in \Omega \times \mathbb{S}^1$, with position $\mathbf{r} := (x, y)$ and (momentum) direction θ. The corresponding quantum mechanical system is the spectral problem of the Laplacian in Ω with homogeneous local

Communications on Pure and Applied Mathematics, Vol. LIX, 0001–0032 (2006)
© 2006 Wiley Periodicals, Inc.

boundary conditions that we may (and will) take to be Dirichlet,

$$-\Delta\phi_n = E_n\phi_n, \tag{1.1}$$

$$\phi_n(\Gamma) = 0. \tag{1.2}$$

Eigenfunctions ϕ_n are real-valued and normalized

$$\langle\phi_n, \phi_m\rangle := \int_\Omega \phi_n(\mathbf{r})\phi_m(\mathbf{r})d\mathbf{r} = \delta_{nm}, \tag{1.3}$$

where $d\mathbf{r} := dx\,dy$ is the usual area element, and the corresponding "energy" (or frequency) eigenvalues are ordered $E_1 < E_2 \leq E_3 \leq \cdots \leq \infty$. We will also write $E_j = k_j^2$ where k_j is the wavenumber. This Dirichlet eigenproblem [35], also known as the membrane or drum problem, has a rich 150-year history of applications to acoustics, electromagnetism, optics, vibrations, and quantum mechanics.

When the classical dynamics (flow) is ergodic, it is well-known that almost all eigenfunctions are "quantum ergodic" in the following sense:

THEOREM 1.1 (Quantum Ergodicity Theorem (QET) [22, 48, 58, 66]) *Let $\Omega \in \mathbb{R}^2$ be a two-dimensional compact domain with piecewise-smooth boundary whose classical flow is ergodic. Then for all n except a subsequence of vanishing density,*

$$\langle\phi_n, \hat{A}\phi_n\rangle - \bar{A} \to 0 \quad \text{as } n \to \infty \tag{1.4}$$

for all well-behaved functions $A : \Omega \to \mathbb{R}$.

The operator $\hat{A}$ is multiplication by A, a "test function" whose spatial average is $\bar{A} := (1/\mathrm{vol}(\Omega)) \int_\Omega A(\mathbf{r})d\mathbf{r}$. The physical interpretation is that almost all quantum expectation values (diagonal matrix elements) of the observable $\hat{A}$ tend to their classical expectation $\bar{A}$, an example of the correspondence principle of quantum mechanics. Note the choice "almost all" need not depend on A [58]. Equivalently, almost all probability densities $|\phi_j|^2$ tend to the uniform function $1/\mathrm{vol}(\Omega)$, weakly in $L^1(\Omega)$ (see Figure 1.2). The proof relies on the machinery of semiclassical analysis including Fourier integral operators (see [39, 65] and references within). In our context the "well behaved" requirement is, loosely speaking, that A not be oscillatory on the (vanishing) wavelength scale $1/k_j$. (Formally $\hat{A}$ must be a zeroth-order pseudodifferential operator, a Weyl quantization of the principal symbol $A(\mathbf{r})$). We investigate only multiplication operators, that is, no dependence on momentum, and, as we will see below, by further restricting ourselves to piecewise-constant test functions we will exploit a huge numerical efficiency gain.

Our study is motivated by the fact that QET tells us nothing about the rate of convergence of $\langle\phi_n, \hat{A}\phi_n\rangle$ or the density of the excluded subsequence, both of which are needed to understand the practical applicability of the QET in quantum or other eigenmode systems. We are interested in how the size of the deviation

$\langle \phi_n, \hat{A}\phi_n \rangle - \bar{A}$ varies with eigenvalue E_n. We define its "local variance" (mean square value) at energy E by

$$V_A(E) := \frac{1}{N_L(E)} \sum_{n: E_n \in [E, E+L(E)]} \left| \langle \phi_n, \hat{A}\phi_n \rangle - \bar{A} \right|^2, \tag{1.5}$$

where $N_L(E) := N(E + L(E)) - N(E)$ and the level counting function is

$$N(E) := \#\{n : E_n \leq E\}. \tag{1.6}$$

Here we envisage an energy window width $L(E) = O(E^{1/2})$, that is, a wavenumber window of width $O(1)$: this contains $O(E^{1/2})$ eigenvalues by Weyl's law [29, 65]. For practical reasons (Section 4.1) we will in fact use other $L(E)$ widths, which nevertheless contain many ($\approx 10^3$) eigenvalues. We will test the following asymptotic form for the variance.

Conjecture 1.2 (Power-Law Diagonal Variance). For ergodic flow, as $E \to \infty$, there is the asymptotic form, for some a and γ,

$$V_A(E) \sim a E^{-\gamma}. \tag{1.7}$$

A random-wave model for eigenfunctions (Section 2.1) predicts $\gamma = \frac{1}{2}$ and a certain prefactor a_{RW}. We will also test a more elaborate heuristic from the physics literature (Section 2.2) that predicts $\gamma = \frac{1}{2}$ and the following different prefactor.

Conjecture 1.3 (Feingold-Peres Diagonal Variance [25, 27]). For ergodic flow with no symmetries other than time reversal,

$$V_A(E) \sim \frac{g\tilde{C}_A(0)}{\mathrm{vol}(\Omega)} E^{-1/2} \tag{1.8}$$

where the symmetry factor is $g = 2$.

Here the prefactor $\tilde{C}_A(\omega)$ is the power spectral density (also known as classical variance [46], spectral measure [63, 65], or fluctuations intensity [21]),

$$\tilde{C}_A(\omega) := \int_{-\infty}^{\infty} C_A(\tau) e^{i\omega\tau}\, d\tau, \tag{1.9}$$

the time autocorrelation of A being

$$C_A(\tau) := \lim_{T\to\infty} \frac{1}{T} \int_0^T A(\mathbf{r}(t)) A(\mathbf{r}(t+\tau)) dt - \bar{A}^2, \tag{1.10}$$

where $\mathbf{r}(t)$ is any uniformly distributed (ergodic) trajectory.

Remark 1.4. In the physics literature much stronger conjectures are often discussed, such as *individual* matrix elements $\langle \phi_n, \hat{A}\phi_n \rangle$ being pseudorandomly distributed with variance given as above. We present evidence for this in Section 4.2. At the other extreme, proven theorems (and some numerical work [6]) often involve sums of the form $S_p(E; A) := N(E)^{-1} \sum_{j: E_j \leq E} |\langle \phi_j, \hat{A}\phi_j \rangle - \bar{A}|^p$. Note that

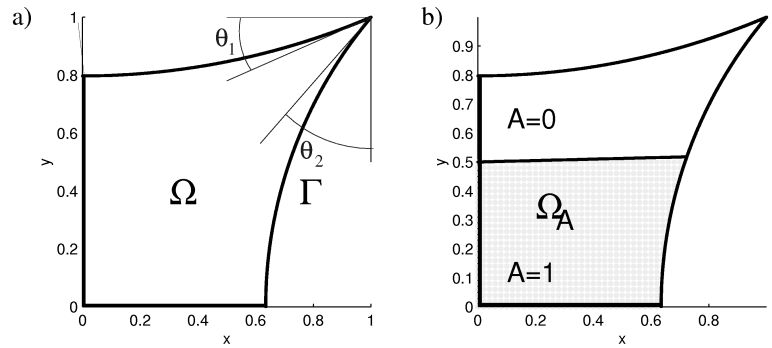


FIGURE 1.1. Billiard and test function used in this study. (a) Quarter generalized Sinai billiard formed by two circular arcs that meet at (1, 1) at angles $\theta_1 = 0.4$ to the horizontal and $\theta_2 = 0.7$ to the vertical. The straight sections lie on the axes and meet the arcs at right angles. b) Piecewise-constant $A(\mathbf{r})$ that takes the value 1 inside the region Ω_A and 0 elsewhere. $\partial\Omega_A \setminus \Gamma$ is a straight line that does not meet either wall at right angles.

asymptotic decay $S_p(E; A) \sim bE^{-p\gamma/2}$ is equivalent to Conjecture 1.2. However, we will study $V_A(E)$ rather than $S_p(E; A)$ for the following important practical reasons:

(1) Narrow spectral windows are needed rather than the complete spectrum, allowing much higher eigenvalues to be included in the statistics.
(2) Asymptotic behavior will emerge sooner since data at high E_n are not averaged with that from the lower part of the spectrum.

About 10% agreement[1] with Conjecture 1.3 has been shown in the quantum bakers map [25], and rough agreement with $\gamma = \frac{1}{2}$ has been found in hyperbolic polygons [2]. For general ergodic flow, proven bounds on the power-law rate γ in Conjecture 1.2 are quite wide: Zelditch [61] has shown that $S_p(E; A) = O((\log E)^{-p/2})$, implying $\gamma > 0$, and shown [62] that for generic A (more precisely, one with nonzero mean subprincipal symbol), $\gamma \leq 1$.

There are special "arithmetic" manifolds with ergodic flow for which powerful number-theoretic tools [44, 46] allow much more to be proven. For the quotient manifold $\mathbf{H}^2/\,\mathrm{SL}(2,\mathbb{Z})$ Luo and Sarnak [38] recently showed $S_2(E; A) \sim B(A)E^{-1/2}$, where the prefactor is a quadratic form $B(\cdot)$ diagonalized by the eigenfunctions themselves. It takes the value $B(\phi_n) = \tilde{C}(0)L(\frac{1}{2}, \phi_n)$, where $L(\,\cdot\,,\cdot\,)$ is an L-function [46]. Thus the power law appearing in Conjecture 1.3 holds but

[1] These researchers also studied the hydrogen atom in a strong magnetic field (nearly completely ergodic with sticky islands), and found some agreement at the 20% level; however, they admit that the agreement was "unexpectedly good" since it depended on a choice of a smoothing parameter.

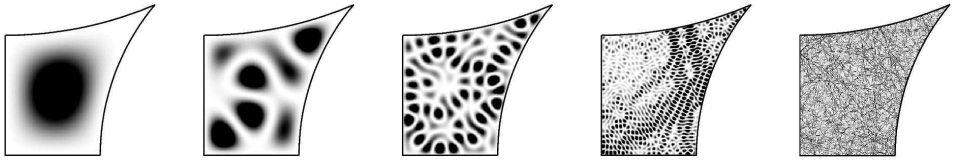


FIGURE 1.2. Illustration of spatial equidistribution of $|\phi_n|^2$ (shown as density plots; larger values are darker) with increasing eigenvalue: n^{th} Dirichlet eigenfunction for mode numbers $n = 1, 10, 100, 1000$ and $n \approx 50\,000$.

the prefactor differs from $g = 2$. This is hardly surprising; Sarnak [46] notes that a simple reflection symmetry is enough to cause $g = 0$. Arithmetic systems are very special and have many symmetries: all periodic orbits possess a single Lyapunov exponent, and eigenvalue spacing statistics are unusual for ergodic systems [17]. This makes the study of a planar Euclidean billiard system *without* symmetry (where Lyapunov exponents differ), for which no number-theoretic analytic tools exist, particularly interesting. We choose such a generic billiard (with hyperbolic, i.e., Anosov, flow [49]) for our numerical experiments (see Figure 1.1). Numerical tests of Conjecture 1.2 (in the form of $S_1(E; A)$, see Remark 1.4) exist for low eigenvalues ($n < 6000$) in the Anosov cardioid billiard [6]: various powers were found in the range $\gamma = 0.37$ to 0.5, and up to 20% deviations from the prefactor in Conjecture 1.3.

We address two other questions, the first regarding the excluded subsequence in Theorem 1.1, the second regarding off-diagonal matrix elements.

Conjecture 1.5 (Quantum Unique Ergodicity (QUE) [43]). There is no excluded subsequence in Theorem 1.1.

"Unique" refers to the existence of only one "quantum limit" (any measure to which $|\phi_n|^2$ tends weakly). Made in the context of general negatively curved manifolds, this conjecture has remarkably been proven for arithmetic manifolds [37]. In contrast, QUE has been proven *not* to hold for certain dimensional quantizations of Arnold's cat map [26]. This begs the question: for which ergodic billiards, if any, does this conjecture hold? For instance, there is strong evidence [4, 52] (but no proof [24]) that a sequence of "bouncing ball" modes (eigenfunctions concentrated on cylindrical or neutrally stable orbits that occupy zero measure in phase space) persists to arbitrarily high eigenvalues in ergodic billiards such as Bunimovich's stadium. Since such modes are not spatially uniform, the conjecture would not hold for this shape. There are also more subtle nonuniformity effects: in the physics community enhancements of $|\phi_n|^2$, dubbed "scars" by Heller [30, 31],

were discovered along isolated unstable periodic orbits, and there has been a long-standing debate as to whether these may prevent QUE. In Section 4.4 we discuss how our results relate to scarring.

Finally, we consider the size of off-diagonal matrix elements $\langle \phi_n, \hat{A}\phi_m \rangle$. Define $\Delta_k(E) := 2\pi/[E^{1/2}\,\mathrm{vol}(\Omega)]$ as the mean level spacing (in wavenumber).

Conjecture 1.6 (Feingold-Peres Off-Diagonal Variance [27, 57]). Fix $\omega \in \mathbb{R}$. Then for ergodic flow, as $E \to \infty$, there is the asymptotic result

$$\begin{aligned} V_A(E;\omega) &:= \frac{\Delta_k(E)}{2\epsilon(E)N_L(E)} \sum_{\substack{m,n:E_n\in[E,E+L(E)]\\ |k_m-k_n-\omega|\le\epsilon(E)}} |\langle \phi_n, \hat{A}\phi_m\rangle|^2 \\ &\sim \frac{\tilde{C}_A(\omega)}{\mathrm{vol}(\Omega)}\, E^{-1/2} \end{aligned} \tag{1.11}$$

where $0 < \epsilon(E) = O(E^{-1/4})$.

$V_A(E;\omega)$ measures the mean off-diagonal quantum variance a "distance" (in wavenumber units) ω from the diagonal, which is thus given by classical variance at frequency ω (see Section 2.2 for a heuristic argument). As above, the choice $L(E) = O(E^{1/2})$ is envisaged. The rate at which the wavenumber window $\epsilon(E)$ vanishes includes a growing number $O(E^{1/4})$ of modes. The result (with equivalent choices of $L(E)$ and $\epsilon(E)$) has been proven for Schrödinger operators with smooth confining potential using coherent states [23]. Conjecture 1.6 is a stronger version of the spectral measure theorem [53, 63]

$$\lim_{E\to\infty} \frac{1}{N(E)} \sum_{\substack{m,n:E_n\in[0,E]\\ \alpha<k_m-k_n<\beta}} |\langle \phi_n, \hat{A}\phi_m\rangle|^2 = \int_\alpha^\beta \frac{\tilde{C}_A(\omega)}{2\pi}\, d\omega, \tag{1.12}$$

which holds for $\alpha < \beta$ independent of ergodicity. It is known that $V_A(E;\omega)$ vanishes for ergodic weak-mixing flows (weak-mixing ensures $\tilde{C}_A(\omega)$ is a bounded function), but without any proven rate [59]. The most accurate previous numerical test of Conjecture 1.6 is the 10% agreement found for billiards with $\hat{A}$ a singular boundary operator [10]. Other tests have included the quartic oscillator [3], bakers map [19], and limaçon billiard [40].

Existing numerical studies of all the above conjectures share the features of low and poorly quantified accuracy (i.e., lack of statistical rigor in the tests), and relatively low mode numbers ($n \sim 10^3$ to 10^4). In this work we remedy both these flaws by performing a large-scale study using nonstandard cutting-edge numerical techniques that excel at very high eigenvalues. In Section 2.2 we review heuristic arguments for Conjectures 1.2, 1.3, and 1.6. Then in Section 3 (which refer to the

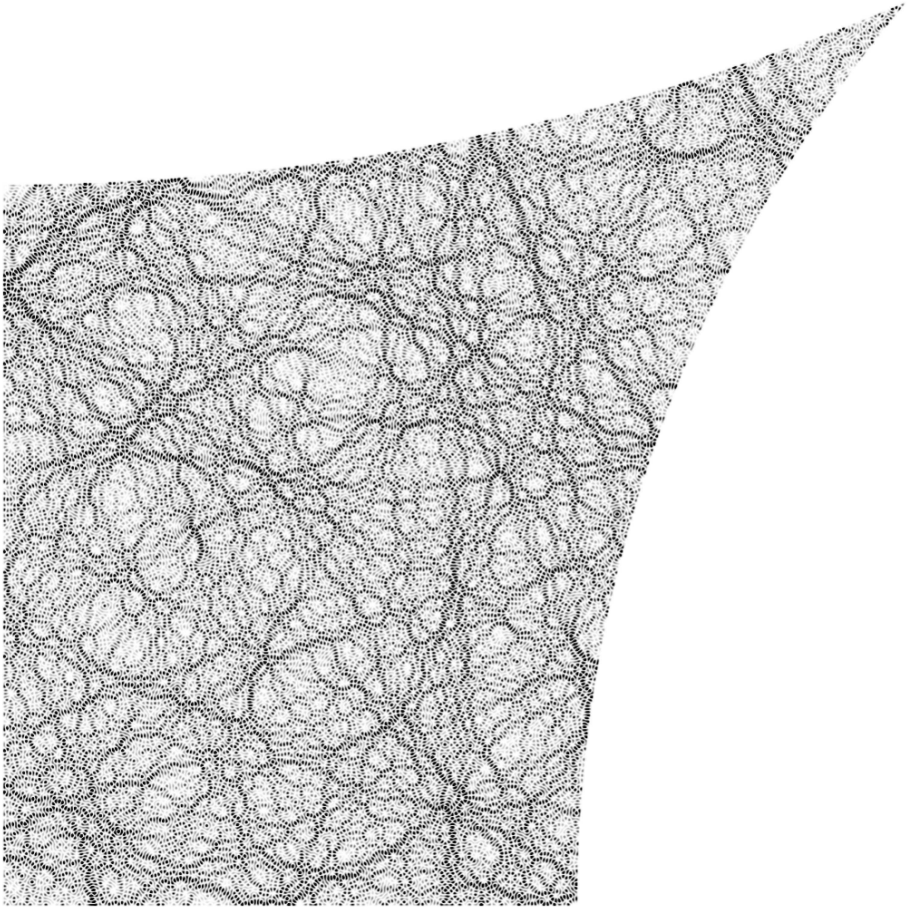


FIGURE 2.1. Density plot of $|\phi_n|^2$ for an eigenfunction with $k_n = 999.90598\ldots$, that is, $E_n \approx 10^6$ and level number $n \approx 5 \times 10^4$. There are about 225 wavelengths across the diagonal.

three appendices) we outline numerical methods; we emphasize there are several innovations. However, the reader interested in the results on the four conjectures and discussion may skip directly to Section 4. We summarize our conclusions in Section 5.

2 Heuristic Arguments for Ergodicity Rate

Here we review some arguments for Conjectures 1.2, 1.3, and 1.6 from the physics literature. We feel these are appropriate since quantum chaos is somewhat of a crossover area between mathematics and physics.

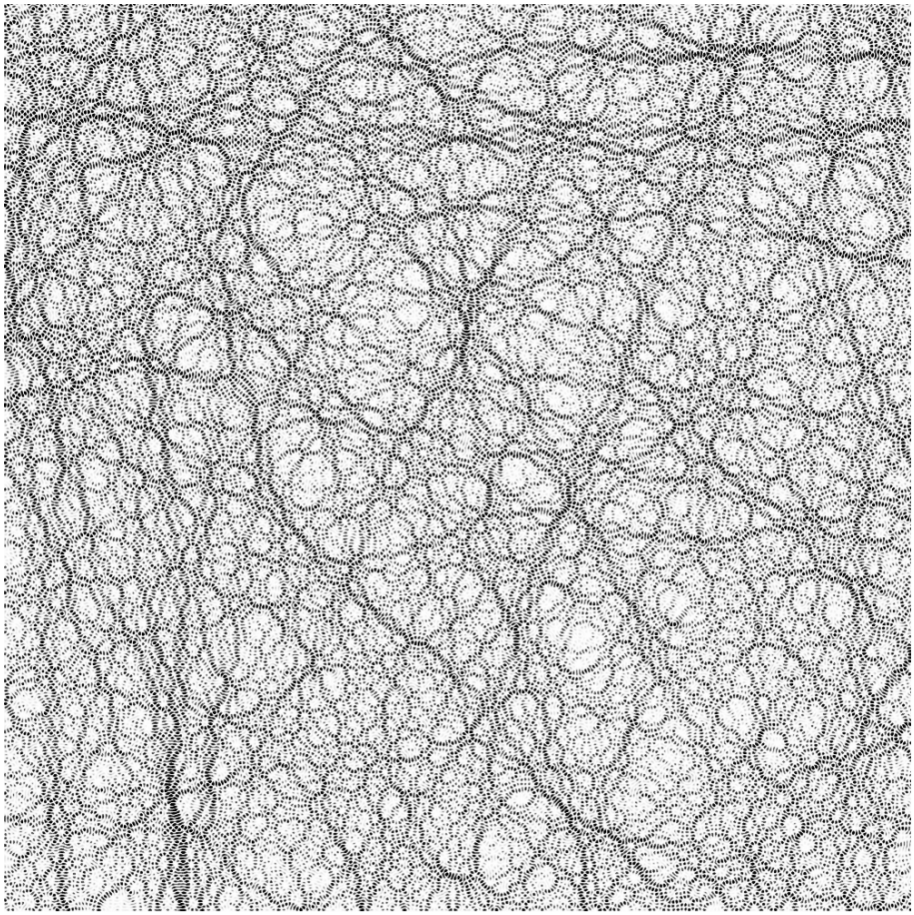


FIGURE 2.2. Density plot of one sample from the ensemble of random plane waves with the same wavenumber magnitude $|\mathbf{k}| = k$ and mean intensity as the eigenfunction in Figure 2.1, shown in a square region of space. (Note there are no boundary conditions imposed.) The "stringy" structures visible to the eye are a feature of the constant wave vector magnitude; they disappear if a finite range of k is included [31].

2.1 Random Wave Model

Berry [12] put forward the conjecture that chaotic eigenfunctions should look locally like a superposition of plane waves of fixed energy E, traveling in all directions, with random amplitudes and phases (compare Figures 2.1 and 2.2). We will show (following [25]) that this model satisfies Conjecture 1.2 with certain γ and a. We note that Zelditch has more rigorously considered models of random orthonormal bases on manifolds and shown that they possess similar ergodicity properties to quantum chaotic eigenfunctions [60, 64].

Using the notation $\langle \cdot \rangle$ to denote averaging over the ensemble of plane wave coefficients, the two-point correlation of a random-wave field is [12]

$$\langle \phi(\mathbf{r}_1)\phi(\mathbf{r}_2)\rangle = \frac{1}{\mathrm{vol}(\Omega)} J_0(k|\mathbf{r}_1 - \mathbf{r}_2|), \tag{2.1}$$

where we use the normalization $\mathrm{vol}(\Omega)\langle \phi^2 \rangle = 1$ appropriate to the billiard area. This applies to both real (time-reversal symmetric) and complex (non-time-reversal symmetric) waves; from now we stick to the real case. Because the model is a statistical one, it is meaningful to speak of the variance of a particular matrix element within the ensemble, namely $\mathrm{var}(A_{nm})$ where we use $A_{nm} := \langle \phi_n, \hat{A}\phi_m \rangle$. A simple calculation using (2.1) and Wick's theorem for Gaussian random variables gives

$$\langle \phi_n(\mathbf{r}_1)\phi_m(\mathbf{r}_1)\phi_n(\mathbf{r}_2)\phi_m(\mathbf{r}_2)\rangle = \frac{1}{\mathrm{vol}(\Omega)^2}\left[\delta_{nm} + g_{nm} J_0^2(k|\mathbf{r}_1 - \mathbf{r}_2|)\right] \tag{2.2}$$

where

$$g_{nm} := \begin{cases} g, & n = m, \\ 1, & n \neq m. \end{cases} \tag{2.3}$$

Here the symmetry factor is $g = 2$; we will shortly see it gives the ratio of diagonal to off-diagonal variance. A key assumption made was that ϕ_n and ϕ_m are statistically *independent* members of the ensemble when $n \neq m$, a natural assumption in the context of random matrix theory (RMT) ($g = 2$ for time-reversal symmetry is a standard result in the Gaussian orthogonal ensemble [18]). The independence assumption, like the random-wave model itself, remains a heuristic one, albeit one with numerical support. Clearly orthogonality (1.3) dictates that, when considered as functions over all of Ω, ϕ_n and ϕ_m cannot be independent! (As was mentioned there are improved models that take this into account [64]). An intuitive argument can be made if a restriction to a small subregion of Ω is made: the eigenfunctions behave like random orthogonal vectors, and projections of such vectors onto a much smaller-dimensional subspace become approximately independent.

Now we use (2.2) to evaluate the variance diagonal and off-diagonal matrix elements, writing the variance as the mean square minus the square mean,

$$\begin{aligned} \mathrm{var}(A_{nm}) &= \left\langle \left| \int_\Omega \phi_n(\mathbf{r})\phi_m(\mathbf{r})A(\mathbf{r})d\mathbf{r} \right|^2 \right\rangle - \langle A_{nm}\rangle^2 \\ &= \frac{g_{nm}}{\mathrm{vol}(\Omega)^2} \int_\Omega \int_\Omega J_0(k_n|\mathbf{r}_1 - \mathbf{r}_2|) J_0(k_m|\mathbf{r}_1 - \mathbf{r}_2|) d\mathbf{r}_1\, d\mathbf{r}_2 \\ &\approx \frac{g_{nm}}{\pi\, \mathrm{vol}(\Omega)} E^{-1/2} \int_\Omega \int_\Omega \frac{A(\mathbf{r}_1)A(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|}\, d\mathbf{r}_1\, d\mathbf{r}_2, \end{aligned} \tag{2.4}$$

where we used $\langle A_{nm}\rangle = \delta_{nm}\bar{A}$. In the final step two approximations have been made: (1) $L|k_n - k_m| \ll 1$ where L is the largest spatial scale of $A(\mathbf{r})$, meaning that the two Bessel functions always remain in phase and so can be set equal, and (2) the asymptotic form $J_0(x) \sim (2/\pi x)^{1/2}\cos(x - \pi/4)$ was used, and $\cos^2$ replaced by its average value $\frac{1}{2}$, giving a semiclassical expression valid when $kl \gg 1$, where l is the smallest relevant spatial scale in $A(\mathbf{r})$. Considering the diagonal and off-diagonal cases, (2.4) implies

$$\operatorname{var}(A_{nn}) = g \operatorname{var}(A_{nm}) \tag{2.5}$$

in a region $n \approx m$ close enough to the diagonal. The diagonal case of (2.4) gives the power law $\gamma = \frac{1}{2}$,

$$V_A(E) \approx a_{\mathrm{RW}} E^{-1/2}, \tag{2.6}$$

where the prefactor takes the form of a Coulomb interaction energy of the "charge density" $A(\mathbf{r})$,

$$a_{\mathrm{RW}} = \frac{g}{\pi \operatorname{vol}(\Omega)} \int_\Omega \int_\Omega \frac{A(\mathbf{r}_1)A(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|}\, d\mathbf{r}_1\, d\mathbf{r}_2. \tag{2.7}$$

Note that this model takes no account of the billiard shape or boundary conditions.

2.2 Classical Autocorrelation Argument (FP)

Feingold and Peres [27] were the first to derive a semiclassical expression for diagonal variance in chaotic systems. There are two steps:

(1) relating off-diagonal variance to $\tilde{C}_A(\omega)$, the power spectral density, yielding Conjecture 1.6, then
(2) relating diagonal variance to off-diagonal variance close to the diagonal, yielding Conjecture 1.3.

Step 1. Our presentation is loosely based on Cohen [21]; we emphasize that this is not a mathematical proof but rather one form of a heuristic common in physics literature (cf. [25, 32, 40, 57]). The autocorrelation (1.10) of the "signal" $A(\mathbf{r}(t))$ associated with a uniformly distributed ergodic unit-speed trajectory $\mathbf{r}(t)$ is $C_A(\tau) = \overline{A(0)A(\tau)}$, where the ergodic theorem is used to rewrite the time average as an average over initial phase-space locations $(\mathbf{r}_0, \theta_0)$.

Fixing τ gives the function of phase space $\mathcal{A}(\mathbf{r}_0, \theta_0) := A(0)A(\tau)$. Applying QET to this function $\mathcal{A}$ gives, as $E \to \infty$,

$$N_L(E)^{-1} \sum_{n:E_n \in [E, E+L(E)]} \langle \phi_n, \hat{A}(0)\hat{A}(\tau)\phi_n\rangle \sim C_A(\tau), \tag{2.8}$$

where $\hat{A}(t)$ is the quantization of A shifted in time (according to the Heisenberg picture of quantum mechanics). A window $L(E) = O(E^{1/2})$ is sufficient for va-

lidity of QET [65]. Assuming a wave dispersion relation $\omega = k$, the operator $\hat{A}(t)$ is expressed in the eigenfunction basis,

$$\langle \phi_n, \hat{A}(t)\phi_m \rangle = A_{nm} e^{-i(k_m - k_n)t}. \tag{2.9}$$

Using this and inserting a sum over projections onto all eigenfunctions into (2.8) gives

$$N_L(E)^{-1} \sum_{n: E_n \in [E, E+L(E)]} \sum_{m=1}^{\infty} |A_{nm}|^2 e^{-i(k_m - k_n)\tau} \sim C_A(\tau). \tag{2.10}$$

Recognizing $\lim_{\epsilon \to 0} (2\epsilon)^{-1} 1_{[-\epsilon, \epsilon]}$ as the Dirac delta function, taking the inverse Fourier transform of the definition in (1.11), and using (2.10) gives

$$\frac{1}{2\pi} \int e^{-i\omega\tau} V_A(E; \omega) d\omega \sim \frac{\Delta_k(E)}{2\pi} C_A(\tau). \tag{2.11}$$

Finally, taking a Fourier transform and substituting $\Delta_k(E)$ gives Conjecture 1.6.

Step 2. To approach the diagonal we take the limit $\omega \to 0$; here $\tilde{C}_A(\omega)$ is well-defined and bounded since the flow is weak mixing [63]. The expectation that diagonal variance should exceed off-diagonal variance by the time-reversal invariance symmetry factor (2.5) with $g = 2$ gives Conjecture 1.3. Feingold and Peres [27] justify this by considering $\phi_\pm := (\phi_n \pm \phi_m)/\sqrt{2}$ for eigenfunctions ϕ_n and ϕ_m with sufficiently small $E_n - E_m$. Matrix elements are then $\langle \phi_+, \hat{A}\phi_- \rangle = (A_{nn} - A_{mm})/2$, where $A = A^*$ was used; these are expected (by rotational invariance) to have the same variance as A_{mn}. Treating these quantities as *independent* statistical variables, taking variances, and recognizing that mean values $\bar{A}$ cancel on the right gives $\mathrm{var}(A_{mn}) = \mathrm{var}(A_{+-}) = \mathrm{var}(A_{nn})/2$, that is, $g = 2$. Although this was not discussed by Feingold and Peres, the system must be assumed to be without further symmetry, or other values of g may result.

Finally, the physical significance of the correlation functions used above should be noted: the right-hand side of (1.11) is proportional to dissipation (heating) rate in a classical system driven at frequency ω with the forcing function A, and the left-hand side to quantum dissipation rate under equivalent forcing (with $\hat{A}$) within linear response theory (for reviews, see [9, 11, 21]).

3 Computing Quantum Matrix Elements

Our nonsymmetric billiard Ω is shown in Figure 1.1(a) and defined in the caption. Its classical autocorrelation has already been studied [11]. In Appendix A we present the method used to compute $\tilde{C}_A(\omega)$ to an accuracy of a fraction of a percent (this is much less time consuming than the following quantum computations). Note that since two walls are the coordinate axes, it is a desymmetrized version of a billiard of four times the area formed by unfolding reflections in x and y. Desired

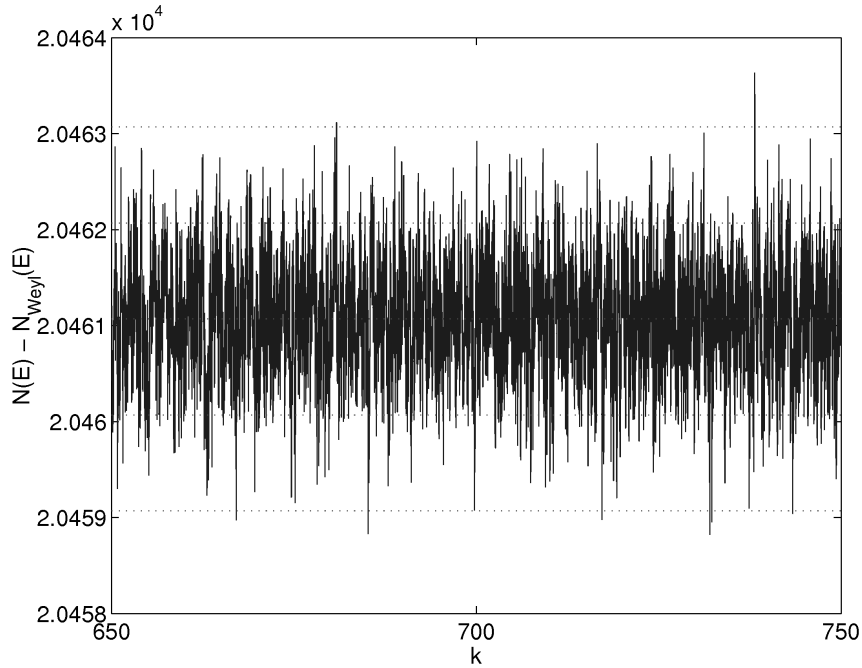


FIGURE 3.1. Demonstration that none of the 6812 modes are missing in the wavenumber window $k_n \in [650, 750]$. The level counting function (1.6) is plotted after the first two Weyl terms $N_{\text{Weyl}}(E) = (\text{vol}(\Omega)/4\pi)E - (|\Gamma|/4\pi)\sqrt{E}$ have been subtracted. The horizontal axis shows wavenumber $k = E^{1/2}$. Spectral rigidity ensures that a single missing or extra mode can be detected [29]; this would be visible as a permanent jump of size 1 (the gap between the dotted horizontal lines). No such jump occurs.

eigenfunctions are then the subset of odd-odd symmetric eigenfunctions of the unfolded billiard, a fact that enables use of a smaller symmetrized basis set (hence higher eigenvalues) and a reduction of the effective boundary to Γ_{desym}, the part of Γ that excludes the symmetry lines (axes), accelerating the quantum calculation by a factor of about 4 [9].

We chose the test function $A(\mathbf{r})$ to be the characteristic function of the subdomain Ω_A shown in Figure 1.1(b) that falls on one side of the straight line $\partial\Omega_A \setminus \Gamma$. Our choice of the shape of Ω_A was informed by the issue of boundary effects raised by Bäcker et al. [6], the main point being that within a boundary layer of order a wavelength, there are Gibbs-type phenomena associated with spectral projections, and by choosing a large angle of intersection of the line with Γ, their contribution is minimized. Our classical mean is $\bar{A} = \text{vol}(\Omega_A)/\text{vol}(\Omega) \approx 0.55000$. Matrix

elements $\langle \phi_n, \hat{A}\phi_m \rangle = \langle \phi_n, \phi_m \rangle_{\Omega_A}$ are computed using integrals of eigenfunctions in an efficent manner described shortly.

Eigenfunctions and eigenvalues were found with the "scaling method" [9, 56], outlined further in Appendix B. This is a little-known basis approximation method, a variant of the "method of particular solutions" [15, 35], which uses collocation on Γ_{desym} to extract all eigenvalues lying in an energy window $E_n \in [E, E + L(E)]$, where $L(E)$ is $O(E^{1/2})$. The equivalent wavenumber window is $O(1)$. The spectrum in larger intervals can then be found by collecting from sufficiently many windows. We are certain that all modes and no duplicates have been found in the desired intervals, as demonstrated by comparing the counting function against Weyl's law in Figure 3.1. At energy E (wavenumber k) the required basis size $N = O(N_{\text{sc}})$, the "semiclassical basis size," is

$$N_{\text{sc}} := \frac{k|\Gamma_{\text{desym}}|}{\pi}, \tag{3.1}$$

where $|\Gamma_{\text{desym}}|$ indicates the desymmetrized perimeter. Rather than evaluating ϕ_n at a set of points covering all of Ω, which would be very expensive, only *basis coefficients* of ϕ_n are computed, from which $\phi_n(\mathbf{r})$ can later be computed at any desired location. Computational effort is $O(N^2) = O(E)$ per mode, assuming $O(N^3)$ theoretical effort for dense matrix diagonalization (in fact, because of memory limitations, for large N it was not this favorable). Because of the simultaneous computation of many modes, the scaling method is faster by $O(N)$ than any other known method (see overview in [9, 35]) such as boundary integral equations [5]. We needed $N \approx 3500$ at the largest wavenumber reached ($k_n \approx 4000$ at $n \approx 7 \times 10^5$) for the billiard under study; in this case the resulting efficiency gain is roughly a factor of a thousand! Similar efficiency gains have been reported in other studies of the Dirichlet eigenproblem at extremely high energy in both two [20, 54, 56] and three [41, 42] dimensions. Only a couple of studies in billiards have computed eigenfunctions at greater n, and they invariably involved shapes without corners (for example, [20]). Note that in Appendix B we outline the basis set innovation that allows us to handle nonconvex shapes and (nonreentrant) corners effectively. Despite its success the scaling method has not yet been analyzed in a rigorous fashion [8].

Once a large set of eigenfunctions (in the form of their basis coefficients) have been found, such as the example plotted in Figure 2.1, matrix elements may be efficiently computed as we now show. The following is proved in Appendix C.

LEMMA 3.1 *Fix $E > 0$ and let $-\Delta u = Eu$ and $-\Delta v = Ev$ hold in a Lipschitz domain $\Omega_A \in \mathbb{R}^2$. Let $\mathbf{n}$ be the outwards unit normal vector at boundary location $\mathbf{r} \in \partial\Omega_A$ ($\mathbf{r}$ is measured relative to some fixed origin) and ds be surface measure.*

Then

$$\langle u, v\rangle_{\Omega_A} = \frac{1}{2E} \oint_{\partial\Omega_A} (\mathbf{r}\cdot\mathbf{n})(Euv - \nabla u\cdot\nabla v) + (\mathbf{r}\cdot\nabla u)(\mathbf{n}\cdot\nabla v) + (\mathbf{r}\cdot\nabla v)(\mathbf{n}\cdot\nabla u)ds.$$

This boundary integral identity, with the substitution $u = v = \phi_n$, allows diagonal matrix elements to be calculated using one-dimensional rather than two-dimensional numerical integration. Since typically 10 quadrature points per wavelength are needed for integration, and our system is up to hundreds of wavelengths in size, this is an enormous efficiency gain of order $O(N)$, or 10^3. Note that the boundary integrand is nonzero only on the line $\partial\Omega_A \setminus \Gamma$. Off-diagonal matrix elements are found via the following identity, which is a simple consequence of the divergence theorem.

LEMMA 3.2 *Let* $-\Delta u = E_u u$ *and* $-\Delta v = E_v v$ *hold with* $E_u \neq E_v$, *and other conditions be as in Lemma* 3.1; *then*

$$\langle u, v\rangle_{\Omega_A} = \frac{1}{E_u - E_v} \oint_{\partial\Omega_A} (u\mathbf{n}\cdot\nabla v - v\mathbf{n}\cdot\nabla u)ds.$$

Again we use $u = \phi_n$, $E_u = E_n$, $v = \phi_m$, and $E_v = E_m$, and the integrand is nonzero only on $\partial\Omega_A \setminus \Gamma$.

Thus values and first derivatives of eigenfunctions on boundaries alone are sufficient to evaluate all matrix elements. Note the eigenfunctions need never be evaluated in the interior of Ω_A. For the boundary integrals, $O(N)$ quadrature points are needed, and at each point $O(N)$ basis evaluations are needed to find ϕ_n and its gradient, giving $O(N^2)$ effort per eigenfunction, the same effort required to find modes by the scaling method. The calculations reported in this work took only a few CPU-days (1-GHz Pentium III equivalent, 1–2 GB RAM) in total. The effort is divided roughly equally between evaluating basis (Bessel) functions and their gradients at the quadrature points, and dense matrix diagonalization.

4 Results and Discussion

4.1 Diagonal Variance (Conjectures 1.2 and 1.3)

Figure 4.1 shows a sample of raw diagonal matrix element data, using only eigenvalues in certain intervals, up to $k \approx 10^3$ ($E \approx 10^6$). From this we chose a sequence of E-values and computed $V_A(E)$ at each. Computing matrix elements at high eigenvalues is very costly, so it would be inefficient and impractical numerically to grow the interval width as $L(E) = cE^{1/2}$ for some constant c. This would either involve ignoring or breaking into very short intervals (which would introduce large relative fluctuations) low eigenvalue data that is cheap to collect, or requiring vast numbers of modes at high eigenvalue, which is too expensive.

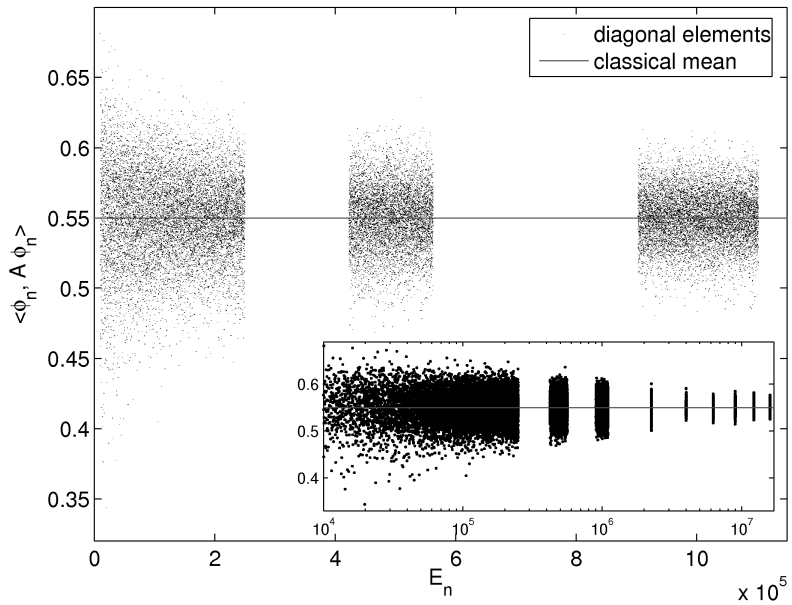


FIGURE 4.1. Scatter plot of diagonal matrix elements $\langle \phi_n, \hat{A}\phi_n \rangle$ plotted against energy eigenvalue E_n. The gaps are due to the fact that only certain windows on the E-axis have been computed; within each window all eigenvalues are found. The windows shown in the main plot correspond to wavenumbers $k_n \in [100, 500]$, $k_n \in [650, 750]$, and $k_n \in [950, 1050]$, giving a total of 28171 modes. The classical mean $\bar{A}$ is shown as a horizontal line. The inset shows the complete energy range (including the 2718 higher modes not shown in the main plot) on a log scale, with larger points to make extreme values evident.

Rather, we chose a convenient value of $L(E)$ at each E that allowed a large number of modes M to be summed at that E, while still allowing access to the highest E-values possible ($E \approx 1.6 \times 10^7$). Specifically, we split the lowest interval shown ($k_n \in [100, 500]$) into intervals containing about 10^3 modes each, kept the two other intervals shown intact (each containing an extremely large number of order 10^4), and at the higher eigenvalues (see inset of Figure 4.1) chose intervals containing 200–700 modes each.

One may ask how the $V_A(E)$ values obtained this way would differ from those obtained using a strictly growing $L(E) = cE^{1/2}$. To indicate the fluctuations in $V_A(E)$ expected from summing over a finite sample size M, we included an *error bar* of relative size $\sqrt{2/M}$. For example $M = 20\,000$ corresponds to 1% error bar (illustrating that high statistical accuracy is computationally intensive). This model does assume that deviations $|\langle \phi_n, \hat{A}\phi_n \rangle - \bar{A}|$ are roughly statistically independent,

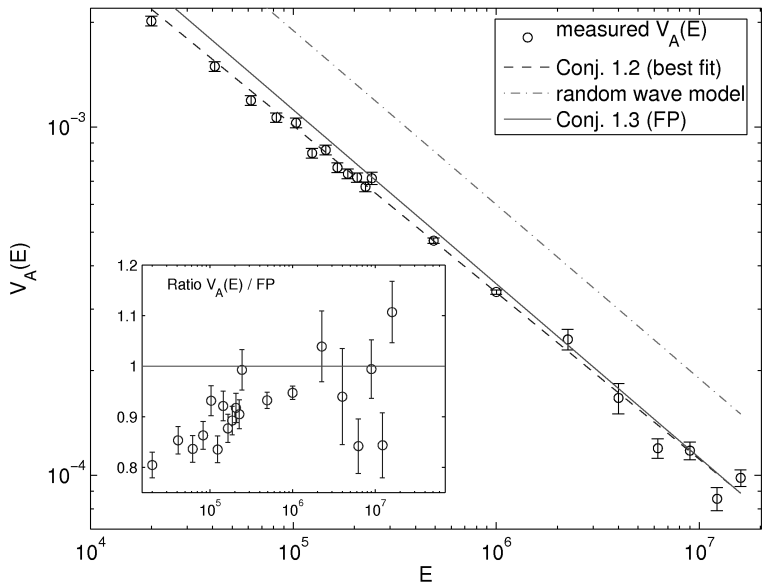


FIGURE 4.2. Log-log plot of diagonal variance $V_A(E)$ (with $\pm 1\sigma$ error bars) as a function of energy E. The smallest error bar is about 1.4% (at $E \approx 10^6$). For Conjecture 1.2 the best-fit power law ($\gamma_{\mathrm{BF}}, a_{\mathrm{BF}}$) is shown as the dashed line. The random wave prediction (2.6) and (2.7) is the dash-dotted line. Conjecture 1.3 (FP) is shown as a solid line, and the ratio of $V_A(E)$ to this prediction is shown in the inset.

an assumption motivated by RMT [18]. We find (see inset of Figure 4.2) that observed fluctuations in $V_A(E)$ fit this assumption well, so that even if correlations are present they do not affect our conclusions much. Thus if longer sums were indeed computed using $L(E) = cE^{1/2}$, it is likely that their values would lie within the error bars shown.

The resulting local variance is presented in Figure 4.2. A power-law dependence is immediately clear. We fitted the power-law model of Conjecture 1.2, obtaining a best-fit power

$$\gamma_{\mathrm{BF}} = 0.479 \pm 0.009, \tag{4.1}$$

differing by only $4 \pm 2\%$ from the random-wave model and Conjecture 1.3 value $\gamma = \frac{1}{2}$. For this fit we used weighted least squares, weighted using the numbers of modes M (equivalently, error bars) for each interval. In a standard maximum-likelihood framework [50] we marginalized over a to obtain the quoted error bar in γ. We also excluded a low-eigenvalue regime found to be nonasymptotic, taking only data with $E > E_{\mathrm{min}} = 1.6\times 10^5$. (Our criterion here was that if any lower E_{min} was used, the fitted γ_{BF} was found to be dependent on E_{min}). This best-fit power law is shown in Figure 4.2; the data are completely consistent with Conjecture 1.2.

	Best-fit $a_{\mathrm{BF}}^{(1/2)}$	Random wave a_{RW}	Feingold-Peres a_{FP}
Prefactor	0.334 ± 0.003	0.5995 ± 0.001	0.3550 ± 0.0004
Deviation from $a_{\mathrm{BF}}^{(1/2)}$	—	$79 \pm 1\%$	$6.5 \pm 0.9\%$

TABLE 4.1. Comparison of best-fit prefactor a in Conjecture 1.2, assuming fixed power $\gamma = \frac{1}{2}$, against the random wave prediction (2.7), and the Feingold-Peres prefactor $a_{\mathrm{FP}} := 2\tilde{C}_A(0)/\operatorname{vol}(\Omega)$ of Conjecture 1.3.

The random wave model and Conjecture 1.3 both involve the power $\gamma = \frac{1}{2}$. Therefore to test their validity the power was held fixed at this value while fitting only for the prefactor a, with results given in Table 4.1. Notice that the random wave model is a poor predictor; this might be expected since this model takes no account of the boundary conditions, yet the support of the test function A extends to the boundary and covers a large fraction of the volume. Intuitively speaking, images (boundary reflections) of A are not being taken into account, and evidently this is a large effect.

The prefactor (which is dominated by the large numbers of modes at $E \approx 10^6$) is overestimated by Conjecture 1.3 by only 6.5%; however, this is statistically significant (a 7σ effect). The systematic deviation is highlighted in the inset of Figure 4.2, where it appears that the deviation decreases with E. This is especially clear for $E \approx 10^6$ or below, and asymptotic agreement with Conjecture 1.3 is not inconsistent with the large scatter of the highest six data points (since they have larger error bars). If Conjecture 1.3 is correct, our results show that convergence must be alarmingly slow. This may explain why previous numerical billiard studies [2, 6] found various power laws and prefactors differing by up to 20% from Conjecture 1.3, depending on the choice of billiard and A: they had failed to reach the asymptotic regime. We achieve this with mode numbers 100 times higher than these studies. We might model slow convergence to Conjecture 1.3 by a correction, for example,

$$V_A(E) \sim \frac{g\tilde{C}_A(0)}{\operatorname{vol}(\Omega)} E^{-1/2}(1 - bE^{-\beta} + o(E^{-\beta})), \tag{4.2}$$

with β sufficiently small and b sufficiently large. Currently we do not have enough data for a meaningful fit for β and b. However, the data suggests that $0.25 \leq \beta \leq 0.5$ and that b is several times greater than 1. We note that in the physics literature there exist periodic-orbit corrections to Conjecture 1.3, derived by Wilkinson [57], but these were later claimed by Prosen not to contribute at this order [40]. Clearly more numerical and theoretical work is needed on this issue of slow convergence.

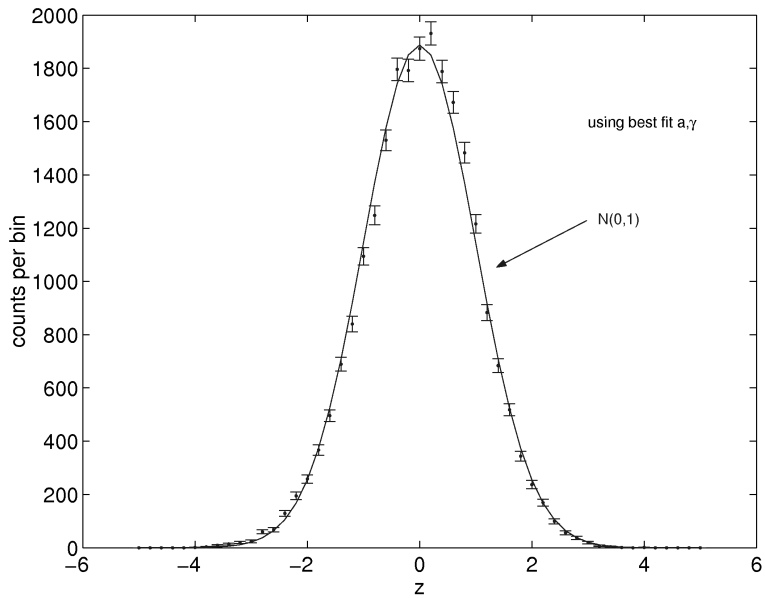


FIGURE 4.3. Histogram of rescaled diagonal deviations

$$z_n := \frac{\langle \phi_n, \hat{A}\phi_n \rangle - \bar{A}}{V_A(E_n)^{1/2}},$$

using the best-fit power-law form for Conjecture 1.2, compared to an appropriately normalized Gaussian distribution of unit variance.

4.2 Quantum Unique Ergodicity (Conjecture 1.5)

We now examine individual diagonal matrix elements. The inset of Figure 4.1 enables extreme values to be seen: it is clear that there are no anomalous extreme values which fall outside of a distribution that is condensing to the classical mean. Both the maximum 0.6811 and minimum 0.3437 of $\langle \phi_n, \hat{A}\phi_n \rangle$ occur at $E_n < 2 \times 10^4$, visible at the far left side. This is strong evidence for QUE in this system. Since about 30 000 modes are tested, the density of any excluded sequence can be given an approximate upper bound of 3×10^{-5}. We note that it is possible (although unlikely) that by an unfortunate choice of A, nonuniform or scarred modes occur that do *not* have anomalous $\langle \phi_n, \hat{A}\phi_n \rangle$ values. The only way to eliminate this possibility would be to repeat the experiment with a selection of different A-functions.

What is the distribution that the deviations $\langle \phi_n, \hat{A}\phi_n \rangle - \bar{A}$ follow? We have rescaled these deviations according to the best-fit form of the variance in Conjecture 1.2, and histogram the results in Figure 4.3. The distribution is consistent with a Gaussian, with an excellent quality of fit. (The slightly fatter tail on the low

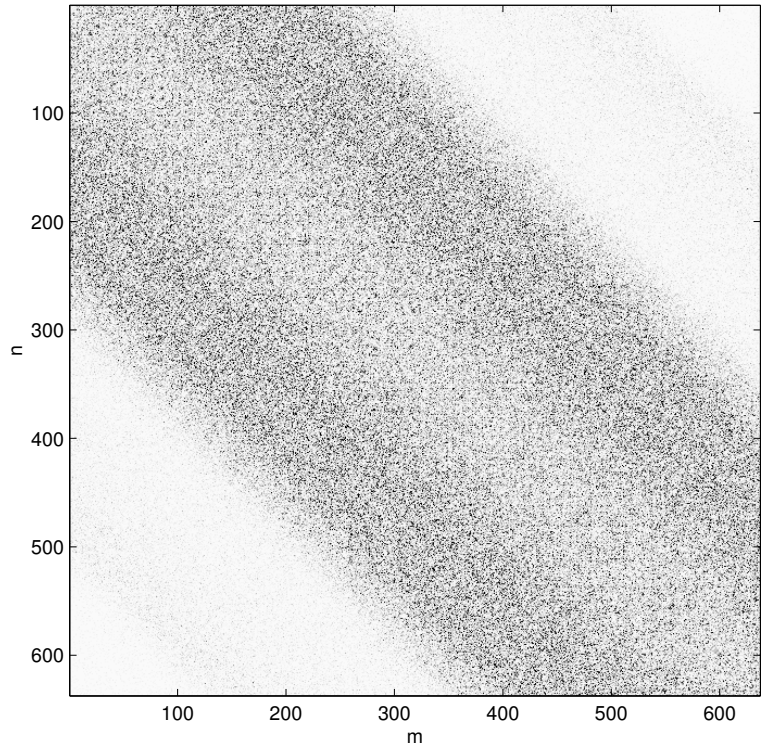


FIGURE 4.4. Density plot of the block of squared matrix elements $|\langle\phi_n, \hat{A}\phi_m\rangle - \delta_{nm}\bar{A}|^2$ for the 637 modes lying in $k \in [650, 660]$. The range white to black indicates 0 to 1.7×10^{-3}. Individual elements appear uncorrelated, the only visible structure being the intensity varying with spectral measure (band profile).

side is entirely due to low-lying modes; as can be seen in Figure 4.1, these have a skewed distribution).

4.3 Off-Diagonal Variance (Conjecture 1.6)

We decided, for reasons of numerical practicality, to test Conjecture 1.6 over a range of ω, but at a single (large) value of E. This was performed using the single sequence of 6812 eigenfunctions with $k_n \in [650, 750]$, from which about 2.3 million individual off-diagonal elements $\langle\phi_n, \hat{A}\phi_m\rangle$ were calculated, namely those in the block $n, m : k_n, k_m \in [650, 660]$, the block $n, m : k_n, k_m \in [660, 670]$, etc., with 10 blocks in total. The first such matrix block is plotted in Figure 4.4; note the strong diagonal band structure ("band profile" [21]). The mean off-diagonal element variances lying in successive ω-intervals of width 0.1 were then collected. Thus we are testing Conjecture 1.6 with a window of $L(E) \approx 1.4 \times 10^5$ and $\epsilon(E) = 0.05$. Our extremely large choice of $L(E)$ allowed statistical fluctuations

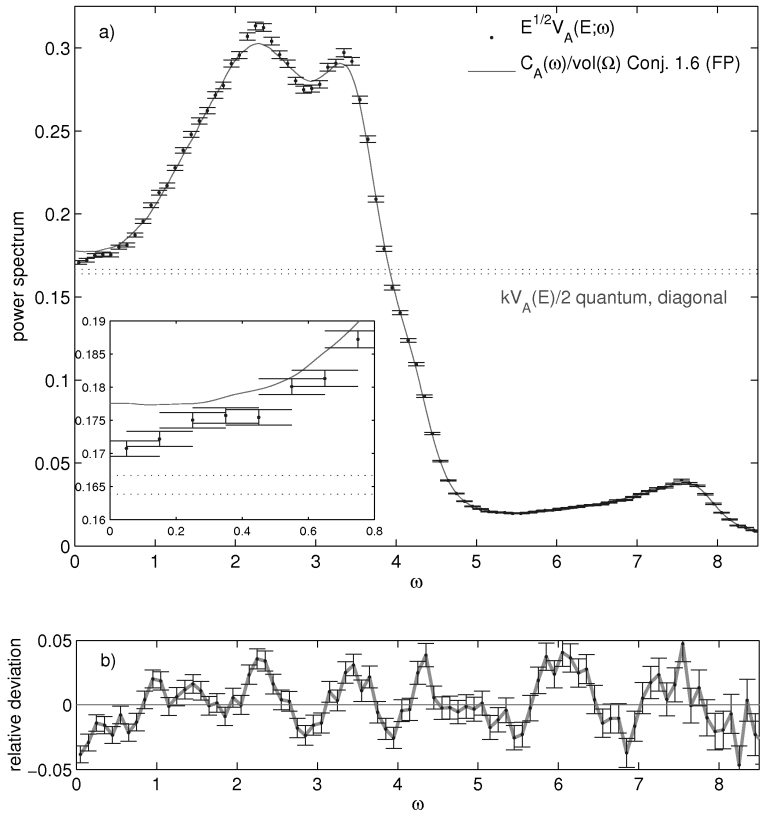


FIGURE 4.5. (a) Off-diagonal quantum variance $V_A(E;\omega)$ (error bars) at a single $E \approx 4.9 \times 10^5$, as a function of distance ω (wavenumber units) from the diagonal, compared against Conjecture 1.6 estimated using the classical power spectrum (solid line). Error bars are 0.7% for quantum (near the diagonal) and 0.2% for classical. Also, half the diagonal variance is shown as two horizontal dotted lines indicating ± 1 standard deviation. Inset is a zoom on the $\omega \to 0$ region. (b) Relative deviation (difference from 1 of ratio) between quantum variance and Conjecture 1.6.

to be minimized (error bars were estimated as in Section 4.1 and are generally less than 1%). We believe this is the most accurate test of the conjecture ever performed.

The resulting band profile $V_A(E;\omega)$ is compared against Conjecture 1.6 in Figure 4.5. The agreement is excellent, with generally less than 3% discrepancy. Let us emphasize that there are no fitted parameters. There appears to be statistically significant deviations: the peaks and valleys (points of highest curvature, including $\omega = 0$) are exaggerated more in the quantum variance than the classical. Care has been taken to ensure that this was not due to numerical errors in $\tilde{C}_A(\omega)$ (it

was smoothed only on a much finer scale; see Appendix A). As with the diagonal variance, this may reflect slow convergence.

The diagonal quantum variance, divided by $g = 2$, is also shown. If step 2 of Section 2.2 applied exactly then this would coincide with $V_A(E; 0)$, but a $3 \pm 1\%$ difference is found. It is unlikely, but possible, that this is merely a statistical fluctuation (a null result). Thus, at this energy, the diagonal variance prefactor discrepancy of 7% seems to result from the addition of two roughly equal effects: step 1 has about 4% discrepancy, and step 2 about 3%. Our data does not rule out the possibility that g is closer to 1.85 than to 2, which would imply slight positive correlations between neighboring eigenfunctions.

In terms of individual elements, we find no anomalously large values. This strongly supports "off-diagonal QUE": the vanishing of every single off-diagonal element, a stronger result than [59]. A preliminary examination suggests uncorrelated Gaussian distribution of elements, with variance given by ω-distance from the diagonal, but we postpone analysis for future work.

4.4 Discussion on Existence of Scars

Here we connect our results to the question of scarring of eigenfunctions of chaotic systems. We have space for no more than an overview of this topic (see [26, 33, 34] for reviews). In the physics community, a "scar" is taken to mean, broadly speaking, a statistical enhancement of the intensity $|\phi_n|^2$ near an isolated unstable periodic orbit (UPO). Note that the requirement that the UPO be isolated excludes neutrally stable bouncing ball modes (eigenfunction families condensing on cylindrical sections of a manifold). More precisely, the "strength" of a scar was originally defined [30, 31] as the projection of ϕ_n onto a coherent state test function (normalized Gaussian wavepacket) with center wavenumber $k = k_n$ and phase-space coordinates lying on (i.e., launched along) the UPO. An eigenfunction is said to be scarred on a UPO if this scar strength significantly exceeds its expected value given the random wave model of Section 2.1 (it was later realized that other eigenfunctions must show a corresponding *decrease* in scar strength, the so-called antiscar [33]). Heller's explanation [30] relied on the argument that, if the UPO has a small Lyapunov exponent Λ, in the semiclassical limit such a wavepacket has an infinite-time average return probability enhanced by a factor $1/\Lambda$. This probability in turn determines typical scar strengths. Note that this implies that UPOs that are less unstable show stronger scarring.

Scar theory has been reformulated [1, 13, 14, 16], and there have been many experimental observations and applications of scarring in optical, electromagnetic, and quantum systems (see [34, 47] and references within). More refined test functions have followed, addressing the issue of shape parameters of the wavepacket and using time-evolved wavepacket superpositions [33, 34]. A common feature is that the test wavepacket (or related test function) is *not* held fixed as the limit

$E \to \infty$ is taken; rather it is chosen to collapse microlocally onto a phase-space point on the UPO, with a coordinate space width $O(E^{-1/4})$. (This scaling arises since in our units the quantity playing the role of $\hbar$ is $1/k$, and both the coordinate and momentum width of the wavepacket scale like $O(\sqrt{\hbar})$ [33, 34].) Using this criterion, it is believed in the physics community that scar strength does not die out in the semiclassical limit [33]. Thus typical scar intensities (enhancements of $|\phi_n|^2$ along the UPO) do not decay; however, it has always been realized [30] that their width, hence the associated probability measure, does vanish as $O(E^{-1/4})$.

In contrast, in the mathematics community scarring is defined in terms of the limiting measure of an infinite subsequence of eigenfunctions. The measure may be in coordinate space (as we discuss here), or in phase space (via the Wigner or Husimi distribution of ϕ_j). Given uniformly hyperbolic flow, a subsequence $\{n_j\}_{j=1,\ldots,\infty}$ is said to scar on a UPO if the limit measure of the ϕ_{n_j} is not uniform but has a singular component on the UPO [26, 43]. The presence of any such scar invalidates Conjecture 1.5; likewise, validity of Conjecture 1.5 precludes any scarring. A consequence of scarring is that matrix elements of a generic test function A would tend to a number differing from the mean, that is, $|\langle\phi_{n_j}, \hat{A}\phi_{n_j}\rangle - \bar{A}| = \Omega(1)$, where Ω is the Hardy-Littlewood lower-bound symbol. Thus we might distinguish a *physicist's scar*, where a probability mass vanishing as $O(E^{-1/4})$ is associated with the UPO, from a *mathematician's scar* (following [43] we call this a "strong scar") that carries $\Omega(1)$ probability mass.

Our results (Section 4.1) provide compelling evidence that strong scars do not persist asymptotically in two-dimensional, uniformly hyperbolic billiards, in agreement with previous lower-energy studies [6] and the conjecture on negatively curved manifolds [43]. Our ergodicity rate also matches physicists' expectations: a scar probability mass of $O(E^{-1/4})$ would cause a similar matrix element deviation $|\langle\phi_n, \hat{A}\phi_n\rangle - \bar{A}| = O(E^{-1/4})$, which implies for $V_A(E)$ the power-law decay of $\gamma = \frac{1}{2}$, consistent with what is observed. The apparently strong scars (at least visually) in Heller's numerical demonstrations, at mode numbers $n \approx 2 \times 10^3$, are, in light of our work, well below the asymptotic regime. There still exist controversies about the width of scars in billiards with bouncing-ball families [36], and in related quantum models the mechanism of scarring is an active research area [47]. Mathematically, there remains a wide open territory between the no-scarring result of Lindenstrauss in modular forms [37] and the scarred-modes result of Faure, Nonnenmacher, and De Biévre in special quantizations of the Arnold's cat map [26]. For general billiard systems nothing is proven.

5 Conclusions

We have studied a generic Euclidean Dirichlet billiard whose classical dynamics is Anosov, and studied quantum ergodicity of both diagonal and off-diagonal

matrix elements. By accessing very high eigenvalues (100 times higher than in previous studies) by using the scaling method and boundary integral formulae for $\langle \phi_n, \hat{A}\phi_m \rangle$ in the case of a piecewise-constant test function A, we believe we have reached the asymptotic regime for first time. We also have unprecedented statistical accuracy due to the large number of modes computed. A summary of evidence found for the four conjectures from the introduction is as follows:

- Conjecture 1.2: Diagonal variance shows excellent agreement, with power law $\gamma = 0.479$.
- Conjecture 1.3: Diagonal variance is consistent with the Feingold-Peres prediction, but with quite slow convergence (for instance, a 7% overestimate remains at $E = 10^6$). This contrasts with a random wave model, which overestimates the fitted prefactor by 80%.
- Conjecture 1.5: There is compelling evidence for QUE in this system (density of exceptional subsequence $< 3 \times 10^{-5}$).
- Conjecture 1.6: We have excellent agreement for off-diagonal variance (of order 3% discrepancies at $E = 5 \times 10^5$) and evidence for off-diagonal QUE.

Performing this large-scale study required limiting ourselves to one billiard Ω and one operator A. In order to complete the picture, the discrepancy between the predictions of Conjectures 1.3 and 1.6 and numerically measured variances should be studied as a function of Ω and A.

Appendix A: Classical Power Spectrum

We use standard techniques [28] to estimate $\tilde{C}_A(\omega)$. For a particular trajectory, launched with certain initial location in phase space, $A(t)$ is a noisy function (stochastic stationary process). We define its windowed Fourier transform as

$$\tilde{A}(\omega) := \int_0^T A(t) e^{i\omega t}\, dt, \tag{A.1}$$

where the window is a "top hat" function from 0 to T. Using (1.10) and (1.9), and taking care with the order of limits, we have the Wiener-Khinchin theorem,

$$\tilde{C}_A(\omega) = \lim_{T\to\infty} \frac{1}{2\pi T}\, \tilde{A}^*(\omega)\tilde{A}(\omega). \tag{A.2}$$

For this single trajectory, $\tilde{A}(\omega)$ is a rapidly fluctuating random function of ω, with zero mean (for $\omega \neq 0$), variance given by $2\pi T \tilde{C}_A(\omega)$, and correlation length in ω of order $2\pi/T$. (As $T \to \infty$, the ω-correlation becomes a delta-function). Thus (A.2) converges only in the weak sense, that is, when smoothed in ω by a finite-width test function.

A given trajectory is found by solving the particle collisions with the straight and circular sections of Γ, and $A(t)$ is sampled at intervals $\Delta t = 0.02$ along the trajectory (recall we assume the particle has unit speed). Then $\tilde{A}(\omega)$ is estimated using the discrete Fourier transform (implemented by an FFT library) of this sequence of samples, giving samples of the spectrum at ω-values separated by $\Delta\omega = 2\pi/T$. The correlation in ω is such that each sample is (nearly) independent. Δt was chosen sufficiently short so that aliasing (reflection of high-frequency components into apparently low frequencies) was insignificant. A trajectory length $T = 10^4$ (about 1.8×10^4 collisions) was used. The finiteness of T causes relative errors of order t_{corr}/T, where $t_{\text{corr}} \approx 2$ (for our domain) is the time scale for exponential (since the billiard is Anosov) decay of correlations. Thus more sophisticated window functions are not needed.

Given $\tilde{A}(\omega)$ we use (A.2), with the T given above, to estimate $\tilde{C}_A(\omega)$. We smooth in ω by a Gaussian of width $\omega_{\text{sm}} = 0.03$. This width is chosen to be as large as possible to average the largest number of independent samples from the neighborhood of each ω, but small enough to cause negligible convolution of the sharpest features of $\tilde{C}_A(\omega)$.

Finally, in order to reduce further the random fluctuations in the estimate, $n_r = 6000$ independent trajectory realizations with random initial phase space locations were averaged. An estimate for the resultant relative error ϵ in $\tilde{C}_A(\omega)$ can be made by counting the number N of independent random samples that get averaged and using the fact that the variance of the square of a Gaussian zero-mean random variable (i.e., χ^2 distribution with 1 degree of freedom) is twice the mean. This gives

$$\epsilon = \left(\frac{2}{N}\right)^{1/2} \approx \left(\frac{2\pi}{n_r \omega_{\text{sm}} T}\right)^{1/2}, \tag{A.3}$$

which numerically has been found to be a conservative estimate. In our case $\epsilon \approx 2 \times 10^{-3}$, that is, about 0.2% error.

The zero-frequency limit $\tilde{C}_A(0)$ is found using the smoothed $\tilde{C}_A(\omega)$ graph at $\omega = 0$, and therefore is an average of frequencies within $O(\omega_{\text{sm}})$ of 0. This is justified because correlation decay (weak mixing) causes all moments of $C_A(\tau)$ to be finite; hence there is no singularity in $\tilde{C}_A(\omega)$ at $\omega = 0$ (it can be expanded in an even Taylor series about $\omega = 0$ with finite coefficients).

Appendix B: Scaling Method for the Dirichlet Eigenproblem

The scaling method for the solution of the Dirichlet eigenproblem in star-shaped domains was invented by Vergini and Saraceno [56], and considering its great efficiency it has received remarkably little attention. Here we give only an outline.

The method relies on the remarkable fact that the normal derivatives of eigenfunctions lying close in energy are "quasi-orthogonal" (nearly orthogonal) on the boundary with respect to the boundary weight function $r_n := \mathbf{r} \cdot \mathbf{n}$.

LEMMA B.1 *Let $\Omega \in \mathbb{R}^d$, $d \geq 2$, be a Lipschitz domain with boundary Γ and Dirichlet spectral data $\{E_j\}$ and $\{\phi_j\}$. Let $\mathbf{n}$ and $\mathbf{r}$ be defined as in Lemma* 3.1*, with $r_n := \mathbf{r} \cdot \mathbf{n}$. Then, for all $i \geq 1$, $j \geq 1$,*

$$Q_{ij} := \oint_\Gamma r_n(\mathbf{n} \cdot \nabla\phi_i)(\mathbf{n} \cdot \nabla\phi_j)\, ds = 2E_i\delta_{ij} + \frac{(E_i - E_j)^2}{4}\langle \phi_i, r^2\phi_j\rangle_\Omega. \tag{B.1}$$

This is proved in Appendix C. A corollary is that, since $r^2 := \mathbf{r} \cdot \mathbf{r}$ is a bounded operator on the domain, off-diagonal elements of Q must vanish quadratically as one approaches the diagonal. Thus the matrix with elements $Q_{ij}/2E_i$ approximates the identity matrix, when restricted to an energy window $E_i, E_j \in [E-\epsilon_0, E+\epsilon_0]$, if the window size remains relatively narrow $\epsilon_0 = o(E^{1/2})$.

We choose a "center" wavenumber $k = E^{1/2}$, near which we are interested in extracting eigenfunctions, and relative to which the wavenumber shift of mode i is $\omega_i(k) := k - k_i$. Consider an eigenfunction ϕ_i for which $\omega_i < 0$ and $|\omega_i| \ll O(1)$. We create a version spatially rescaled (dilated about the origin) by an amount needed to bring its wavenumber to k, that is, $\chi_i(k, \mathbf{r}) := \phi_i(k\mathbf{r}/k_i)$. We call this function "k-rescaled." Thus we have $-\Delta\chi_i = E\chi_i$ everywhere inside Ω, with $\chi_i(k, \mathbf{r}) = 0$ on the rescaled boundary (i.e., for all $k\mathbf{r}/k_i \in \Gamma$). The rescaled eigenfunction can be Taylor-expanded in ω_i,

$$\begin{aligned}\chi_i(k, \mathbf{r}) &= \phi_i\left(\mathbf{r} + \frac{\omega_i}{k_i}\mathbf{r}\right) = \phi_i(\mathbf{r}) + \frac{\omega_i}{k_i}\mathbf{r} \cdot \nabla\phi_i + O(\omega^2) \\ &= \frac{\omega_i}{k_i} r_n \mathbf{n} \cdot \nabla\phi_i + O(\omega^2) \quad \text{for } \mathbf{r} \in \Gamma,\end{aligned} \tag{B.2}$$

where Dirichlet boundary conditions were applied. We construct a basis of N functions $\xi_l(k, \mathbf{r})$, satisfying $-\Delta\xi_l = E\xi_l$ inside Ω, no particular boundary conditions on Γ, and nonorthogonal over Ω. We assume they approximately span the linear space in which rescaled eigenfunctions live, so that

$$\chi_i(k, \mathbf{r}) = \sum_{l=1}^{N} X_{li}\xi_l(k, \mathbf{r}) + \epsilon_i(\mathbf{r}) \quad \text{for all } i \text{ of interest}, \tag{B.3}$$

where the error ϵ_i can be made negligibly small for some N. In practice N need exceed N_{sc} defined in (3.1) by only a small factor (2 or less).

Our goal is then to solve for a shift ω_i and the corresponding i^{th} column of the coefficient matrix X. We can do this by simultaneous diagonalization of quadratic

forms. We define two symmetric bilinear forms on the boundary,

$$f(u,v) := \oint_\Gamma \frac{1}{r_n} uv\, d\mathbf{s}, \tag{B.4}$$

$$g(u,v) := \frac{1}{k}\oint_\Gamma \frac{1}{r_n}(u\mathbf{r}\cdot\nabla v + v\mathbf{r}\cdot\nabla u) d\mathbf{s}. \tag{B.5}$$

Note that defining these forms brings the extra requirement that the domain be strictly star-shaped about the origin ($r_n > 0$), which from now on we assume. In the rescaled eigenbasis, f is, via (B.2) and (B.1),

$$f(\chi_i,\chi_j) = \frac{\omega_i\omega_j}{k_ik_j} Q_{ij} + O(\omega^3) = 2\omega_i^2\delta_{ij} + O(\omega^3), \tag{B.6}$$

a matrix that is close to diagonal, because of the closeness of Q to the identity. In the same basis, recognizing that for k-rescaled functions g is equivalent to df/dk, the derivative of (B.4) with respect to the center wavenumber, and using $d\omega_i/dk = 1$, we have

$$g(\chi_i,\chi_j) = \frac{\omega_i+\omega_j}{k_ik_j} Q_{ij} + O(\omega^2) = 4\omega_i\delta_{ij} + O(\omega^2), \tag{B.7}$$

so g is also close to diagonal. Thus the set $\{\chi_i\}$ with small $|\omega_i|$ approximately diagonalizes both bilinear forms, with the approximation error growing as a power of $|\omega_i|$. As we explain below, in practice the converse applies, that is, by simultaneously diagonalizing f and g we can extract the set of eigenfunctions $\{\chi_i\}$ with smallest $|\omega_i|$. Therefore, loosely speaking, when the boundary weight function $1/r_n$ is used, domain eigenfunctions are given by the simultaneous eigenfunctions of the (squared) boundary norm and its k-derivative.

We perform the diagonalization in the basis (B.3). That is, matrices $F_{lm} := f(\xi_l,\xi_m)$ and $G_{lm} := g(\xi_l,\xi_m)$, with $l,m = 1,\dots,N$, are filled. This requires basis and first-derivative evaluations on the boundary. It is an elementary fact that given a positive matrix F and a symmetric matrix G there exists a square matrix Y and a diagonal matrix $D := \mathrm{diag}\{\mu_i\}$ that satisfy $Y^\mathsf{T} F Y = I$ and $Y^\mathsf{T} G Y = D$. The matrices Y and D can be found by standard numerical diagonalization algorithms in $O(N^3)$ time.

If (B.6) and (B.7) held without error terms, and the number of modes i for which they held were equal to (or exceeded) the basis size N, then we would be able to equate the columns of Y directly with the desired columns of X (barring permutations). In this case $\omega_i = 2/\mu_i$ would also hold, from which the desired wavenumbers k_i follow. However, by using Weyl's law and (3.1) it follows that such a large number of modes requires that the largest $|\omega_i|$ is of order unity, in which case errors in (B.6) and (B.7) would become unacceptable. It is an empirical observation found through numerical study that in fact columns of Y corresponding

to the *largest magnitude* generalized eigenvalues μ_i (and therefore the smallest shifts $|\omega_i|$) do accurately match columns of X. Thus perturbations by other vectors in the span of basis functions are small. Further discussion is postponed to a future publication [8].

We mention a couple of other implementation details. Because the generalized eigenproblem turns out to be singular, it is truncated to its nonsingular part [9, 56]. If columns of Y are normalized such that $Y^{\mathsf{T}}FY = I$ holds, then the resulting eigenfunctions can be normalized over Ω by dividing the ith column of Y by $\sqrt{2}\omega_i$. Depending on the choice of basis, spurious solutions can result; they are easily identified because their norm over Ω, computed by the following Rellich identity, is not close to 1.

LEMMA B.2 (Rellich) *With the definitions of Lemma* B.1*, for all* $j \geq 1$*,*

$$\frac{1}{2E_j} \oint_\Gamma r_n (\mathbf{n} \cdot \nabla \phi_j)^2 \, ds = 1. \tag{B.8}$$

This identity is a special case of Lemma 3.1 found by substituting $\Omega_A = \Omega$, $u = v = \phi_j$, and $E = E_j$, and recognizing $\mathbf{r} \cdot \nabla \phi_j = r_n \mathbf{n} \cdot \nabla \phi_j$. The maximum $|\omega_i|$ in which levels of useful accuracy are found is of order $0.2/R$ where R is the largest radius of the domain. The lack of missing modes obtained with this method is illustrated by Figure 3.1. There are several implementation issues and improvements that we do not have space to discuss here [8, 9, 55, 56].

A word about the basis set choice $\{\xi(k, \mathbf{r})\}$, $1 \leq i \leq N$, is needed. Until now plane waves (including evanescent plane waves [55, 56]) or regular Bessel functions [20] have been used. These fail for nonconvex domain shapes or those with corners; thus to tackle the domain in this study the author developed a basis of irregular Bessel (i.e., Neumann) functions, placed at equal intervals along a curve Γ^+ exterior to Ω. The curve Γ^+ is defined as the set of points whose nearest distance to Γ is D, with $kD = 7$ (roughly one wavelength distant). This was found to handle (nonreentrant) corners successfully. It performs extremely well for all shapes that have been attempted so far. The basis size N is about $1.5N_{\text{sc}}$ (see (3.1)); thus, depending on required accuracy, about $N/20$ useful modes are found per dense matrix diagonalization ($O(N^3)$ effort). This is $O(N)$ faster than other boundary methods; we remind the reader that N is larger than 10^3 in our work.

Appendix C: Identities Involving Eigenfunctions of the Laplacian

Let there be constants $E_u > 0$ and $E_v > 0$, and let $-\Delta u = E_u u$ and $-\Delta v = E_v v$ hold in a Lipschitz domain $\Omega_A \in \mathbb{R}^d$ for some general dimension $d \geq 2$. The following expressions for the divergence of certain vector fields result from

elementary calculus (by $\nabla\nabla u$ we mean the second-derivative tensor [dyad]):

$$\nabla\cdot(v\nabla u) = -E_u uv + \nabla u\cdot\nabla v, \tag{C.1}$$

$$\nabla\cdot(\mathbf{r}uv) = duv + u\mathbf{r}\cdot\nabla v + v\mathbf{r}\cdot\nabla u, \tag{C.2}$$

$$\nabla\cdot(\mathbf{r}\nabla u\cdot\nabla v) = d\nabla u\cdot\nabla v + \nabla u\cdot\nabla\nabla v\cdot\mathbf{r} + \nabla v\cdot\nabla\nabla u\cdot\mathbf{r}, \tag{C.3}$$

$$\nabla\cdot[(\mathbf{r}\cdot\nabla u)\nabla v] = \nabla u\cdot\nabla v - E_v v\mathbf{r}\cdot\nabla u + \nabla v\cdot\nabla\nabla u\cdot\mathbf{r}, \tag{C.4}$$

$$\nabla\cdot(r^2 v\nabla u) = 2v\mathbf{r}\cdot\nabla u - E_u r^2 uv + r^2\nabla u\cdot\nabla v. \tag{C.5}$$

PROOF OF LEMMA 3.1: Consider the following four equations: (C.2), (C.3), (C.4), and its counterpart obtained by swapping u and v. Integrating each of these over Ω_A and then applying the divergence theorem gives four expressions for surface integrals in terms of domain integrals. Substituting these into the right-hand side of the expression in Lemma 3.1 and setting $E_u = E_v = E$ gives, after cancellation, the left-hand side. (This is also discussed in appendix H of [9]). □

PROOF OF LEMMA B.1: This relies on a similar but more complicated technique. First we consider modes i and j that are nondegenerate, that is, $E_i \neq E_j$. Eight equations are needed: the four mentioned above, then (C.1) and (C.5) and their counterparts swapping u and v. Each should be integrated over Ω and then the divergence theorem applied. The following identity may then be verified by substitution of the eight resulting equations into its right-hand side. Using the abbreviations $\varepsilon := E_u - E_v$, $\mathcal{E} := E_u + E_v$, $u_n = \mathbf{n}\cdot\nabla u$, and $u_r = \mathbf{r}\cdot\nabla u$, the identity to be checked is

$$\begin{aligned}\frac{\varepsilon^2}{4}\int_\Omega r^2 uv\,d\mathbf{r} = \oint_\Gamma \frac{d-2}{2}(u_n v + v_n u) + \left(\frac{\mathcal{E}}{\varepsilon} - \frac{\varepsilon}{4}r^2\right)(u_n v - v_n u)\\ + r_n\left(\frac{\mathcal{E}}{2}uv - \nabla u\cdot\nabla v\right) + u_r v_n + v_r u_n\,ds.\end{aligned} \tag{C.6}$$

Making the substitutions $u = \phi_i$, $v = \phi_j$, $E_u = E_i$, and $E_v = E_j$ and applying Dirichlet boundary conditions turns the right-hand side into Q_{ij}. More details about how such identities are found using a symbolic matrix method will be postponed to a future publication [7].

Finally, the other possibility is that $E_i = E_j$ (which need not imply $i = j$). We take Lemma 3.1 with the choices $\Omega_A = \Omega$, $u = \phi_i$, $v = \phi_j$, and $E_i = E_j = E$, apply Dirichlet boundary conditions, and use orthonormality (1.3). Thus Lemma B.1 is proved for all choices of i, j. □

Acknowledgments. This work was inspired by questions of Peter Sarnak, with whom the author has had enlightening and stimulating interactions. We are also delighted to have benefited from discussions with Steve Zelditch, Percy Deift, Fanghua Lin, Eduardo Vergini, Doron Cohen, Eric Heller, Kevin Lin, and the detailed and helpful comments of the anonymous reviewer. When the bulk of this work was performed, the author was supported by the Courant Institute at New York University. The author is now at Dartmouth College and supported by National Science Foundation Grant DMS-0507614.

Bibliography

[1] Agam, O.; Fishman, S. Semiclassical criterion for scars in wave functions of chaotic systems. *Phys. Rev. Lett.* **73** (1994), no. 6, 806–809.

[2] Aurich, R.; Taglieber, M. On the rate of quantum ergodicity on hyperbolic surfaces and for billiards. *Phys. D* **118** (1998), no. 1-2, 84–102.

[3] Austin, E. J.; Wilkinson, M. Distribution of matrix elements of a classically chaotic system. *Europhysics Lett.* **20** (1992), 589–593.

[4] Bäcker, A.; Schubert, R.; Stifter, P. On the number of bouncing ball modes in billiards. *J. Phys. A* **30** (1997), no. 19, 6783–6795.

[5] Bäcker, A. Numerical aspects of eigenvalue and eigenfunction computations for chaotic quantum systems. *The mathematical aspects of quantum maps,* 91–144. Lecture Notes in Physics, 618. Springer, Berlin, 2003.

[6] Bäcker, A.; Schubert, R.; Stifter, P. Rate of quantum ergodicity in Euclidean billiards. *Phys. Rev. E (3)* **57** (1998), no. 5, part A, 5425–5447. Erratum: Rate of quantum ergodicity in Euclidean billiards. *Phys. Rev. E (3)* **58** (1998), no. 4, 5192.

[7] Barnett, A. H. Quasi-orthogonality on the boundary for euclidean laplace eigenfunctions. arXiv: math-ph/0601006, 2006.

[8] Barnett, A. H. The scaling method for the Dirichlet eigenproblem. In preparation.

[9] Barnett, A. H. *Dissipation in deforming chaotic billiards*. Doctoral dissertation, Harvard University, 2000. Also available at `http://www.math.dartmouth.edu/~ahb/thesis_html`.

[10] Barnett, A. H.; Cohen, D.; Heller, E. J. Deformations and dilations of chaotic billiards: dissipation rate, and quasi-orthogonality of the boundary wavefunctions. *Phys. Rev. Lett.* **85** (2000), 1412. arXiv:nlin.CD/0003018, 2000.

[11] Barnett, A.; Cohen, D.; Heller, E. J. Rate of energy absorption for a driven chaotic cavity. *J. Phys. A* **34** (2001), no. 3, 413–438. arXiv:nlin.CD/0006041, 2000.

[12] Berry, M. V. Regular and irregular semiclassical wavefunctions. *J. Phys. A* **10** (1977), no. 12, 2083–2091.

[13] Berry, M. V. Quantum scars of classical closed orbits in phase space. *Proc. Roy. Soc. London Ser. A* **423** (1989), no. 1864, 219–231.

[14] Berry, M. Some quantum-to-classical asymptotics. *Chaos et physique quantique (Les Houches, 1989),* 251–304. North-Holland, Amsterdam, 1991.

[15] Betcke, T.; Trefethen, L. Reviving the method of particular solutions. *SIAM Rev.* **47** (2005), no. 3, 469–491.

[16] Bogomolny, E. B. Smoothed wave functions of chaotic quantum systems. *Phys. D* **31** (1988), no. 2, 169–189.

[17] Bogomolny, E. B.; Georgeot, B.; Giannoni, M.-J.; Schmit, C. Chaotic billiards generated by arithmetic groups. *Phys. Rev. Lett.* **69** (1992), no. 10, 1477–1480.

[18] Bohigas, O. Random matrix theories and chaotic dynamics. *Chaos et physique quantique (Les Houches, 1989)*, 87–199. North-Holland, Amsterdam, 1991.

[19] Boose, D.; Main, J. Distributions of transition matrix elements in classically mixed quantum systems. *Phys. Rev. E* **60** (1999), 2813–2844.

[20] Casati, G.; Prosen, T. The quantum mechanics of chaotic billiards. Classical chaos and its quantum manifestations (Toulouse, 1998). *Phys. D* **131** (1999), no. 1-4, 293–310.

[21] Cohen, D. Chaos and energy spreading for time-dependent Hamiltonians, and the various regimes in the theory of quantum dissipation. *Ann. Physics* **283** (2000), no. 2, 175–231.

[22] Colin de Verdière, Y. Ergodicité et fonctions propres du laplacien. *Comm. Math. Phys.* **102** (1985), no. 3, 497–502.

[23] Combescure, M.; Robert, D. Semiclassical sum rules and generalized coherent states. *J. Math. Phys.* **36** (1995), no. 12, 6596–6610.

[24] Donnelly, H. Quantum unique ergodicity. *Proc. Amer. Math. Soc.* **131** (2003), no. 9, 2945–2951.

[25] Eckhardt, B.; Fishman, S.; Keating, J.; Agam, O.; Main, J.; Müller, K. Approach to ergodicity in quantum wave functions. *Phys. Rev. E* **52** (1995), 5893–5903.

[26] Faure, F.; Nonnenmacher, S.; De Bièvre, S. Scarred eigenstates for quantum cat maps of minimal periods. *Comm. Math. Phys.* **239** (2003), no. 3, 449–492.

[27] Feingold, M.; Peres, A. Distribution of matrix elements of chaotic systems. *Phys. Rev. A (3)* **34** (1986), no. 1, 591–595.

[28] Gardiner, C. W. *Handbook of stochastic methods for physics, chemistry and the natural sciences.* 3rd ed. Springer Series in Synergetics, 13. Springer, Berlin, 2004.

[29] Gutzwiller, M. C. *Chaos in classical and quantum mechanics*. Interdisciplinary Applied Mathematics, 1. Springer, New York, 1990.

[30] Heller, E. J. Bound-state eigenfunctions of classically chaotic Hamiltonian systems: scars of periodic orbits. *Phys. Rev. Lett.* **53** (1984), no. 16, 1515–1518.

[31] Heller, Eric J. Wavepacket dynamics and quantum chaology. *Chaos et physique quantique (Les Houches, 1989),* 547–664. North-Holland, Amsterdam, 1991.

[32] Hortikar, S.; Srednicki, M. Random matrix elements and eigenfunctions in chaotic systems. *Phys. Rev. E (3)* **57** (1998), no. 6, 7313–7316.

[33] Kaplan, L.; Heller, E. J. Linear and nonlinear theory of eigenfunction scars. *Ann. Physics* **264** (1998), no. 2, 171–206.

[34] Kaplan, L.; Heller, E. J. Measuring scars of periodic orbits. *Phys. Rev. E (3)* **59** (1999), no. 6, 6609–6628.

[35] Kuttler, J. R.; Sigillito, V. G. Eigenvalues of the Laplacian in two dimensions. *SIAM Rev.* **26** (1984), no. 2, 163–193.

[36] Li, B.; Hu, B. Statistical analysis of scars in stadium billiard. *J. Phys. A* **31** (1998), no. 2, 483–504.

[37] Lindenstrauss, E. Invariant measures and arithmetic quantum unique ergodicity. *Ann. of Math. (2)* **163** (2006), no. 1, 165–219.

[38] Luo, W.; Sarnak, P. Quantum variance for Hecke eigenforms. *Ann. Sci. École Norm. Sup. (4)* **37** (2004), no. 5, 769–799.

[39] Martinez, A. *An introduction to semiclassical and microlocal analysis.* Universitext. Springer, New York, 2002.

[40] Prosen, T. Statistical properties of matrix elements in a Hamilton system between integrability and chaos. *Ann. Physics* **235** (1994), no. 1, 115–164.

[41] Prosen, T. Quantization of a generic chaotic 3D billiard with smooth boundary. I. Energy level statistics. *Phys. Lett. A* **233** (1997), no. 4-6, 323–331.

[42] Prosen, T. Quantization of generic chaotic 3D billiard with smooth boundary. II. Structure of high-lying eigenstates. *Phys. Lett. A* **233** (1997), no. 4-6, 332–342.

[43] Rudnick, Z; Sarnak, P. The behaviour of eigenstates of arithmetic hyperbolic manifolds. *Comm. Math. Phys.* **161** (1994), no. 1, 195–213.

[44] Sarnak, P. Arithmetic quantum chaos. *The Schur lectures (1992) (Tel Aviv)*, 183–236. Israel Mathematical Conference Proceedings, 8. Bar-Ilan University, Ramat Gan, Israel, 1995.

[45] Sarnak, P. Spectra and eigenfunctions of Laplacians. *Partial differential equations and their applications (Toronto, ON, 1995),* 261–276. CRM Proceedings & Lecture Notes, 12. American Mathematical Society, Providence, R.I., 1997.

[46] Sarnak, P. Spectra of hyperbolic surfaces. *Bull. Amer. Math. Soc. (N.S.)* **40** (2003), no. 4, 441–478.

[47] Schanz, H.; Kottos, T. Scars on quantum networks ignore the lyapunov exponent. *Phys. Rev. Lett.* **90** (2003), 234101.

[48] Šnirel′man, A. I. Ergodic properties of eigenfunctions. (Russian) *Uspehi Mat. Nauk* **29** (1974), no. 6 (180), 181–182.

[49] Sinaĭ, Ya. G. Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards. *Uspehi Mat. Nauk* **25** (1970), no. 2 (152), 141–192.

[50] Sivia, D. S. *Data analysis. A Bayesian tutorial.* Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1996.

[51] Steiner, F. *Quantum chaos*. arXiv:chao-dyn/9402001, 1994.

[52] Tanner, G. How chaotic is the stadium billiard? A semiclassical analysis. *J. Phys. A* **30** (1997), no. 8, 2863–2888.

[53] Tate, T. Some remarks on the off-diagonal asymptotics in quantum ergodicity. *Asymptot. Anal.* **19** (1999), no. 3-4, 289–296.

[54] Veble, G.; Robnik, M.; Liu, J. Study of regular and irregular states in generic systems. *J. Phys. A* **32** (1999), no. 36, 6423–6444.

[55] Vergini, E. Estudio cuántico y semiclásico de billares clásicamente caóticos. Doctoral dissertation, Universidad de Buenos Aires, 1995.

[56] Vergini, E.; Saraceno, M. Calculation by scaling of highly excited states of billiards. *Phys. Rev. E* **52** (1995), no. 3, 2204–2207.

[57] Wilkinson, M. A semiclassical sum rule for matrix elements of classically chaotic systems. *J. Phys. A* **20** (1987), no. 9, 2415–2423.

[58] Zelditch, S. Uniform distribution of eigenfunctions on compact hyperbolic surfaces. *Duke Math. J.* **55** (1987), no. 4, 919–941.

[59] Zelditch, S. Quantum transition amplitudes for ergodic and for completely integrable systems. *J. Funct. Anal.* **94** (1990), no. 2, 415–436.

[60] Zelditch, S. Quantum ergodicity on the sphere. *Comm. Math. Phys.* **146** (1992), no. 1, 61–71.

[61] Zelditch, S. On the rate of quantum ergodicity. I. Upper bounds. *Comm. Math. Phys.* **160** (1994), no. 1, 81–92.

[62] Zelditch, S. On the rate of quantum ergodicity. II. Lower bounds. *Comm. Partial Differential Equations* **19** (1994), no. 9-10, 1565–1579.

[63] Zelditch, S. Quantum mixing. *J. Funct. Anal.* **140** (1996), no. 1, 68–86.

[64] Zelditch, S. A random matrix model for quantum mixing. *Internat. Math. Res. Notices* **1996**, no. 3, 115–137.

[65] Zelditch, S. Quantum ergodicity and mixing of eigenfunctions. *Elsevier Encyclopedia of Mathematical Physics*, to appear. arXiv:math-ph/0503026, 2005.

[66] Zelditch, S.; Zworski, M. Ergodicity of eigenfunctions for ergodic billiards. *Comm. Math. Phys.* **175** (1996), no. 3, 673–682.

ALEXANDER BARNETT
Department of Mathematics
Dartmouth College
6188 Bradley Hall
Hanover, NH 03755
E-mail: ahb@gauss.dartmouth.edu

Received July 2004.
Revised February 2006.

Effective scattering coefficient of the cerebral spinal fluid in adult head models for diffuse optical imaging

Anna Custo, William M. Wells III, Alex H. Barnett, Elizabeth M. C. Hillman, and David A. Boas

An efficient computation of the time-dependent forward solution for photon transport in a head model is a key capability for performing accurate inversion for functional diffuse optical imaging of the brain. The diffusion approximation to photon transport is much faster to simulate than the physically correct radiative transport equation (RTE); however, it is commonly assumed that scattering lengths must be much smaller than all system dimensions and all absorption lengths for the approximation to be accurate. Neither of these conditions is satisfied in the cerebrospinal fluid (CSF). Since line-of-sight distances in the CSF are small, of the order of a few millimeters, we explore the idea that the CSF scattering coefficient may be modeled by any value from zero up to the order of the typical inverse line-of-sight distance, or approximately 0.3 mm^{-1}, without significantly altering the calculated detector signals or the partial path lengths relevant for functional measurements. We demonstrate this in detail by using a Monte Carlo simulation of the RTE in a three-dimensional head model based on clinical magnetic resonance imaging data, with realistic optode geometries. Our findings lead us to expect that the diffusion approximation will be valid even in the presence of the CSF, with consequences for faster solution of the inverse problem. © 2006 Optical Society of America

OCIS codes: 170.6960, 170.3660, 170.5280, 170.6920.

1. Introduction

Diffuse optical imaging (DOI) is a relatively new method used to image blood oxygenation *in vivo*. It uses near-infrared light and has the advantage of low cost and portability. The success of DOI techniques is due to the properties of near-infrared light in biological tissue. The absorption coefficient (μ_a) depends on the total hemoglobin concentration and oxygenation within the tissue; therefore calculating μ_a provides useful information about the physiological conditions of the tissue.[1] For instance, during the past few years DOI has been tested for application to imaging breast cancer[2–9] and brain function.[10–16]

In DOI, near-infrared light is scattered in a medium with optical properties x, and some fluence y is recorded at the detectors' positions. Solving an imaging problem [described as $f(x) = y$; solved for x] requires a good combination of a forward and an inverse model. The forward problem models the process producing the set of measurements [establishing the rules to calculate $f(x)$]. The inverse problem arises when it is necessary to recover an image of the optical properties of the medium from the observed data [i.e., $x = f^{-1}(y)$]. The DOI image reconstruction problem is ill posed; hence the inverse procedure typically involves the use of regularization techniques.[17,18] Because of its relatively poor spatial resolution, DOI is increasingly combined with other imaging techniques, such as magnetic resonance imaging (MRI) and x ray, which provide high-resolution structural information to guide the characterization of the unique physiological information offered by DOI.[8,19–22]

Several papers have been published by Okada *et al.*,[23] Fukui *et al.*,[24] Hayashi *et al.*,[25] Koyama *et al.*,[26] Arridge *et al.*[27,28] Firbank *et al.*,[29] and Hielscher *et al.*[30] that explore a variety of models of light propagation in highly scattering media with properties

A. Custo (custom@csail.mit.edu) is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 and the Athinoula A. Martinos Center for Biomedical Engineering, Massachusetts General Hospital, Harvard Medical School, Charlestown, Massachusetts 02129. W. M. Wells III is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 and the Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115. A. H. Barnett, E. M. C. Hillman, and D. A. Boas are with the Athinoula A. Martinos Center for Biomedical Engineering, Massachusetts General Hospital, Harvard Medical School, Charlestown, Massachusetts 02129.

Received 4 November 2005; revised 29 December 2005; accepted 3 January 2006; posted 18 January 2006 (Doc. ID 65791).
0003-6935/06/194747-09$15.00/0
© 2006 Optical Society of America

similar to those found in the human head. The low-scattering properties of the cerebral spinal fluid (CSF) surrounding the space between the brain and the dura mater[31] have been of particular concern in the development of an accurate photon migration forward problem for the human head because the diffusion equation is known to provide inaccurate solutions under such circumstances.[23,30,31] Controversy has continued as to whether this low-scattering region (often called the void region) affects sensitivity in the underlying brain tissue.[24,32,33] As a result, several papers have been published exploring the implementation of the radiative transport equation[30,34] (RTE) or hybrid combinations of the radiosity equation and the diffusion equation.[25,26] A number of papers have shown that in some idealized geometries the transport equation is necessary to accurately describe photon migration in the presence of a low-scattering region.[25,30,33,35,36] Transport-based modeling approaches, such as Monte Carlo (MC) simulations, typically require excessively long processing times. This can make them unsuitable for use as the forward model in an effective image reconstruction algorithm. It is therefore desirable to adopt a faster alternative forward model with comparable accuracy. A solution of the diffusion equation by finite difference (FD), for instance, offers greater computational speed at the cost of reduced modeling accuracy.[30,37] It has been suggested[23,37] that the diffusion equation may be sufficiently accurate given the folding nature of the low-scattering region between the brain and the skull and the fact that this space is filled with connective tissue and blood vessels.[38]

The irregularity in the thickness of the CSF layer is of particular interest because it limits the average straight-line distance that a photon would travel in the void region. Thus even if the scattering coefficient μ_s' in the void region were zero, an increase in μ_s' up to the value of the inverse typical straight-line distance is expected to cause little change in photon transport. Barnett *et al.*[37] hypothesized that, since the line of sight of light propagation through the CSF is likely to be an average ≤ 3 mm, approximating the CSF reduced scattering coefficient with a value not larger than its inverse (which is $\sim 0.3\ \mathrm{mm}^{-1}$) would not introduce a large error. Therefore keeping the overall goal of rapid diffusion modeling in mind, we conjecture that we can approximate the CSF scattering coefficient with values between 0.1 and $0.3\ \mathrm{mm}^{-1}$ without introducing an error larger than 20%.

DOI can be achieved by using continuous wave (cw), time domain (TD), and frequency domain measurements. TD measurements of functional brain activity have recently been shown to provide dramatically enhanced sensitivity to cortical activation.[39–42] In this paper we consider both cw and TD propagation of light through the human head. We expect that our conclusions about the suitability of a larger scattering coefficient to approximate CSF will extend to frequency domain measurements.

We recently began exploring the use of 3D MRI of the human head as a spatial prior step in an iterative reconstruction process of the optical properties of the different structures in the human head[37] as well as for improving quantitative functional imaging.[22,42] This work required creation of a MRI-based anatomically correct 3D model of the head and brain. In this paper we used MC modeling, which implements the transport equation,[44,45] to accurately simulate light propagation through this MRI-based 3D model of the human head. We simulated both time-resolved and cw measurements to validate our hypothesis of an effective CSF scattering coefficient by measuring the error introduced when using a CSF scattering coefficient larger than a near zero value (we used the reference CSF scattering coefficient of $0.001\ \mathrm{mm}^{-1}$). Given this new larger effective μ_s', the conditions for validity of the diffusion approximation may hold to a much greater degree than previously thought, allowing accurate diffusion solution.

We calculated the deviation from our reference measurements of the photon fluence detected on the surface of the head and of the sensitivity to the brain. The results presented in this paper support our hypothesis that the nonscattering CSF region can be treated with a larger scattering coefficient ($0.3\ \mathrm{mm}^{-1}$) with only a 20% difference in the measurements. This result will also have relevance for the debate on the exact CSF scattering coefficient, since it demonstrates that, as long as it is less than the inverse typical line-of-sight distances, its exact value is not relevant. Thus it is possible in principle that the diffusion equation may provide sufficiently accurate modeling of photon migration through the human head, greatly reducing the computational expense.

2. Methods

A. Head Model and Probe Placement

We used segmented MRI data to create the head geometry that we employed for this study. Within this adult head model we distinguished three tissue types (extracerebral, CSF, and cerebral, as described in Table 1). The whole volume is voxelized in a cube with 128 voxels/side (a total of 128^3 voxels, $2 \times 2 \times 2\ \mathrm{mm}^3$ each). A coronal slice of the 3D head model is shown in Fig. 1 with a voxel size of 1 mm. This voxel size was increased to 2 mm for the simulation, decreasing the simulation run time and the memory requirement. In data not reported we established that this increase in voxel size did not significantly affect the simulated data. The gray diamond and the circles on the surface of the head show the placement of the source and detectors, respectively.

The optical properties provided in Table 1 for each tissue type were taken from the literature.[46–48] The refractive index was the same for all tissues and assumed to be 1, and the scattering anisotropy $g = 0.01$. We chose a value of μ_s' for CSF between 0.001 and $1\ \mathrm{mm}^{-1}$ to test our hypothesis that the value of the CSF scattering coefficient could be mod-

Table 1. Optical Properties of the Adult Head Model

Tissue Type	Reduced Transport Scattering Coefficient (mm^{-1})	Absorption Coefficient (mm^{-1})	Tissue Thickness (mm)
Scalp and skull	0.86	0.019	3–8 (scalp) 7–8 (skull)
CSF	0.001, 0.01, 0.1, 0.2, 0.3, 0.7, 1	0.004	2–4
Gray and white matter	1.11	0.01	4–10 (gray) >40 (white)

eled as the inverse of the average CSF layer thickness.[37] In our model we calculated $\mu_{s,\mathrm{CSF}}' = 1/\text{average thickness}_{\mathrm{CSF}}$, which gave us a μ_s' of $\sim 0.25\ \mathrm{mm}^{-1}$.

For the purpose of this study, we used a linear probe geometry, in which the sources and detectors were positioned along a line on the surface of the head, all placed in the same coronal plane. The probe featured a single source and 25 detectors placed at 12, 14, 16, 18 . . . 60 mm from the source position. The source was located at the top of the head (see Fig. 1). We chose this particular arrangement of optodes to analyze the effect of the CSF scattering coefficient on the measured photon fluence as a function of the separation between the source and the detector. Furthermore, the same probe has been used in similar studies by Okada *et al.*[23]

B. Solution of the Radiative Transport Equation

MC modeling is a simple method that offers a great deal of freedom in defining geometries and optical properties based on the RTE.[44,45] The method models photon trajectories through heterogeneous tissues, reproducing the randomness of each scattering event in a stochastic fashion (a random seed is employed). When a photon is detected, its partial optical path length for each of the tissue types through which it passed is recorded in a history file. MC methods have the disadvantage of being computationally expensive to use while obtaining a good signal-to-noise ratio (SNR) in highly scattering thick tissues. We typically ran 10^9 photons to achieve an appropriate SNR for the results presented in this paper. We ran 11 independent MC simulations of 10^8 photons for each μ_s' configuration so that we could calculate the standard deviation across the independent runs. Each simulation took approximately 12 h (see Boas *et al.*[45] for more details). By recording the path length of each photon, our MC data could be converted to cw or TD.

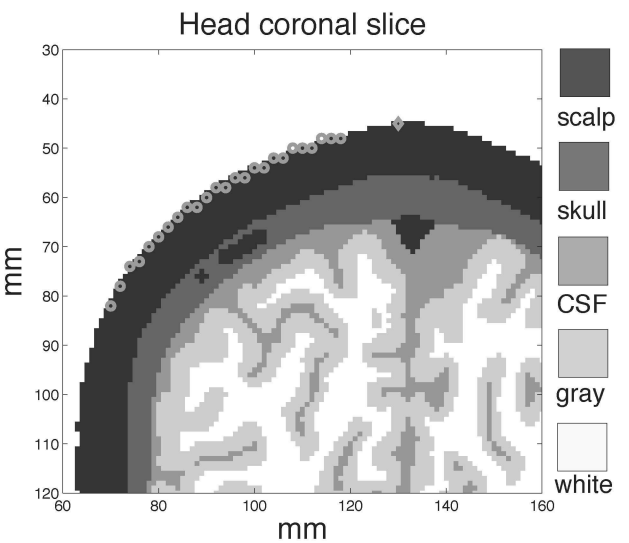


Fig. 1. Head geometry and probe placement. The gray diamond on the top of the head indicates the position of the single source and the gray circles show the position of the 25 detectors.

C. Calculation of the Total Fluence in Time Domain and Continuous Wave

To consider the time-resolved problem from the data recorded by the MC simulation we calculated the temporal point spread function by using Eq. (2) from Ref. 45:

$$\Phi_j(t_i) = \frac{1}{N_j(t_i)\Delta t} \sum_{l=1}^{N_j(t_i)} \prod_{m=1}^{N_R} \exp(-\mu_{a,m} L_{j,l,m}), \qquad (1)$$

where $\Phi_j(t_i)$ is the measured photon fluence at detector j, $N_j(t_i)$ is the number of photons collected at detector j in a time gate of width Δt centered at time t_i, $\exp(-\mu_{a,m} L_{j,l,m})$ accounts for the effects of absorption in each region in which $L_{j,l,m}$ is the path length of photon l through region m, and the photon migration time is related to the photon path length by the speed of light in the medium. N_R is the number of regions through which the photons migrate. The cw fluence is calculated by averaging over the time index i.

D. Calculation of Partial Optical Path Length Factor

Tissue scattering causes the photons to travel a greater distance than the geometric distance between the source and the detector. The partial path length factor (PPF) of light through each of the tissue types is defined as[24,38,39,49]

$$\mathrm{PPF}_m = \partial \mathrm{OD}/\partial \mu_{a,m}, \qquad (2)$$

where $\mathrm{OD} = -\ln(\Phi/\Phi_0)$, Φ/Φ_0 is given by Eq. (1), and Φ_0 is the incident number of photons. The partial path length is thus easily derived from Eq. (1):

$$\mathrm{PPF}_{j,m} = \frac{\sum_{l=1}^{N_j(t_i)} \Pi_{m=1}^{N_S} L_{j,l,m} \exp(-\mu_{a,m} L_{j,l,m})}{\sum_{l=1}^{N_j(t_i)} \Pi_{m=1}^{N_R} \exp(-\mu_{a,m} L_{j,l,m})}. \qquad (3)$$

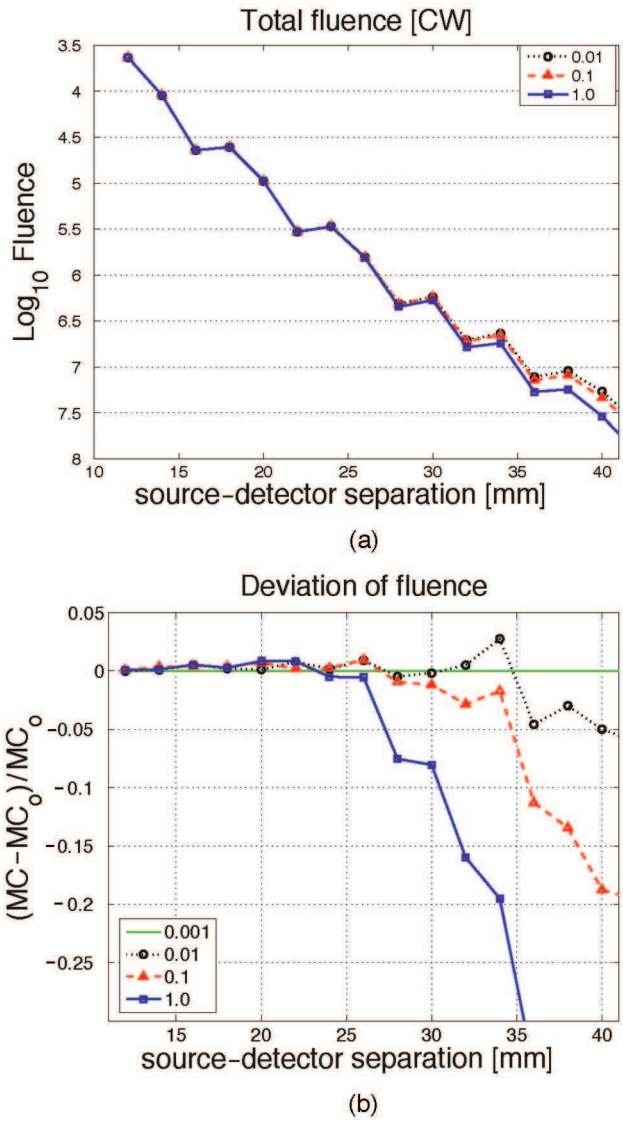


Fig. 2. (Color online) (a) Total detected fluence simulated with Monte Carlo in cw. (b) Relative fluence in cw calculated with respect to MC_o, which is the MC prediction when $\mu_{s,CSF}' = 0.001$ mm^{-1}.

3. Results

A. Fluence and Partial Path Length for Continuous Wave

In Fig. 2 we show the total measured fluence versus the source–detector separation for different $\mu_{s,CSF}'$ [Fig. 2(a)] and the fractional difference relative to $\mu_{s,CSF}' = 0.001$ mm^{-1} [Fig. 2(b)]. The total fluence is normalized by the incident fluence, i.e., the total number of photons launched in the MC simulation. We see that the detected fluence varies little when the $\mu_{s,CSF}' \leq 0.1$ mm^{-1} but that a significant difference occurs at larger $\mu_{s,CSF}'$ for separations greater than 25 mm. In Fig. 2(b) we see that at small separations (12–22 mm) the fractional difference for increasing $\mu_{s,CSF}'$ is less than 2%, becoming significant (>20%) at larger separations (>32 mm) when $\mu_{s,CSF}'$ is 1.0 mm^{-1}. This is to be expected because smaller separations are predominantly sensitive to the scalp–skull and therefore do not probe the CSF layer.

In Fig. 3(a) we show the cw partial path length for the scalp–skull region and the brain region versus the source–detector separation for different scattering coefficients in the CSF region. The results show that the sensitivity of the measurement to absorption changes does not change as μ_s' in the CSF space increases from 0.01 to 0.1 mm^{-1}, but that a change is observed with $\mu_s' = 1.0$ mm^{-1}. This is consistent with the hypothesis that the sensitivity will change as the scattering length becomes smaller than the typical line-of-sight distance through the CSF, which is approximately 3 mm in our model.

In Figs. 3(b) and 3(c) we plotted the fractional difference in the PPF for the scalp–skull region and the brain region relative to that when $\mu_{s,CSF}' = 0.001$ mm^{-1}. In Fig. 3(b) we observe a difference greater than 3% at larger separations (>30 mm) for a CSF model with $\mu_s' = 1.0$ mm^{-1}, and a difference less than 1% for $\mu_{s,CSF}' = 0.1$ and 0.01 mm^{-1}. Figure 3(c) shows the same fractional difference observed in the brain. In the brain, even at small separations, we observed a large difference only when $\mu_{s,CSF}' = 1.0$ mm^{-1}, increasing from 20% to 47% as the separation increases. For the smaller $\mu_{s,CSF}'$ we observed differences less than 10% at all separations.

To investigate the variation in more detail, we plotted in Figs. 4 and 5 the deviation in detected photon fluence at the surface of the scalp (Fig. 4) and the deviation in the partial path length in the scalp–skull and brain (Fig. 5) versus μ_s' of the CSF at source–detector separations of 20, 30, and 40 mm. These results show that a change greater than 20% is not observed until $\mu_s' > 0.3$ mm^{-1}, except at a 40 mm separation. We believe that the large discrepancy observed at 40 mm is due to the weakness of the signal reaching the far detectors. These results support our hypothesis that we can treat this voidlike region with a larger scattering coefficient ($0.1 < \mu_s' < 0.3$ mm^{-1}) and obtain similar results (errors between 10% and 20% for cw measurements with a source–detector spacing of <40 mm).

We also note from Figs. 3(a) and 5 that the brain PPF, which corresponds to the sensitivity of the cw measurement to absorption changes in the brain, is higher when the CSF μ_s' is low than when the CSF μ_s' matches that of the brain.

B. Fluence and Partial Path Length in the Time Domain

We further explored this result for time-resolved photon migration. In Fig. 6(a) we show how the partial path length varied with the photon transit time in the medium. These results confirm what we have seen in cw, that is, the time-resolved fluence is approximately the same when $\mu_{s,CSF}'$ is less than 0.1 mm^{-1} but a significant difference is observed for a larger scattering coefficient. In Fig. 6(b) we quantified this change relative to the fluence when $\mu_{s,CSF}' = 0.001$ mm^{-1} and observed a greater than 50% de-

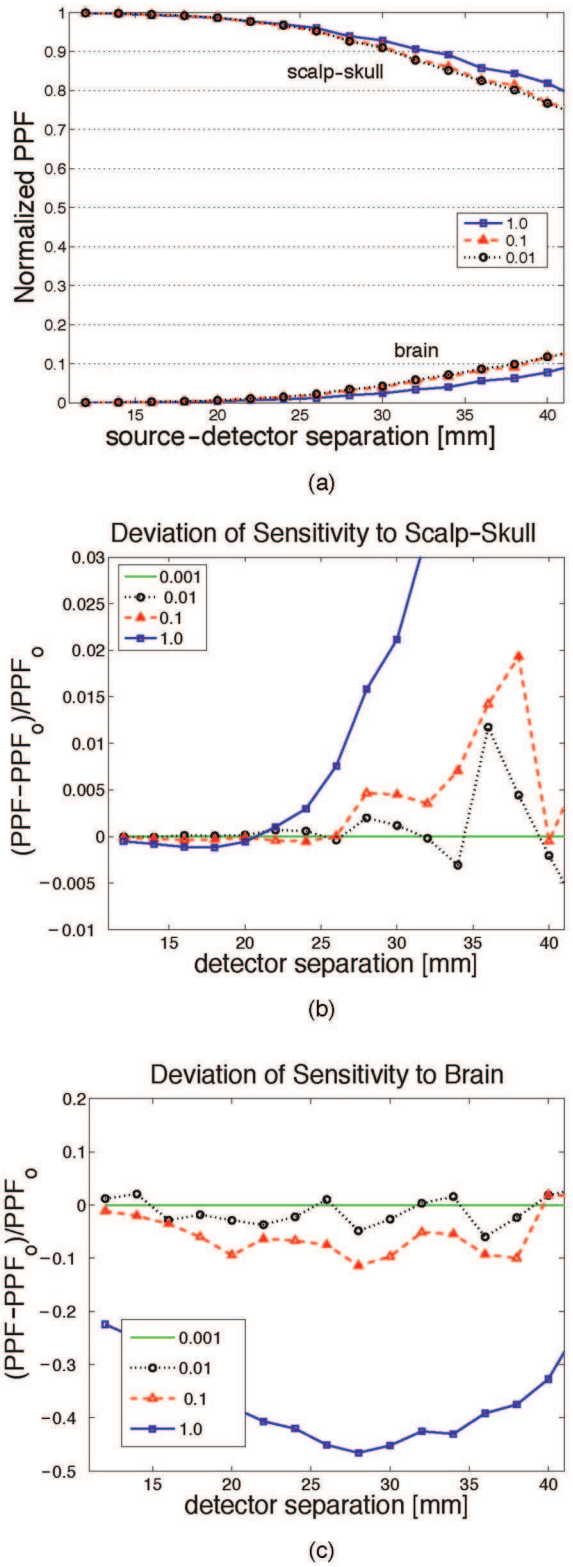


Fig. 3. (Color online) (a) MC normalized path length factor calculated versus separation for three different $\mu_{s,\mathrm{CSF}}'$: 0.01, 0.1, and 1.0 mm^{-1} in cw. The PPF is normalized by the total sensitivity to all tissue types. MC measure of relative sensitivity to (b) scalp–skull layer and (c) brain versus separation when varying $\mu_{s,\mathrm{CSF}}'$ in cw ($\mu_{s,\mathrm{CSF}}'$ = 0.01, 0.1, and 1.0 mm^{-1}). The error is calculated with respect to PPF_o, which is the MC prediction of PPF when $\mu_{s,\mathrm{CSF}}'$ = 0.001 mm^{-1}.

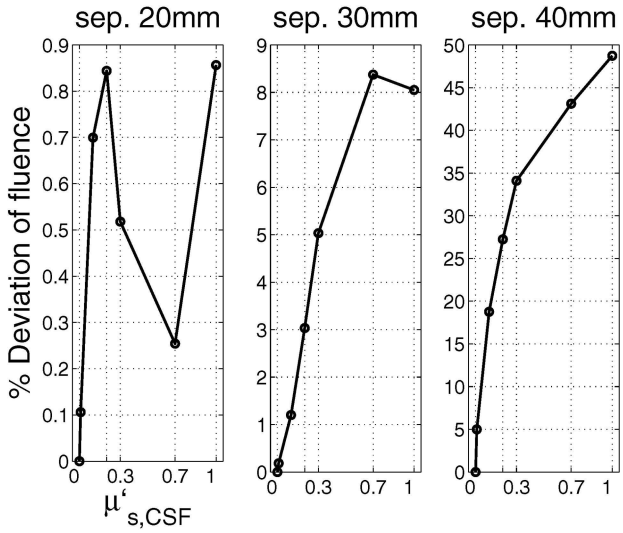


Fig. 4. Fluence changes as a function of $\mu_{s,\mathrm{CSF}}'$. CSF scattering coefficient varies from 0.001 to 1.0 mm^{-1}. The data are calculated via MC simulations in cw using $\mu_{s,\mathrm{CSF}}'$ values of 0.001, 0.01, 0.1, 0.2, 0.3, 0.7, and 1.0.

viation for $\mu_{s,\mathrm{CSF}}'$ = 1.0 mm^{-1} and a mostly less than 8% deviation when $\mu_{s,\mathrm{CSF}}'$ = 0.1 and 0.01 mm^{-1}.

In Fig. 7 we explored how the CSF scattering coefficient affects the partial path length in the superficial layers and the brain versus time delay. The absolute partial path length in Fig. 7(a) shows that no difference is observed for $\mu_{s,\mathrm{CSF}}'$ = 0.1 and 0.01 mm^{-1} but that a difference is observed when μ_s' is further increased to 1.0 mm^{-1}. The deviation with increasing $\mu_{s,\mathrm{CSF}}'$ relative to a $\mu_{s,\mathrm{CSF}}'$ of 0.001 mm^{-1} is shown in Figs. 7(b) and 7(c) for the superficial and brain re-

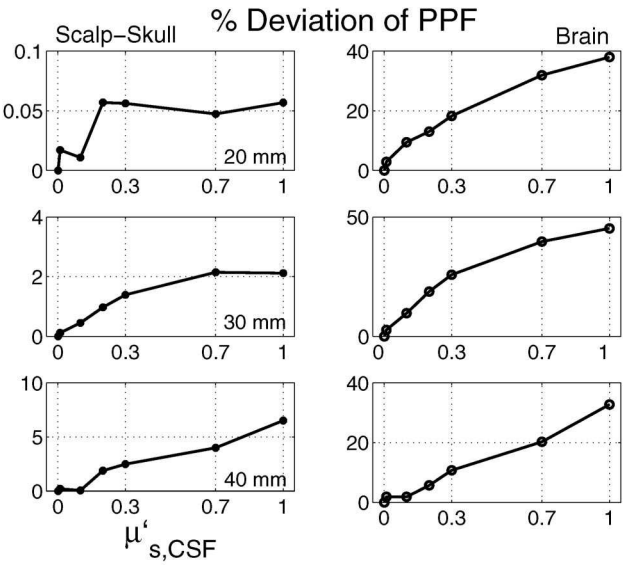


Fig. 5. Partial path length absolute changes as a function of $\mu_{s,\mathrm{CSF}}'$ scattering coefficient varies from 0.001 to 1.0 mm^{-1} taking values of 0.001, 0.01, 0.1, 0.2, 0.3, 0.7, and 1.0, and the PPF is simulated with MC in cw. Results are shown for separations of 20, 30, and 40 mm.

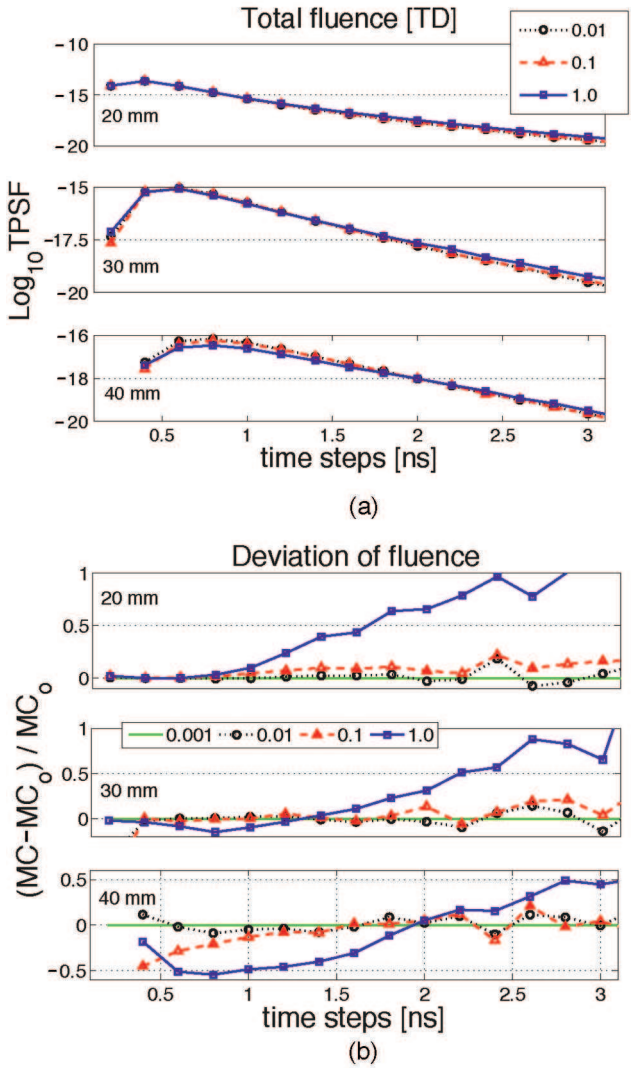


Fig. 6. (Color online) (a) Temporal point spread function predicted by MC and (b) its relative error with respect to the reference measurement MC_o, simulated with $\mu_{s,CSF}' = 0.001\ mm^{-1}$. Results are shown for separations of 20, 30, and 40 mm.

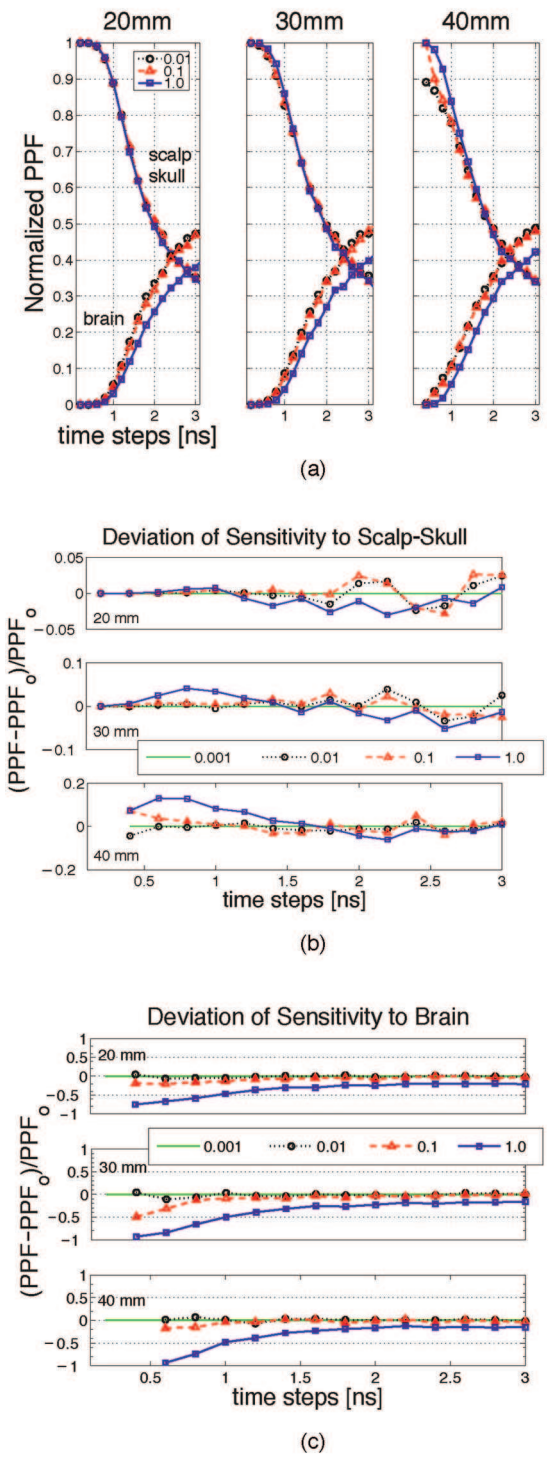


Fig. 7. (Color online) (a) Monte Carlo prediction of the optical pathlength factor at three $\mu_{s,CSF}' = 0.01$, 0.1, and 1.0 mm^{-1} versus time delay normalized by the total PPF. Results are shown for separations of 20, 30, and 40 mm. (b) Relative sensitivity to absorption changes in the scalp–skull layer when $\mu_{s,CSF}' = 0.01$, 0.1, and 1.0 mm^{-1} versus time delay as predicted by Monte Carlo. The reference measure of sensitivity to scalp–skull is given by simulating PPF when $\mu_{s,CSF}'$ is 0.001 mm^{-1}. (c) Time-resolved Monte Carlo predictions of the relative sensitivity to absorption changes in the brain when $\mu_{s,CSF}'$ assumes the values of 0.01, 0.1 and 1.0 mm^{-1}. The reference measure of sensitivity to brain is given by simulating PPF when $\mu_{s,CSF}'$ is 0.001 mm^{-1}. Results are shown for separations of 20, 30, and 40 mm.

gions, respectively. Little deviation is observed in the superficial scalp–skull region since the deviation is never larger than 15%. A more significant difference is observed for the brain, where the partial path length is underestimated by 20%–50% for $\mu_{s,CSF}' = 1.0\ mm^{-1}$, whereas the error is less than 20% when $\mu_{s,CSF}' < 0.3\ mm^{-1}$. In all cases, the error is greater at the earliest photon migration times when brain sensitivity is weakest.

Figures 8 and 9 show the deviation in the detected fluence and PPF, respectively, versus $\mu_{s,CSF}'$ at 20, 30, and 40 mm at a time delay of 2 ns, where we see that changes greater than 20% occur only when $\mu_s' > 0.3\ mm^{-1}$. These results further support the observation that the CSF layer can be accurately characterized by a scattering coefficient value between 0 and $\sim 0.3\ mm^{-1}$ and still provide detected

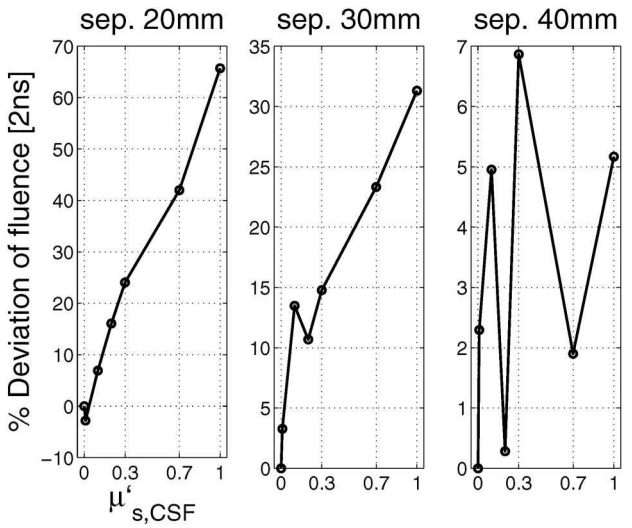


Fig. 8. Fluence changes as a function of $\mu_{s,CSF}'$ for different separations at a time delay of 2 ns. CSF scattering coefficient varies between 0.001 and 1.0 mm^{-1}.

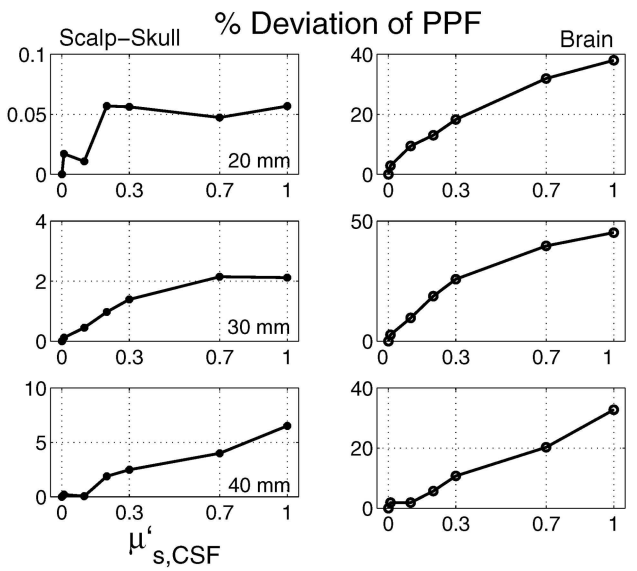


Fig. 9. Absolute changes in the partial optical path length versus $\mu_{s,CSF}'$ at source–detector separations of 20, 30, and 40 mm for a time delay of 2 ns.

fluence and brain sensitivity within an acceptable approximation error (between 10% and 20%).

As with the cw results, we note that for the TD measurements the partial path length in the brain also increases when the CSF reduced scattering coefficient is low [see Figs. 7(a) and 9]. Also, importantly, sensitivity to absorption changes in the brain is significantly enhanced at time delays greater than 1.5 ns with respect to cw sensitivity measurements [compare Figs. 3(a) and 7(a)].

4. Conclusions

We performed extensive simulation studies to quantify the deviation in photon migration measurements and sensitivity to brain activity given a range of CSF scattering coefficient values. Through an analysis of total fluence and partial optical path length using MC simulations in an accurate MRI-based 3D head geometry, we found that the CSF scattering coefficient can increase up to the inverse of its typical thickness without significant variation from a near zero scattering coefficient. The results support our initial hypothesis that an effective CSF scattering coefficient of approximately 0.3 mm^{-1} can be used. Under these circumstances it may be possible to obtain accurate solutions of the forward problem with diffusion approximation. The advantage of using diffusion approximation is that we can utilize faster algorithms to simulate photon migration in the adult head.

Our results also suggest that low-scattering CSF increases DOI measurement sensitivity to brain activity in contrast to previous studies that assumed a simplified smooth CSF layer.[32,42,50,51] CSF may change the depth sensitivity profile, but this does not mean that the signal from the cortex is decreased in the presence of low-scattering CSF. We hypothesize that the presence of the CSF layer has the effect of concentrating measurement sensitivity to the more superficial layers of the cortex, but that the overall sensitivity to cortical hemodynamics is not adversely affected.

In conclusion, our results indicate the following:

(i) Using a diffusion model with a CSF reduced scattering coefficient of approximately 0.3 mm^{-1} leads to measurements with errors no larger than 20%, for both TD and cw.

(ii) The sensitivity of DOI measurements to cortical activity is not adversely affected by the presence of CSF in a realistic 3D head geometry.

(iii) TD measurements can further reduce the effect of the CSF layer by increasing the sensitivity to deeper tissues, in agreement with the previous findings of Steinbrink *et al.*,[47] Montcel *et al.*,[41] and Okada and Delpy.[42]

We thank Jon Stott for useful conversations and guidance in the early stages of this work. This work was supported by the National Institutes of Health, (P41-RR14075 and R01-EB002482) and the National Science Foundation under grant IIS 9610249.

References

1. V. Chernomordik, D. W. Hattery, D. Grosenick, H. Wabnitz, H. Rinneberg, K. T. Moesta, P. M. Schlag, and A. Gandjbakhche, "Quantification of optical properties of a breast tumor using random walk theory," J. Biomed. Opt. **7,** 80–87 (2002).
2. N. Shah, A. E. Cerussi, D. Jakubowski, D. Hsiang, J. Butler, and B. J. Tromberg, "The role of diffuse optical spectroscopy in the clinical management of breast cancer," Dis. Markers **19,** 95–105 (2003).
3. R. Choe, A. Corlu, K. Lee, T. Durduran, S. D. Konecky, M. Grosicka-Koptyra, S. R. Arridge, B. J. Czerniecki, D. L. Fraker, A. DeMichele, B. Chance, M. A. Rosen, and A. G. Yodh, "Diffuse optical tomography of breast cancer during neoadjuvant chemotherapy: a case study with comparison to MRI," Med. Phys. **32,** 1128–1139 (2005).
4. X. Intes, J. Ripoll, Y. Chen, S. Nioka, A. G. Yodh, and B. Chance, "In vivo continuous-wave optical breast imaging en-

hanced with Indocyanine Green," Med. Phys. **30,** 1039–1147 (2003).
5. P. Taroni, G. Danesini, A. Torricelli, A. Pifferi, L. Spinelli, and R. Cubeddu, "Clinical trial of time-resolved scanning optical mammography at 4 wavelengths between 683 and 975 nm," J. Biomed. Opt. **9,** 464–473 (2004).
6. D. Grosenick, H. Wabnitz, K. T. Moesta, J. Mucke, M. Moller, C. Stroszczynski, J. Stossel, B. Wassermann, P. M. Schlag, and H. Rinneberg, "Concentration and oxygen saturation of haemoglobin of 50 breast tumours determined by time-domain optical mammography," Phys. Med. Biol. **49,** 1165–1181 (2004).
7. S. Srinivasan, B. W. Pogue, S. Jing, H. Dehghani, C. Kogel, S. Soho, J. J. Gibson, T. D. Tosteson, S. P. Poplack, and K. D. Paulsen, "Interpreting hemoglobin and water concentration, oxygen saturation, and scattering measured in vivo by near-infrared breast tomography," Proc. Natl. Acad. Sci. U.S.A. **100,** 12349–12354 (2003).
8. A. Li, E. L. Miller, M. E. Kilmer, T. J. Brukilacchio, T. Chaves, J. Stott, Q. Zhang, T. Wu, M. Chorlton, R. H. Moore, D. B. Kopans, and D. A. Boas, "Tomographic optical breast imaging guided by three-dimensional mammography," Appl. Opt. **42,** 5181–5190 (2003).
9. M. A. Franceschini, K. T. Moesta, S. Fantini, G. Gaida, E. Gratton, H. Jess, W. W. Mantulin, M. Seeber, P. M. Schlag, and M. Kaschke, "Frequency-domain techniques enhance optical mammography: initial clinical results," Proc. Natl. Acad. Sci. U.S.A. **94,** 6468–6473 (1997).
10. A. Villringer, J. Planck, C. Hock, L. Schleinkofer, and U. Dirnagl, "Near infrared spectroscopy (NIRS): a new tool to study hemodynamic changes during activation of brain function in human adults," Neurosci. Lett. **154,** 101–104 (1993).
11. Y. Hoshi and M. Tamura, "Detection of dynamic changes in cerebral oxygenation coupled to neuronal function during mental work in man," Neurosci. Lett. **150,** 5–8 (1993).
12. M. A. Franceschini, S. Fantini, J. H. Thompson, J. P. Culver, and D. A. Boas, "Hemodynamic evoked response of the sensorimotor cortex measured non-invasively with near infrared optical imaging," Psychophysiology **40,** 548–560 (2003).
13. H. Koizumi, T. Yamamoto, A. Maki, Y. Yamashita, H. Sato, H. Kawaguchi, and N. Ichikawa, "Optical topography: practical problems and new applications," Appl. Opt. **42,** 3054–3062 (2003).
14. M. Pena, A. Maki, D. Kovacic, G. Dehaene-Lambertz, H. Koizumi, F. Bouquet, and J. Mehler, "Sounds and silence: an optical topography study of language recognition at birth," Proc. Natl. Acad. Sci. U.S.A. **100,** 11702–11705 (2003).
15. E. Gratton, V. Toronov, U. Wolf, M. Wolf, and A. Webb, "Measurement of brain activity by near-infrared light," J. Biomed. Opt. **10,** 011008 (2005).
16. T. Wilcox, H. Bortfeld, R. Woods, E. Wruck, and D. A. Boas, "Using near-infrared spectroscopy to assess neural activation during object processing in infants," J. Biomed. Opt. **10,** 011010 (2005).
17. S. R. Arridge, "Optical tomography in medical imaging," Inverse Probl. **15,** 41–93 (1999).
18. M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging* (Institute of Physics, 1998).
19. R. L. Barbour, H. L. Graber, J. Change, S. S. Barbour, P. C. Koo, and R. Aronson, "MRI-guided optical tomography: prospects and computation for a new imaging method," IEEE Comput. Sci. Eng. **2,** 63–77 (1995).
20. B. W. Pogue and K. D. Paulsen, "High-resolution near-infrared tomographic imaging simulations of the rat cranium by use of a priori magnetic resonance imaging structural information," Opt. Lett. **23,** 1716–1718 (1998).
21. V. Ntziachristos, A. G. Yodh, M. Schnall, and B. Chance, "MRI-guided diffuse optical spectroscopy of malignant and benign breast lesions," Neoplasia **4,** 347–354 (2002).
22. D. A. Boas and A. M. Dale, "Simulation study of magnetic resonance imaging-guided cortically constrained diffuse optical tomography of human brain function," Appl. Opt. **44,** 1957–1968 (2005).
23. E. Okada, M. Firbank, M. Schweiger, S. R. Arridge, M. Cope, and D. T. Delpy, "Theoretical and experimental investigation of near-infrared light propagation in a model of the adult head," Appl. Opt. **36,** 21–31 (1997).
24. Y. Fukui, Y. Ajichi, and E. Okada, "Monte Carlo prediction of near-infrared light propagation in realistic adult and neonatal head models," Appl. Opt. **42,** 2881–2887 (2003).
25. T. Hayashi, Y. Kashio, and E. Okada, "Hybrid Monte Carlo-diffusion method for light propagation in tissue with a low-scattering region," Appl. Opt. **42,** 2888–2896 (2003).
26. T. Koyama, A. Iwasaki, Y. Ogoshi, and E. Okada, "Practical and adequate approach to modeling light propagation in an adult head with low-scattering regions by use of diffusion theory," Appl. Opt. **44,** 2094–2103 (2005).
27. S. R. Arridge, M. Schweiger, M. Hiraoka, and D. T. Delpy, "A finite element approach for modeling photon transport in tissue," Med. Phys. **20,** 299–309 (1993).
28. S. R. Arridge, H. Dehghani, M. Schweiger, and E. Okada, "The finite element model for the propagation of light in scattering media: a direct method for domains with nonscattering regions," Med. Phys. **27,** 252–264 (2000).
29. M. Firbank, S. R. Arridge, M. Schweiger, and D. T. Delpy, "An investigation of light transport through scattering bodies with non-scattering regions," Phys. Med. Biol. **41,** 767–783 (1996).
30. A. H. Hielscher, R. E. Alcouffe, and R. L. Barbour, "Comparison of finite-difference transport and diffusion calculations for photon migration in homogeneous and heterogeneous tissues," Phys. Med. Biol. **43,** 1285–1302 (1998).
31. E. M. C. Hillman, *Experimental and theoretical investigations of near infrared tomographic imaging methods and clinical applications* (University College London, 2002).
32. H. Dehghani, S. R. Arridge, M. Schweiger, and D. T. Delpy, "Optical tomography in the presence of void regions," J. Opt. Soc. Am. A **17,** 1659–1670 (2000).
33. E. Okada and D. T. Delpy, "Effect of discrete scatterers in CSF layer on optical path length in the brain," in *Photon Migration, Diffuse Spectroscopy, and Optical Coherence Tomography: Imaging and Functional Assessment*, S. Andersson-Engels and J. G. Fujimoto, eds., Proc. SPIE **4160,** 196–203 (2000).
34. G. Bal and K. Ren, "Generalized diffusion model in optical tomography with clear layers," J. Opt. Soc. A **20,** 2355–2364 (2003).
35. J. D. Riley, S. R. Arridge, Y. Chrysanthou, H. Dehghani, E. M. Hillman, and M. Schweiger, "Radiosity diffusion model in 3D," in *Photon Migration, Optical Coherence Tomography, and Microscopy*, S. Andersson-Engels and M. F. Kaschke, eds., Proc. SPIE **4431,** 153–164 (2001).
36. H. Dehghani, D. T. Delpy, and S. R. Arridge, "Photon migration in non-scattering tissue and the effects on image reconstruction," Phys. Med. Biol. **44,** 2897–2906 (1999).
37. A. H. Barnett, J. P. Culver, A. G. Sorensen, A. Dale, and D. A. Boas, "Robust inference of baseline optical properties of the human head with 3D segmentation from magnetic resonance imaging," Appl. Opt. **42,** 3095–3108 (2003).
38. E. Okada and D. T. Delpy, "Near-infrared light propagation in an adult head model. I. Modeling of low-level scattering in the cerebrospinal fluid layer," Appl. Opt. **42,** 2906–2914 (2003).
39. K. Uludag, M. Kohl, J. Steinbrink, H. Obrig, and A. Villringer, "Cross talk in the Lambert-Beer calculation for near-infrared wavelengths estimated by Monte Carlo simulations," J. of Biomed. Opt. **7,** 51–59 (2002).
40. J. Selb, J. J. Stott, M. A. Franceschini, A. G. Sorenson, and

D. A. Boas, "Improved sensitivity to cerebral dynamics during brain activation with a time-gated optical system: analytical model and experimental validation," J. Biomed. Opt. **10,** 011013 (2005).
41. B. Montcel, R. Chabrier, and P. Poulet, "Detection of cortical activation with time-resolved diffuse optical methods," Appl. Opt. **44,** 1942–1947 (2005).
42. E. Okada and D. T. Delpy, "Effect of a nonscattering layer on time-resolved photon migration paths, in *Photon Propagation in Tissues IV*, D. A. Benaron, B. Chance, M. Ferrari, and M. Kohl-Bareis, eds., Proc. SPIE **3566,** 2–9 (1998).
43. D. A. Boas, A. M. Dale, and M. A. Franceschini, "Diffuse optical imaging of brain activation: approaches to optimizing image sensitivity, resolution, and accuracy," Neuroimage **23,** S275–288 (2004).
44. L. Wang, S. L. Jacques, and L. Zheng, "MCML-Monte Carlo modeling of light transport in multi-layered tissues," Comput. Methods Programs Biomed. **47,** 131–146 (1995).
45. D. A. Boas, J. P. Culver, J. J. Stott, and A. K. Dunn, "Three dimensional Monte Carlo code for photon migration through complex heterogeneous media including the adult human head," Opt. Express **10,** 159–170 (2002).
46. D. A. Boas, G. Strangman, J. P. Culver, R. D. Hoge, G. Jasdzewski, R. A. Poldrack, B. R. Rosen, and J. B. Mandeville, "Can the cerebral metabolic rate of oxygen be estimated with near-infrared spectroscopy?" Phys. Med. Biol. **48,** 2405–2418 (2003).
47. J. Steinbrink, H. Wabnitz, H. Obring, A. Villringer, and H. Rinneberg, "Determining changes in NIR absorption using a layered model of the human head," Phys. Med. Biol. **46,** 879–896 (2001).
48. G. Strangman, M. A. Franceschini, and D. A. Boas, "Factors affecting the accuracy of near-infrared spectroscopy concentration calculations for focal changes in oxygenation parameters," Neuroimage **18,** 865–879 (2003).
49. E. Okada, M. Schweiger, S. R. Arridge, M. Firbank, and D. T. Delpy, "Experimental validation of Monte Carlo and finite-element methods for the estimation of the optical path length in inhomogeneous tissue," Appl. Opt. **35,** 3362–3371 (1996).
50. H. Dehghani and D. T. Delpy, "Near-infrared spectroscopy of the adult head: Effect of scattering and absorbing obstructions in the cerebrospinal fluid layer on light distribution in the tissue," Appl. Opt. **39,** 4721–4729 (2000).
51. S. Takahshi and Y. Yamada, "Simulation of 3D light propagation in a layered head model including a clear CSF layer," OSA Trends Opt. Photonics Ser. **21,** 2–6 (1998).

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Computational Physics 201 (2004) 771–797
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

A fast numerical method for time-resolved photon diffusion in general stratified turbid media

Alex H. Barnett *

Courant Institute of Mathematical Sciences, New York University, 251 Mercer St, New York, NY 10012, USA

Received 24 October 2003; received in revised form 17 June 2004; accepted 30 June 2004
Available online 13 August 2004

Abstract

Efficient forward models of photon migration in complex geometries are important for noninvasive imaging of tissue in vivo with diffuse optical tomography (DOT). In particular, solving the inverse problem requires multiple solutions of the forward model and is therefore computationally intensive. We present a numerical algorithm for the rapid solution of the time-dependent diffusion equation in a semi-infinite inhomogeneous medium whose scattering and absorption coefficients are arbitrary functions of depth, given a point source impulsive excitation. Such stratified media are biomedically important. A transverse modal representation leads to a series of one-dimensional diffusion problems which are solved via finite-difference methods. A novel time-stepping scheme allows effort to scale independently of total time (for fixed system size). Tayloring to the DOT application gives run times of order 0.1 s. We study convergence, computational effort, and validate against known solutions in the case of 2-layer media. The method will be useful for other forward and inverse diffusion problems, such as heat conduction and conductivity measurement.
© 2004 Elsevier Inc. All rights reserved.

MSC: 65M06; 65M12; 78A70; 80M20; 92C55
Keywords: Parabolic; Diffusion; Photon migration; Diffuse optical tomography; Time-resolved; Inverse problem; Stratified

1. Introduction

Diffuse optical tomography (DOT) [1,9] is receiving growing attention as a noninvasive method for functional imaging of living tissue. Near-infrared light propagates diffusively inside tissue for long distances, allowing surface measurements to be sensitive to absorption at depths of order centimeters. In this wavelength range the small absorption is dominated by the spectra of oxy- and deoxy-hemoglobin. The large

* Corresponding author. Tel.: +1-212-998-3296.
E-mail address: barnett@cims.nyu.edu.

0021-9991/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2004.06.017

difference between the two spectra allows reconstruction of independent spatial maps of oxy- and deoxy-hemoglobin, and therefore of total blood volume and oxygenation, from absorption images at multiple wavelengths. Metabolic activity and hemodynamic response give a window on tissue function, providing information not accessible to other imaging modalities. This, combined with the ability to map the scattering coefficient and a sensitivity to optical contrast agents such as fluorescent dyes [2], results in a wealth of applications. On the clinical side they include breast tumor screening [4], neonatal cerebral monitoring [3] and arthritis assessment [6]. On the functional side they include neuroimaging [2,1] and muscle oximetry [5].

Reconstructing absorption and scattering coefficients from surface flux measurements is an inverse problem [9]. When optical inhomogeneity is large, or when baseline measurements are desired, model-based nonlinear methods [10,13,20,15,36,14] are optimal. They require many solutions to the forward problem, namely the simulation of flux measurements at certain distances from the sources given an optical model of the underlying tissue with certain parameter values. These parameter values are then adjusted iteratively [20,15,36] or sampled in a Bayesian statistical fashion [14] to solve the inverse problem. This can require hundreds or more forward model solutions. Surface light flux measurements may be continuous-wave (unmodulated), radio-frequency modulated, or time-resolved. Of these three measurement types, the latter is considered to provide the most information [1]. The resulting reflectance signal measured at various distances from each impulsive point source is called a *temporal point spread function* (TPSF). Therefore, to simulate such signals an efficient time-dependent forward solution is needed.

In this paper, we provide such a solution for the diffusion approximation (DA) to photon transport in a layered medium. Layered media are ubiquitous in the biomedical field – for instance they approximate the local geometry of scalp, skull, and brain, or skin, fat, and muscle – and recently a growing need for fast layered solvers has been identified [32–34]. The fact that DOT makes use of point-like sources and detectors at various separations means that the problem has cylindrical symmetry rather than being simply one-dimensional (1D). The symmetry allows a much more rapid solution than is possible in general tissue geometries where 3D finite element [9] or finite difference [20,15,14] methods are required. The DA, whose validity is reviewed in Section 2, is orders of magnitude faster to simulate than the more physically accurate radiative transfer equation (RTE). We note that gains due to the symmetry also extend to the RTE, for instance using layered Monte Carlo [18,19], however its simulation remains an intensive task.

Our framework takes advantage of the symmetry of the medium (Fig. 1): using separation of variables into planar coordinate $\boldsymbol{\rho} \equiv (x, y)$ and depth coordinate z, general fluence solutions to the DA can be built up from solutions with a definite in-plane 2D wavevector $\mathbf{k}$. Furthermore, the cylindrical symmetry of the

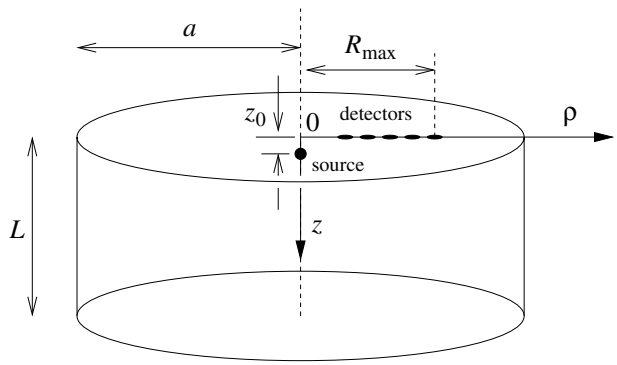


Fig. 1. Coordinate system, and geometry of pill-box used to model the semi-infinite medium. The source fiber (not shown) lies along the negative z-axis; the effective DA point source is shown as a large dot. Detector fibers (not shown) lie parallel to the negative z-axis at the various radii R shown. As explained in Section 4, the pill-box is chosen to be large enough (~100 mm) that the resulting signals differ negligibly from those due to a semi-infinite medium (infinite depth and radius).

source term means that only the 1-parameter family of radially-symmetric solutions is needed (namely the $J_0(k\rho)$ Bessel function parametrized by wavenumber $k = |\mathbf{k}|$, where the radius is $\rho = |\boldsymbol{\rho}|$). At each k there remains to be solved a diffusive 1D partial differential equation (PDE) in z and time. If the system is of infinite extent in the radial coordinate ρ then k is a continuous parameter; integrating the contributions from different k then corresponds to a Hankel transform from k to ρ. This idea is well-known in the literature on multilayer (piecewise homogeneous) stratified media. For instance in the 2-layer medium it is used by Dayan et al. [30] to derive approximate solutions, and, in the case of steady-state solutions, by Kienle et al. [32] who evaluate the Hankel transform via numerical quadrature (the 1D steady-state PDE having an analytic solution).

Our approach, which as far as we are aware is new, involved first truncating the system's radius to a finite value large enough to induce negligible error (because of the nature of the diffusion equation this error can be made exponentially small). Therefore, the integral over k becomes a sum over discrete (quantized) wavenumbers k_m. The choice of method for the 1D evolution of each mode m is very flexible: we choose finite-difference methods. Computational efficiency is achieved (a) by carefully considering the errors and requirements of the time-resolved DOT application, (b) by introducing a novel timestepping scheme appropriate for impulsively-excited parabolic problems (this contrasts the uniform timestepping currently used both in the DOT field [13,20,24,14] and more generally), and (c) by minimizing the length of time for which each mode m is evolved. The last of these relies on the fact that errors due to truncation in the Fourier domain are also exponentially small (see Fig. 3(b)). The desired TPSF signals are then extracted in the usual way from the resulting solution at the surface.

We now categorize some of the different choices made by other authors. In our method sketched above the fluence function $\phi(\rho, z, t)$ is written as a sum of modes $\psi_m(\rho)u_m(z, t)$, which we could call a *transverse mode approach*. The other possibility is a sum of modes $\xi_l(\rho, t)v_l(z)$, or *longitudinal mode approach*, where the roles of ρ and z have been reversed. The latter has been used for steady-state solutions in 2- and 3-layer media [28,29], and time-resolved solutions [34] based upon them. The longitudinal eigenvalues are found by numerically solving a transcendental equation (root-finding), and the $\xi_l(\rho, t)$ obey the analytically-solvable 2D homogeneous diffusion equation. However, a finite total depth is needed to ensure that all eigenvalues are discrete, and computational effort increases with eigenvalue density which is in turn proportional to total depth. Therefore, the approach is ill-adapted to infinitely-thick media. Quantization in both $\boldsymbol{\rho}$ and z is used in the eigenfunction method of Martelli et al. [35], which handles finite media but requires triple summation.

Other approaches include the remarkable analytic expansion of Tualle et al. [33], which relies on a volume source representation of the solution due to a planar medium interface to generate very fast 2-layer solutions. We also note that integral equation techniques lend themselves naturally to the solution of PDEs in piecewise-homogeneous media – in particular since the problem of history-dependence of time-dependent diffusion has been elegantly overcome [47,48] – however we know of no biomedical applications of these promising techniques.

In contrast to all the above layered approaches which are based on boundary matching, our choice of 1D evolution method allows us to model media with arbitrary depth-dependent absorption and scattering coefficients, hence we can tackle an arbitrary number of layers, and continuously-varying coefficients, with no extra complexity. This may be important for tissues such as the skull which appears to have scattering properties which vary continuously with depth.

The remainder of this paper is structured as follows. In Section 2, we present the well-known DA model and discuss its validity. In Section 3, we bring together the requirements placed on our method by the DOT application; this will prove important for maximizing numerical efficiency. Emulation of the infinite system by one of finite size is presented, along with numerical convergence tests in Section 4. The core of the method, which solves the DA in this finite-sized system, is explained in Section 5. Section 5.5 provides an algorithm summary. Numerical convergence of the finite-difference method, and validation against an existing

method, is presented in Section 6. The scaling of computational effort is analyzed in detail in Section 7. We discuss some issues, limitations, and future directions in Section 8, and conclude in Section 9.

2. The diffusion approximation model

Incoherent mono-energetic photon passage through tissue is described by a Boltzmann-type RTE giving the time-evolution of the radiance $\Phi(\mathbf{r}, \boldsymbol{\Omega}, t)$ (photon density in the 5-dimensional phase-space comprising position $\mathbf{r} \equiv (x, y, z)$ and solid angle $\boldsymbol{\Omega}$). By retaining only the first (dipole) term in the angular dependence, and assuming this dependence is small [11,9], we get the DA,

$$\frac{1}{c}\frac{\partial \phi}{\partial t} = \nabla \cdot (\kappa \nabla \phi) - \mu_a \phi + q, \tag{1}$$

where $\phi(\mathbf{r}, t)$ is the fluence (the angular integral of radiance), $q(\mathbf{r}, t)$ is the source term, $\kappa(z)$ is the local diffusion constant (in units of length), $\mu_a(z)$ is the local absorption (in units of inverse length), and c is the speed of light in the medium, defined as the speed in vacuo divided by n the refractive index. κ is related to reduced scattering coefficient μ_s' by the good approximation [16,9] $\kappa = 1/(3\mu_s')$. It is the reduced dimensionality (fluence is a function of 3 spatial dimensions as opposed to 5 for radiance) that makes the DA easier to solve than the RTE. We remind the reader of the coordinate system used for what follows: $z = 0$ defines the air-tissue interface with z increasing with depth inside the medium, and $\boldsymbol{\rho} \equiv (x, y)$ is the in-plane coordinate whose length is the (cylindrical) radial distance ρ (see Fig. 1). We reserve the symbol R to refer to distance of a detector from the source. Note that in this work we have spatial dependence of κ and μ_a but not of n (allowing this would require only simple changes to the finite difference scheme).

Given a source–detector separation R, the DA is a good approximation to the RTE when several conditions are satisfied [39,9,16]: that $\mu_a \ll \mu_s'$ everywhere, that $1/\mu_s'$ is much smaller than both R and any significant geometrical features, that times much greater than R/c are considered, and that appropriate boundary conditions and optode models, discussed below, are used. The size of the error involved in using the DA is a topic of ongoing investigation: it has found to be acceptably small for time-resolved reflectance of layered media [32], although biomedical situations where the steady-state fluence is inaccurate may be found [16]. In particular, the DA breaks down in the clear cerebro-spinal fluid (CSF) layer between the skull and brain, however there exist ways [21,40] to handle this within the DA, and there is recent evidence that the effective μ_s' of CSF is only a few times less than that of surrounding tissue [22].

The best choice of boundary condition for an air-tissue interface appropriate for the time-domain setting is a topic of recent research [38,12,25,31,37]. The extrapolated boundary conditions

$$\phi(\mathbf{r}, t)|_{z=-z_b} = 0, \quad \forall \boldsymbol{\rho}, t, \tag{2}$$

have proved to be reasonable [37]. With $n \approx 1.4$ for tissue [41] the extrapolation distance, taking into account Fresnel reflection [25], is $z_b \approx 1.96 z_0$. Here the scattering depth is $z_0 \equiv 1/\mu_s'(0) = 3\kappa_0$, written in terms of the diffusion constant in the neighborhood of the surface $\kappa_0 \equiv \lim_{z\to 0^+} \kappa(z)$.

The impulsive source q is only nonzero at $t = 0$ (in this work we ignore the finite width of the impulse response of optical fibers and instrumentation since this can be modeled easily by convolution in time). Therefore, we can treat Eq. (1) as an initial-boundary value problem with no source term ($q = 0$). We take the initial conditions appropriate for a narrow normally-incident optical fiber source in contact with the skin to be

$$\phi(\mathbf{r}, 0) = \delta(x)\delta(y)\delta(z - z_0), \tag{3}$$

the commonly-used isotropic DA point source model [25] (see Fig. 1). We note that, alternatively, a gradient source at $z = 0$ (that is, an inhomogeneous Neumann boundary condition) would also be easy to implement within our scheme, requiring only an appropriate change to Eq. (19).

We define $I(R, t)$ to be the detected signal (TPSF) at a source–detector distance R and time t. The best way to extract this from the fluence solution in the time-domain has also recently been discussed [25,31,37]. The general linear combination,

$$I(R,t) = \kappa_0 \left[A \frac{\phi(R,t)}{z_b} + (1-A) \frac{\partial \phi}{\partial z}(R,t) \right] \Bigg|_{z=0}, \tag{4}$$

with $A \in [0, 1]$, covers all possibilities and, by varying A, can account for arbitrary detector fiber numerical aperture. The weighting is chosen so that in the long-time limit the signal is independent of A. Unless stated we will choose $A = 1$, since this is known to perform well (in [37] this choice is named 'R3'). Other choices of boundary condition and measurement type are easy to implement within our method.

3. Requirements of our numerical method

In order to maximize numerical efficiency, it is important to list the requirements which arise in the biomedical DOT application, which we do in this section. Our method is currently designed and used for fitting signals collected by time-correlated single photon counting (TCSPC) apparatus [7,8,1], a popular choice for time-resolved DOT. The goal is fast computation of $I(R, t)$ for each source–detector separation R, to an accuracy sufficient for the fitting process. The first four requirements listed, involving the system parameters, arise generally in time-resolved biomedical DOT. We will show later that requirements (ii) and (iv) have quite a strong influence on the total effort needed for computation; see Eq. (50). The last requirement, namely the choice of acceptable error level, also has a strong effect on total effort, thus we have matched this error level in more detail to our TCSPC fitting application. All five requirements are expected to have analogues in other diffusion applications, such as heat conduction.

(i) *Source-detector separation range*: At each R, we define the time-resolved peak signal $I_{\text{peak}}(R) \equiv \max_t I(R,t)$. The decay of $I_{\text{peak}}(R)$ with R is rapid: for example in a homogeneous medium with zero absorption, $I_{\text{peak}}(R) \sim R^{-5}$. Thermal safety considerations, fiber area and detector sensitivity limit useful source–detector distances of interest to $R_{\max}$, which is of order 40 mm. It transpires that our numerical scheme is insensitive to any minimum R, so we set $R \in [0, R_{\max}]$.
(ii) *Time range of interest*: At early times the DA does a poor job of approximating the RTE [39]. Therefore, there is a time $t_{\min} \sim 100$ ps [17,23,31] before which it is not appropriate [43] to compute the signal. After a time $t_{\max}$ of order a few thousand ps, the signal drops below detectable levels. Therefore, $I(R, t)$ is needed only for $t \in [t_{\min}, t_{\max}]$.
(iii) *Maximum depth of medium*: We are interested in both finite-thickness and semi-infinite media. In this work, we tackle the semi-infinite case, which is more challenging because it requires a decision about how to truncate the medium's infinite spatial extent while keeping errors acceptable (see Section 4).
(iv) *Optical parameter ranges*: Given a medium we define the inhomogeneity ratio $\alpha \equiv \kappa_{\max}/\kappa_{\min} = \mu'_{s,\max}/\mu'_{s,\min}$, where $\kappa_{\max}$ and $\kappa_{\min}$ are the maximum and minimum of $\kappa(z)$. For biomedical applications α rarely exceeds 10, and is usually much less. However, α is not small enough that perturbative methods of solution are possible. For instance, the range of μ'_s reported in in vivo human head tissues, excluding CSF, is 0.6–2.5 mm^{-1} [14,16,42,44]. There is recent evidence [22] that even the clear CSF has an effective μ'_s of 0.16–0.32 mm^{-1}, in which case including it in the DA model causes α hardly to exceed the above limit. Because $\mu_a \ll \mu'_s$ within the DA it will transpire that there are effectively no further constraints on absorption values.
(v) *Acceptable error as a function of time*: The basic principle we apply is that forward model errors (numerical and physical) should not exceed a small fraction of the detection apparatus measurement errors [45]. Given a signal $I(t)$ we define an acceptable absolute error level $I_{\text{err}}(t)$. If absolute numerical errors in

calculating $I(t)$ are below this baseline for all $t \in [t_{\min}, t_{\max}]$, then we define this to be acceptable accuracy. If relative error alone defined acceptability, this would correspond to $I_{\mathrm{err}}(t) = \epsilon I(t)$, where ϵ sets the relative error. However, when signals are very small, such as at early times and large R, in practice relative measurement errors grow (or the signal is below detection levels), and this criterion becomes too stringent. We therefore propose a modified criterion which includes the Poisson shot-noise statistics of single photon counting [7], namely

$$I_{\mathrm{err}}(t;\gamma,\epsilon) \equiv \max\left[\frac{I_{\mathrm{peak}}}{\gamma}, \sqrt{\frac{I_{\mathrm{peak}} I(t)}{\gamma}}, \epsilon I(t)\right], \tag{5}$$

where the max function returns the maximum of its three arguments. Note that for $I(t)$ sufficiently large this expression equals $\epsilon I(t)$. The first and second arguments together approximate the form of the standard deviation of the Poisson distribution ($\sigma \sim \sqrt{n}$ for n detected photons). I_{peak} is relevant because at a given R, in practice the apparatus gain is adjusted to prevent saturation at I_{peak}. The dynamic range of interest is set by γ. In our context, with $\sim 10^5$ photons detected per source–detector pair [14], and a dark count of $\sim 10^{-3} I_{\mathrm{peak}}$, we believe $\gamma \sim 10^4$ or 10^5 is appropriate. The third argument is the original relative error criterion; typically we need $\epsilon \sim 10^{-2}$. Clearly larger γ or smaller ϵ demand smaller numerical errors; this will result in a slower forward solution. In Fig. 2, we show a typical signal $I(t)$ and the resulting acceptable error level $I_{\mathrm{err}}(t)$.

4. Truncation of the spatial domain

We want to solve the initial-boundary value problem given by the sourceless version of Eq. (1), namely

$$\frac{1}{c}\frac{\partial \phi}{\partial t} = \nabla \cdot (\kappa \nabla \phi) - \mu_{\mathrm{a}} \phi, \tag{6}$$

with boundary condition equation (2) and initial condition equation (3), in the semi-infinite slab $\rho \in [0, \infty]$, $z \in [-z_{\mathrm{b}}, \infty]$. Note that the slight extension of the domain into negative z results from the extrapolated boundary condition. Here we give a recipe for truncating this domain while causing negligible changes to the desired signals $I(R, t)$. Although the problems of depth truncation and radial truncation can be treated independently due to the system's separability, they are formally similar so we treat them both here.

We truncate both spatial coordinates to create a finite cylindrical box $\rho \in [0, a]$, $z \in [-z_{\mathrm{b}}, L]$ (see Fig. 1). We impose numerically-convenient (we choose Dirichlet) boundary conditions on the newly-created boundaries:

$$\phi(a, z, t) = 0 \quad \forall z, t, \tag{7}$$

$$\phi(\rho, L, t) = 0 \quad \forall \rho, t. \tag{8}$$

The key idea is that because the form of the time-dependent Green's function has Gaussian tails, the errors introduced by the box die exponentially with increasing box size, at fixed time. The spreading increases with time, therefore we should use $t = t_{\max}$ when bounding the errors. A one-dimensional schematic is given in Fig. 3(a).

We now formalize this more precisely in the similar case of a homogeneous medium of scattering coefficient κ and absorption coefficient μ_{a}, inside the *cuboid* box $x \in [-a, a]$, $y \in [-a, a]$, $z \in [-z_{\mathrm{b}}, L]$. Because of exponential convergence, the general result will carry over to the cylindrical box case; we will also verify this numerically. The exact solution in semi-infinite space is the sum of a heat kernel and its image charge,

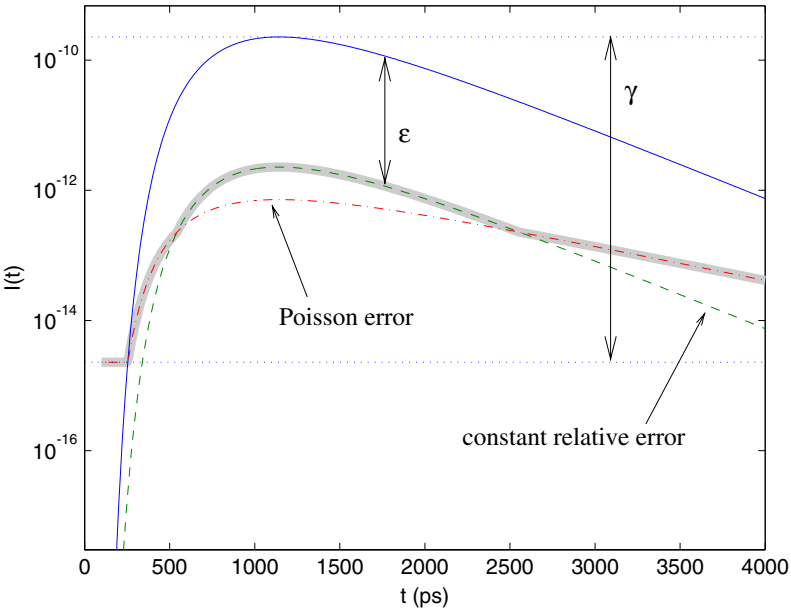


Fig. 2. Typical signal $I(t)$ (thin solid line), and its corresponding acceptable error level $I_{\mathrm{err}}(t)$ (fat grey solid line); see Section 3. $I_{\mathrm{err}}(t)$ is the larger of the Poisson error level (dash-dotted) and a constant relative error level (dashed). Here the relative error level dominates from roughly 500–2500 ps. The horizontal dotted lines show I_{peak} and I_{peak}/γ. The error parameters are a dynamic range of $\gamma = 10^5$ and relative error of $\epsilon = 10^{-2}$. This signal resulted from a homogeneous medium with $\mu_{\mathrm{s}}' = 1\ \mathrm{mm}^{-1}$, $\mu_{\mathrm{a}} = 0.01\ \mathrm{mm}^{-1}$, $R = 40$ mm.

$$\phi(\rho,z,t) = \frac{\mathrm{e}^{-\frac{\rho^2}{4\kappa ct}-\mu_{\mathrm{a}}ct}}{(4\pi\kappa ct)^{3/2}}\left[\mathrm{e}^{-\frac{(z-z_0)^2}{4\kappa ct}} - \mathrm{e}^{-\frac{(z+2z_{\mathrm{b}}+z_0)^2}{4\kappa ct}}\right]. \tag{9}$$

For large t we can Taylor expand the term in square brackets. The first nonzero term gives the dipole approximation

$$\phi_{\mathrm{d}}(\rho,z,t;z_{\mathrm{d}}) \approx d\,\frac{z-z_{\mathrm{d}}}{t^{5/2}}\mathrm{e}^{-\frac{(z-z_{\mathrm{d}})^2+\rho^2}{4\kappa ct}-\mu_{\mathrm{a}}ct}, \tag{10}$$

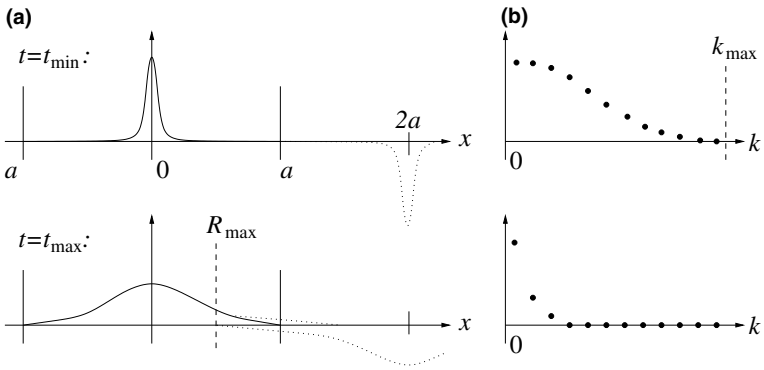


Fig. 3. One-dimensional example of real-space (a) and Fourier space (b) representation of the diffusion equation Green's function, or heat kernel, in a box. As the Gaussian width grows in real-space, it shrinks in Fourier space.

where the dipole strength is $d = (z_b + z_0)/(4\pi)^{3/2}(\kappa c)^{5/2}$ and its vertical location $z_d = -z_b$. We use Eq. (4) with $A = 0$ (the A-dependence is negligible) to give the signal detected at distance R_d in the xy-plane from a dipole vertically located at z_d,

$$I(R_d, t; z_d) = \kappa \left.\frac{\partial \phi_d(R_d, z, t; z_d)}{\partial z}\right|_{z=0}. \tag{11}$$

The cuboid box case can be solved using the method of images [46], using an infinite periodic rectangular 3D lattice of image dipole sources. We are interested in the case where the relative errors are small; this corresponds to a heat kernel width $\sqrt{2\kappa ct}$ which is smaller than a or L. Therefore, the error size will be well approximated by considering only the nearest set of image dipoles (the remaining infinite sum contributing exponentially less, as can be proved by bounding the sum by an exact Gaussian integral [47]).

In the xy-plane the four closest image dipoles have distance $2a$ from the origin, and therefore distance $R_d \geqslant 2a - R$ from the detector. Since R, $R_d \gg z_b$, we can use $z_d \approx 0$. The relative signal error due to these image dipoles is bounded by

$$\epsilon_a \leqslant 4\frac{I(2a - R, t; 0)}{I(R, t; 0)} = 4e^{-\frac{a(a-R)}{\kappa ct}} \leqslant 4e^{-\frac{(a-R)^2}{\kappa ct}}. \tag{12}$$

Within our domain of interest this is largest at $R = R_{max}$ and $t = t_{max}$. Given a desired ϵ_a our choice of a is therefore

$$a \approx R_{max} + \sqrt{\kappa c t_{max} \ln(4/\epsilon_a)}. \tag{13}$$

We have found that the difference between the convergence of ϵ_a with a in the cuboid and cylindrical cases is slight. Therefore, Eq. (13) is also a reasonable choice of radius for the cylindrical box.

Similarly, along the z-axis the two closest image dipoles are at $z_d \approx \pm 2L$ and $R_d = R$, where we again have used $z_b \approx 0$. The relative signal error due to these two dipoles is

$$\epsilon_L \approx 2\frac{I(R, t; 2L)}{I(R, t; 0)} = 2\left(1 + \frac{2L^2}{\kappa ct}\right)e^{-\frac{L^2}{\kappa ct}}. \tag{14}$$

Again using $R = R_{max}$ and $t = t_{max}$, an approximate solution for L is

$$L \approx \sqrt{\kappa c t_{max} \ln[C(\epsilon_L)/\epsilon_L]}, \tag{15}$$

where $C(\epsilon_L) = 2[1 + 2\ln(1/\epsilon_L)]$.

Returning to our inhomogenous system, where no closed form for $\phi(\rho, z, t)$ is known, we need a recipe to choose the smallest safe values of a and L. We use the fact that the spatial spread is bounded (in the xy-plane) by a Gaussian of width $\sqrt{2\kappa_{max} ct}$. This leads us to use Eq. (13) with the crude substitution $\kappa = \kappa_{max}$. A similar consideration leads to Eq. (15) with the same substitution. In other words, we are using a homogeneous system of diffusion coefficient κ_{max} to bound the box size needed for an inhomogeneous system. When the inhomogeneity ratio α is large, these choices may be less than optimal, but because the Gaussian width scales like $\kappa^{1/2}$, the factor by which both a and L could be suboptimal is only $\alpha^{1/2}$ or at most about 3.

No rigorous analytic proof of the validity of the above box size bounds for the inhomogeneous case are known to the author, however Fig. 4 shows verification that these estimates are valid for various inhomogeneous two-layered systems with large inhomogeneity $\alpha = 10$. In practise the exponential convergence ensures that we can set ϵ_a and ϵ_L to values much smaller than the desired relative error ϵ with very little penalty: for the rest of this work we use $\epsilon_a = \epsilon_L = 10^{-6}$. In our application, the typical resulting box size is $a \sim L \sim 10^2$ mm.

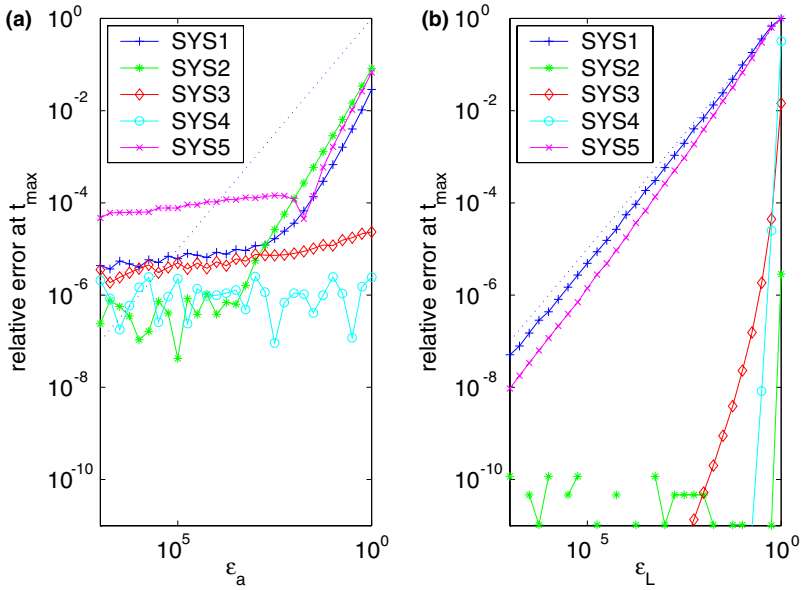


Fig. 4. Test of convergence of relative error at $R = 40$ mm (a typical $R_{\max}$), and $t = 5$ ns (a typical $t_{\max}$) due to box truncation. The five systems are: homogeneous (SYS1) and four inhomogeneous 2-layer systems with maximal range of κ (SYS2–5), as described in Table 1. (a) shows relative error using a given by Eq. (13) while $\epsilon_L = 10^{-15}$, and (b) shows the same using L given by Eq. (15) while $\epsilon_a = 10^{-15}$. The dotted lines show expected behavior if these estimates were correct. The systems are compared against a calculation with $\epsilon_a = \epsilon_L = 10^{-15}$. In (a) the bottoming out at a relative error of about 10^{-4} reflects the overall error of the method; because a changes it cannot be removed. In (b) λ_{shift} was fixed to allow comparison beyond this overall error level. Finite difference parameters were $h = 0.5$ mm, $\beta = 0.05$ (see Sections 5.2 and 5.3).

5. The method

5.1. Transverse mode representation

All solutions of Eq. (6) with the boundary condition equations (2), (7) and (8) can be written

$$\phi(\rho, z, t) = \sum_{m=1}^{M} B_m \psi_m(\rho) u_m(z, t), \tag{16}$$

which is exact in the limit $M \to \infty$, where the radial functions are normalized eigenmodes of the Laplacian in the disc of radius a in the xy-plane,

$$\psi_m(\rho) = \frac{1}{\sqrt{\pi} a J_1(\chi_m)} J_0(k_m \rho). \tag{17}$$

The transverse wavenumbers are given by $k_m = \chi_m/a$, where χ_m is the argument of the mth zero of the J_0 Bessel function. The coefficents corresponding to the correct initial condition equation (3) are $B_m = [\sqrt{\pi} a J_1(\chi_m)]^{-1}$, which can easily be shown by using $\delta(x)\delta(y) = \sum_{m=1}^{\infty} \psi_m(x, y)\psi_m(0, 0)$, a special case of the eigenfunction closure relation. We remind the reader that it is the symmetry of the initial condition that causes only radially-symmetric modes to contribute.

For each m the evolution of $u_m(z, t)$ obeys

$$\frac{1}{c}\frac{\partial u_m}{\partial t} = \frac{\partial}{\partial z}\left(\kappa \frac{\partial u_m}{\partial z}\right) - \mu_{a,m} u_m, \tag{18}$$

$$u_m(z,0) = \delta(z - z_0), \tag{19}$$

$$u_m(-z_b, t) = u_m(L, t) = 0, \quad \forall t, \tag{20}$$

with the effective m-dependent absorption coefficient

$$\mu_{a,m}(z) = \mu_a(z) + k_m^2 \kappa(z). \tag{21}$$

The signal is extracted using Eq. (4) applied to Eq. (16),

$$I(R,t) = \kappa_0 \sum_{m=1}^{M} B_m \psi_m(R) \left[A \frac{u_m(0,t)}{z_b} + (1-A) \frac{\partial u_m}{\partial z}(0,t) \right]. \tag{22}$$

We discuss choice of the number of modes M in Section 5.4.

5.2. Spatial finite-difference discretization

Currently the 1D problem $u_m(z, t)$ is discretized spatially using a finite-difference lattice of cell size h. A uniform lattice is used out of convenience; undoubtedly improved efficiency will result with a nonuniform lattice. Values of fluence $u_m^{(i)}$ are based at nodes (i.e. junctions between cells), labelled by $i = 0, 1, \ldots, n_z$. The extrapolated boundary conditions require that the first cell begins at $z = -z_b$, with the $i = 0$ node fluence value held at zero. A minimum number of cells n_z is chosen such that the last cell ends beyond the box depth $z = L$ (recall that L has been chosen in Section 4 such that the results are not sensitive to L). The $i = n_z$ node fluence value is held at zero to model the box boundary condition. In Eq. (18) the right-hand side is spatially discretized using a 3-point template to give the continuous time-evolution (often called the 'Method of Lines' [26]),

$$\frac{1}{c}\frac{\partial u_m^{(i)}}{\partial t} = \frac{\kappa^{(i-\frac{1}{2})}}{h^2} u_m^{(i-1)} - \left(\frac{\kappa^{(i-\frac{1}{2})} + \kappa^{(i+\frac{1}{2})}}{h^2} + \frac{\mu_{a,m}^{(i-\frac{1}{2})} + \mu_{a,m}^{(i+\frac{1}{2})}}{2} \right) u_m^{(i)} + \frac{\kappa^{(i+\frac{1}{2})}}{h^2} u_m^{(i+1)}, \quad i = 1, 2, \ldots, n_z - 1. \tag{23}$$

Here $\kappa^{(i-\frac{1}{2})}(\kappa^{(i+\frac{1}{2})})$ signifies a diffusion coefficient, and $\mu_{a,m}^{(i-\frac{1}{2})}(\mu_{a,m}^{(i+\frac{1}{2})})$ an effective absorption coefficient, within the cell to the left (right) of node i.

The cell-averaging is not entirely obvious. By considering the steady-state flux conservation law and Fick's law across a cell split into two homogeneous pieces with different κ, it is easy to show that μ_s', or the inverse of κ, is the correct quantity to average. On similar physical grounds μ_a, and therefore for each mode the effective $\mu_{a,m}$, is the correct quantity to average. Therefore, making use of Eq. (21),

Table 1
Homogeneous and 2-layer systems used for numerical tests

Name	Upper layer			Lower layer		
	μ_s'	μ_a	Thickness	μ_s'	μ_a	Thickness
SYS1	1.0	0.01	∞			
SYS2	0.3	0	10	3	0.03	∞
SYS3	3	0	10	0.3	0.03	∞
SYS4	0.3	0.03	10	3	0	∞
SYS5	3	0.03	10	0.3	0	∞
SYS6	0.5	0.01	10	2.0	0.003	∞

Absorption μ_a and reduced scattering coefficient μ_s' are in units of mm^{-1}, while thickness is in units of mm. The first four 2-layer systems cover the four possibilities of large and small absorption and diffusion for each layer, with a large choice of diffusion inhomogeneity $\alpha = 10$. All refractive indices are $n = 1.4$.

$$(\kappa^{(i+\frac{1}{2})})^{-1} \equiv \frac{1}{h}\int_{z_i}^{z_i+h} \kappa(z)^{-1}\,\mathrm{d}z, \tag{24}$$

$$\mu_{\mathrm{a},m}^{(i+\frac{1}{2})} \equiv \frac{1}{h}\int_{z_i}^{z_i+h} [\mu_{\mathrm{a}}(z) + k_m^2\kappa(z)]\mathrm{d}z, \tag{25}$$

where $z_i = \mathrm{i}h - z_{\mathrm{b}}$ is the depth of node i. Note the peculiarity that κ^{-1} is averaged for the diffusion term, but κ itself is averaged within the effective absorption term. The scheme is second-order accurate in h, as demonstrated in Section 6.1.

We can express the single mode evolution equation (23) using the column vector $\mathbf{u}_m \equiv \{u_m^{(i)} : i = 1, 2, \ldots, n_z - 1\}$ as the ordinary differential equation

$$\frac{\partial \mathbf{u}_m}{\partial t} = -S_m \mathbf{u}_m(t), \tag{26}$$

$$\mathbf{u}_m(0) = \mathbf{u}_0. \tag{27}$$

The initial condition is given by a standard linear spatial interpolation of $\delta(z - z_0)$ appearing in Eq. (19) onto the lattice. Similarly, the signal is extracted using linear spatial interpolations of $u_m(0, t)$ and $(\partial u_m/\partial z)(0, t)$ in Eq. (22) on the lattice. These linear interpolations preserve the overall $\mathrm{O}(h^2)$ accuracy.

Furthermore, we may express the whole discretized system as the N-component vector $\mathbf{U} \equiv [\mathbf{u}_1^{\mathrm{T}} \ldots \mathbf{u}_{\mathrm{M}}^{\mathrm{T}}]^{\mathrm{T}}$ which is simply the vectors $\mathbf{u}_m$ stacked vertically. Its evolution (that is, the Method of Lines for the whole system) is then

$$\frac{\partial \mathbf{U}}{\partial t} = -S\mathbf{U}(t), \tag{28}$$

$$\mathbf{U}(0) = \mathbf{U}_0. \tag{29}$$

The $N \times N$ matrix S is block diagonal, with tridiagonal blocks S_m. Signal extraction by Eq. (22) can now be summarized in the form of a projection onto a time-independent vector $\mathbf{P}(R)$,

$$I(R, t) = \mathbf{P}(R)^{\mathrm{T}}\mathbf{U}(t). \tag{30}$$

5.3. Fast time-stepping scheme

The exact solution to Eqs. (28) and (29) is

$$\mathbf{U}(t) = \mathrm{e}^{-St}\mathbf{U}_0 = V\mathrm{diag}(\mathrm{e}^{-\lambda_j t})V^{\mathrm{T}}\mathbf{U}_0, \tag{31}$$

where the matrix exponential has been written in terms of the spectral decomposition of S: the eigenvalues are λ_j and the eigenvector matrix is $V = (\mathbf{v}_1 \cdots \mathbf{v}_N)$, that is, eigenvectors $\mathbf{v}_j$ arranged in columns. Evaluating the evolution exactly in this manner would require computing the full decomposition. It is the job of any time-stepping scheme to approximate the important contributions to this evolution with much less numerical effort.

Traditionally in the DOT field, if a finite-difference method is used then the time-stepping has been uniform, using either forward Euler [14] or Crank–Nicolson (CN) [13,20,24,9] schemes. The reasoning that CN allows a longer time step Δt than forward Euler because it is guaranteed stable for arbitrarily large Δt is often repeated [9] but can be unwittingly abused. What is not commonly understood is that in the limit of stiff lattice modes (high decay rates λ) then CN only approaches *marginal* stability: in Fig. 5(a) the CN curve is asymptotic to -1. Technically CN is 'A-stable' but not 'L-stable' [26]. This manifests itself as

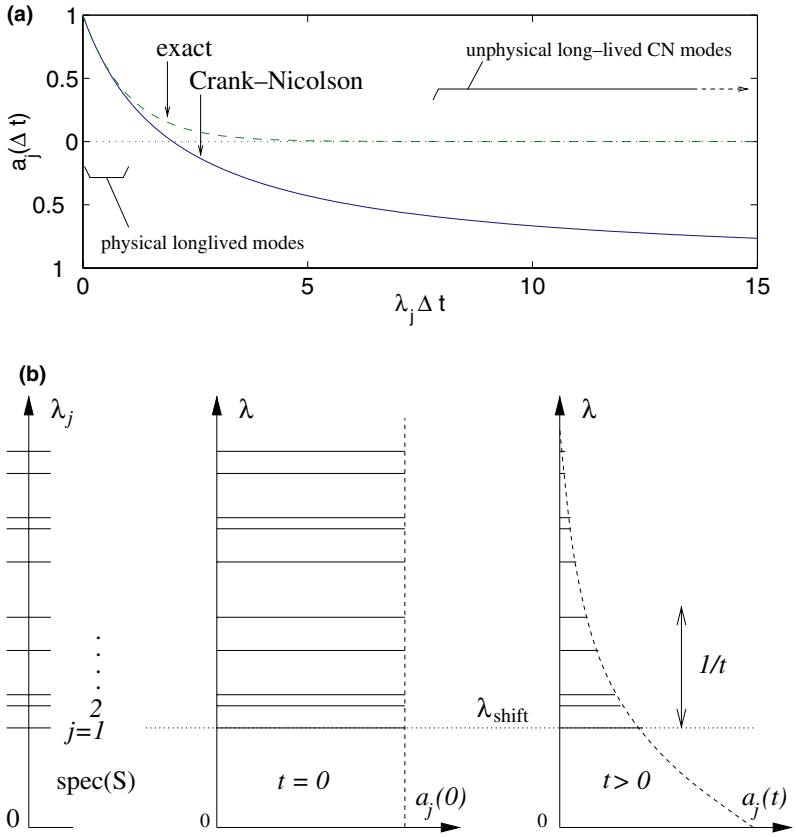


Fig. 5. (a) Decay of amplitude $a_j(\Delta t)$ of a mode j of the system due to a single Crank–Nicolson time-step Δt (solid line), compared to its exact decay $e^{-\lambda_j \Delta t}$ (dotted line). For convenience we took $a_j(0) = 1$. (b) Spectrum of system matrix S. Amplitude coefficients $a_j(0)$ at $t = 0$ are assumed to have similar sizes. At a later time t they have decayed as an exponential function of their eigenvalue λ_j.

unphysical extremely long decay times of the most oscillatory modes of the lattice, which can decay more slowly than the physical low-frequency modes, thereby causing an exponentially erroneous solution. (Contrary to one claim [51], these modes cannot easily be damped out without reducing accuracy or requiring a large increase in numerical effort). These oscillatory modes are strongly excited by a spatially-discontinuous initial condition such as Eq. (19). This fundamental problem has not been addressed in the DOT numerical literature. The scheme we present overcomes this problem while also vastly reducing computational effort.

Eq. (31) states that $\mathbf{U}(t)$ is the sum of eigenvectors with exponentially-decaying coefficients,

$$\mathbf{U}(t) = \sum_{j=1}^{N} a_j(t)\mathbf{v}_j. \tag{32}$$

The coefficients $a_j(t) = e^{-\lambda_j t} a_j(0)$ have initial values $a_j(0) = \mathbf{v}_j^{\mathrm{T}}\mathbf{U}_0$. Consider the spectrum of the system matrix S and the coefficient decay illustrated in Fig. 5(b). We assume (i) that the density of eigenvalues is approximately uniform along the λ axis (this is to be expected since it is the generic behavior for a wave eigenproblem in 2D [49]), and (ii) that all the $a_j(0)$ are of the same order of magnitude. At a fixed time $t > 0$ the decay $e^{-\lambda t}$ is exponential in λ. Therefore, at this time only the coefficients corresponding to eigenvalues

lying in the range from a minimum of λ_1 to a maximum of $\lambda_{\max}(t) \approx \lambda_1 + O(1/t)$ are needed; dropping coefficients corresponding to λ above this maximum would result in exponentially small error. Thus we see that as t increases, a shrinking piece of the spectrum of size $O(1/t)$ is relevant.

A single CN timestep of duration Δt performs approximate time-evolution via the formal replacement

$$e^{-S\Delta t} = \frac{1 - \frac{1}{2}S\Delta t}{1 + \frac{1}{2}S\Delta t} + O(\Delta t^3) + \cdots, \tag{33}$$

where $O(\Delta t^3)$ and higher terms are ignored. In practice the matrix division is handled by solution of the tridiagonal linear system of equations, which requires $O(N)$ computational effort [27]. The CN approximation in the eigenbasis can be written in terms the coefficients thus

$$a_j(\Delta t) \approx a_j^{\mathrm{CN}}(\Delta t) = \frac{1 - \frac{1}{2}\lambda_j\Delta t}{1 + \frac{1}{2}\lambda_j\Delta t} a_j(0). \tag{34}$$

This is plotted in Fig. 5(a): the approximation is good when $\lambda\Delta t \ll 1$, where the dominant error term is $-(\lambda\Delta t)^3/12$. For $\lambda\Delta t \gg 1$ the approximation is very poor. However, if we choose Δt such that $\lambda_{\max}(t)\Delta\varpi$ $t < O(1)$, then all the relevant coefficients will be evolved accurately. Coefficients corresponding to larger eigenvalues will have totally incorrect evolution, however we have already established above that this causes negligible error. This suggests a t-dependent (i.e. nonuniform) timestep $\Delta t(t) \sim [\lambda_1 + O(1/t)]^{-1}$.

The long-time limit of $\Delta t(t)$ presented above is a constant, $1/\lambda_1$. This situation can easily be improved so that there is no upper limit. We shift the spectrum of S by defining

$$S' \equiv S - \lambda_{\mathrm{shift}} I, \tag{35}$$

with shift $\lambda_{\mathrm{shift}} = \lambda_1$. Now the minimum eigenvalue λ_1' of S' is zero, and the relevant spectral range is from zero to $\lambda_{\max}'(t) \sim O(1/t)$. Thus we have the modified scheme

$$\Delta t(t) = \beta t, \tag{36}$$

where the choice of constant $\beta \ll 1$ is discussed in Section 6.2. We call this 'logarithmic timestepping' since it is uniform in $\ln(t)$. We use CN with this timestepping scheme to approximate the evolution

$$\mathbf{U}'(t) = e^{-S't}\mathbf{U}_O \tag{37}$$

in the usual way (namely Eq. (33) with the substitution S' for S). It then requires very little numerical effort to undo the shift:

$$\mathbf{U}(t) = e^{-\lambda_{\mathrm{shift}}t}\mathbf{U}'(t). \tag{38}$$

In practice λ_{shift} can be computed (using the LAPACK routine DSTEVX [50]) in negligible $O(n_z)$ time since λ_1 of S is simply the lowest eigenvalue of the matrix S_m for the first mode $m = 1$. For each CN step the tridiagonal system of equations is solved using LAPACK [50], separately for each block S_m comprising S.

This logarithmic scheme evolves modes of a given decay constant accurately during the time that they contribute, then inaccurately (without increasing their amplitudes) once they reach a negligible (exponentially small) contribution. However, we need a recipe to initialize the method, since Eq. (36) as it stands would require infinite numerical effort to handle small times $t \to 0$. One solution would be to start with an analytic solution for $\mathbf{U}(t)$ at some small $t = t_0$, and evolve from there. This would require simple forms for $\kappa(z)$ and $\mu_a(z)$ in the uppermost part of the medium. Instead we decided to use an initial CN phase from $t = 0$ to $t = t_0$, with constant step size Δt_0 sufficiently small that even the highest decay-rate modes of the lattice decay to safe amplitudes. We call this phase X (see Fig. 6(b)). A reliable recipe for the choice of t_0 and Δt_0 is given in Appendix A.

This timestepping scheme is second-order accurate in β, as verified in Section 6.2. The exponential decay of modes means that either assumption (i) or (ii) would have to be very strongly violated to invalidate the

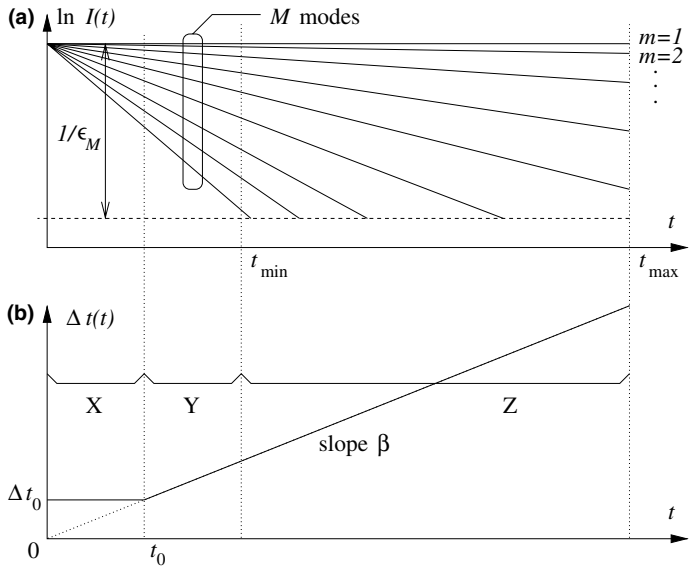


Fig. 6. (a) Schematic of mode decay and error parameter ϵ_M. A constant M modes are evolved from $t = 0$ to $t = t_{min}$; after this they are successively 'killed' when their contribution falls below the error level ϵ_M. (b) CN timestep $\Delta t(t)$ as a function of current time t. Assuming that $t_0 \leqslant t_{min}$, the evolution falls into three phases labelled X, Y, and Z.

above method. However, once the above-calculated $\mathbf{U}(t)$ is substituted into Eq. (30), we note a third assumption becomes required: (iii) that after the projection equation (30), coefficients contribute approximately equally to the signal without significant cancellation. Unfortunately we will see in Section 6.2 that for a certain corner of the (R, t) plane assumption (iii) is not valid and therefore larger errors result.

5.4. Time-dependent truncation of the mode sum

We truncate the number of terms in the sum Eq. (16) to a fixed finite number M as we evolve from $t = 0$ to $t = t_{min}$ (phases X and Y). We then successively shrink this number for the remaining segment of time $t \in [t_{min}, t_{max}]$ (phase Z). This is sketched in Fig. 6(a). This latter truncation in phase Z results in a significant improvement in the scaling of numerical effort (from $\log t_{max}$ to independent of t_{max}).

The truncation can be performed with exponentially small errors. This is evident from Fig. 7(a) where it is clear that at each fixed time t, the mode contributions die exponentially with m, and that as t increases, fewer and fewer modes are needed to give an accurate answer. This exponential convergence of the mode sum can be understood as the result of the growth with m of the effective absorption coefficient $\mu_{a,m}$ in Eq. (21). Alternatively it can be viewed as the result of the exponential tails of the transverse Fourier representation of the heat kernel [47] as it contracts in time, as sketched in Fig. 3(b) and explained in Appendix B.

We define a desired truncation error level ϵ_M. Either by considering the decay rate equation (21) in the top layer of the medium, or using the reasoning in Appendix B, we have an estimate for the transverse wavenumber at which truncation can occur, at a given time t,

$$k_{max}(t) = \sqrt{\frac{\ln(1/\epsilon_M)}{c\kappa_0 t}}. \tag{39}$$

Notice that it decays with time like $t^{-1/2}$. Using the asympotic form for the locations of Bessel zeros, we have a linear relation between wavenumber and mode number,

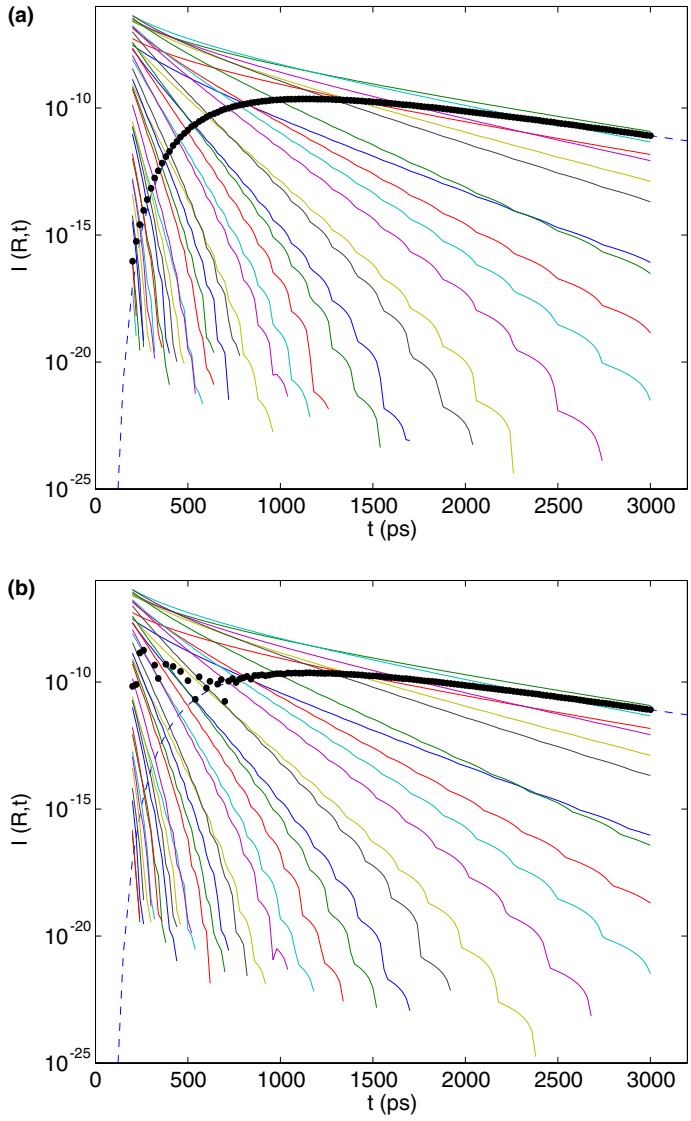


Fig. 7. Contribution due to each mode m (shown as lines) which are summed to give the computed signal $I(R, t)$ (shows as dots), for a homogeneous medium (SYS1), at $R = 40$ mm, $t_{\min} = 200$ ps, giving $M = 38$. The dashed line shows the exact analytic answer. Note the logarithmic vertical scale. (a) Uses consistent time-stepping scheme presented in Section 5.3, and (b) uses a very slight modification which results in inconsistency and therefore huge relative errors at early times where cancellation is important (see Section 8.1).

$$M \approx \frac{k_{\max} a}{\pi}. \tag{40}$$

Therefore, we compute M using Eqs. (39) and (40) with the substitution $t = t_{\min}$, finally rounding M up to the nearest integer. We also decide when to 'kill' each mode m in phase Z by rearranging Eq. (39) to give

$$t_{\mathrm{kill},m} = \frac{\ln(1/\epsilon_{\mathrm{M}})}{c\kappa_0 k_m^2}, \tag{41}$$

where instead of $k_{\max}$ the transverse wavenumber k_m appropriate for each mode has been used.

Our estimates in this section have been based on a homogeneous system with κ equal to its surface value in the inhomogeneous system. This will only cause inaccuracies if $\kappa(z)$ changes substantially over a z range equal to the kernel width $\sqrt{2\kappa_0 c t_{\min}}$. In practise we have found no problems in inhomogeneous systems of interest. We ensure that the error level is set many orders of magnitude lower than the dynamic range γ; we typically use $\epsilon_{\mathrm{M}} = 10^{-12}$, giving in our setting $M_0 \sim 10^2$.

It could also be possible to avoid the above estimation formulae by adding modes successively until convergence of $I(R, t)$ is achieved within the entire desired domain of the (R, t) plane, however we have not implemented this.

5.5. Algorithm summary

We are given the functions $\kappa(z)$ and $\mu_{\mathrm{a}}(z)$ defining the medium, the parameters $t_{\min}$, $t_{\max}$ and $R_{\max}$ defining the region of interest in the (R, t) plane, the detector weighting A, and the error parameters ϵ_{a}, ϵ_{L}, ϵ_{M}, and ϵ_{X} whose values have all been discussed earlier. The finite-difference parameters h and β are then chosen as discussed in Section 6. The box sizes a and L are chosen via Eqs. (13) and (15). The maximum mode number M is chosen as in Section 5.4. We then compute the wavenumbers k_m and the Bessel coefficients $B_m\psi_m(R)$ appearing in Eq. (22), using lookup tables of χ_m and $J_1(\chi_m)$. The 1D system tridiagonal matrices S_m are then set up as in Section 5.2, and λ_{shift} is found as in Section 5.3. The phase X parameters t_0 and Δt_0 are found as in Appendix A. Each mode m is then initialized and evolved as in Section 5.3 until its time to be 'killed' given by Eq. (41), or $t_{\max}$, whichever comes sooner. We note that, until it is 'killed', every mode undergoes an identical CN timestepping scheme. The signals from each mode are summed using Eq. (22).

The signal output $I(R, t)$ is computed for a requested set of R values, on a uniform time grid. Fig. 11(a) shows representative resulting $I(R, t)$. Linear interpolation is used to resample from the CN timesteps to the uniform time grid. This resampling is $\mathrm{O}(\beta^2)$ accurate, so preserves the basic scaling with β (see Section 6.2).

A software implementation (C callable from a MATLAB interface) is freely available online at http://www.cims.nyu.edu/~barnett/software.html.

6. Convergence and accuracy

In this section, we demonstrate convergence with respect to h (finite-difference lattice spacing) and β (timestepping parameter), and validate against a known solution. For convergence with respect to h (Section 6.1) we use relative signal errors, however for the remaining tests we introduce an error measure more relevant for our application. We define the error measure

$$E(t) \equiv \frac{I(t) - I_{\mathrm{exact}}(t)}{I_{\mathrm{err}}(t)}, \tag{42}$$

where $I_{\mathrm{exact}}(t)$ is some 'exact' (or reference) signal. If $E(t) \leqslant 1$ for all t we define this to be acceptable error. Note that the denominator is the acceptable error level $I_{\mathrm{err}}(t)$ defined in Eq. (5) and is controlled by two error parameters: we will use the choices $\epsilon = 10^{-2}$ and $\gamma = 10^5$ throughout. Therefore, for large signals, $E(t)$ is a measure of % relative error.

6.1. Lattice spacing h

The finite-difference scheme is second-order accurate in h, as demonstrated by the 2-layer results for relative error shown in Fig. 8. The convergence prefactor depends strongly on time and system optical properties. For these results, the system with the largest errors of all six systems of Table 1 has been chosen.

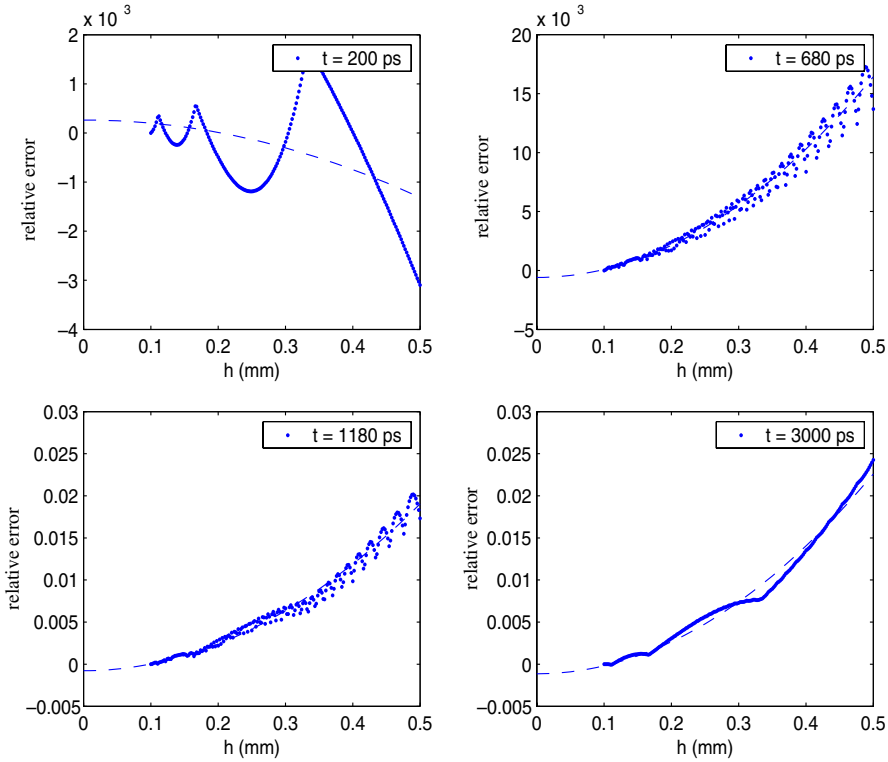


Fig. 8. Convergence of computed signals $I(R, t)$ with finite-difference lattice spacing h, at four choices of t, at $R = 20$ mm. The system was SYS5, which had the largest convergence errors of all the systems from Table 1. The time-stepping scheme was held fixed ($\beta = 0.05$, $\Delta t_0 = 0.04$ ps, see Section 5.3). Normalized changes in signal relative to the value for $h = 0.1$ mm are shown. The dashed curves show best-fit $O(h^2)$ convergence, and are intended only as a guide to the eye.

Rather than having smoothly-varying optical properties, as finite-difference methods are usually intended to handle, these 2-layer systems have a maximal discontinuity in both optical coefficients at the interface between the upper and lower layer. Therefore, we consider this to be a stringent test of the method. Given a desired relative error $\epsilon = 10^{-2}$, this indicates $h = 0.3$ mm is sufficient in this worse-case scenario.

Notice that there is sometimes rapid variation of error as a function of h, resulting in quasi-random scatter, which also dies like $O(h^2)$. This results from the fixed depth of the interface falling at different fractions across a cell as the lattice changes size. Correct cell-averaging (Eqs. (24) and (25)) is needed to ensure this effect does not spoil the $O(h^2)$ convergence. Note that the choice of absorption diagonal term in Eq. (23), namely a $\mu_{a,m}$ that has been averaged across two cells, may also not be optimal.

6.2. Timestepping parameter β

Our CN logarithmic timestepping scheme is second-order accurate in β, as demonstrated by the results in Fig. 9 for a homogeneous system. We postpone the theoretical analysis of this result to a future publication. We only note that the result is reasonable since the constant-Δt CN method is $O(\Delta t^2)$ accurate over a fixed interval of time, and in any small window of time our method resembles this method. It is also important to note that β can be held fixed as the lattice spacing h is reduced: timestepping error at fixed β, since it is related to the physical modal structure, is independent of h. In the figure, oscillations in error are visible,

due to the linear interpolation between the CN timesteps. An improved interpolation scheme (for instance, cubic) could improve the error prefactor but not its basic scaling.

The error $E(t)$ has a peak at early times, which becomes more dominant and wide as R grows (see Fig. 9). This peak is due to breakdown of assumption (iii) from Section 5.3. As shown in Fig. 7(a), the signal at larger R takes some time to grow from very small values; this is simply because the diffusing light has not yet reached that radius (we are in the exponential tail of the spreading fluence distribution). Therefore, at large R and small t a large number of modes m cancel each other almost exactly, allowing the contribution of rapidly-decaying modes to be important. These rapidly-decaying modes were the ones inaccurately evolved in our timestepping scheme, causing the larger errors. Note that at later times the cancellation disappears and higher accuracy is achieved. It is this early peak which sets our acceptable choice of β. Given an $R_{\max}$ of 40 mm and typical tissue optical parameters, this peak is only an order of magnitude larger than later errors, and $\beta = 0.04$ is acceptable. However, if $R_{\max}$ becomes much larger then the necessary β becomes smaller rendering the scheme less efficient. Our form of acceptable error level defined in Eq. (5), as well as our choice of ϵ and γ, have significant effects on the largest acceptable β. For instance, decreasing γ to 10^4 would allow $\beta = 0.07$ to be acceptable, with a corresponding increase in calculation speed.

Fixing $\beta = 0.04$, we have verified that errors are acceptable in a variety of 2-layer systems, as shown in Fig. 10. Because of the large inhomogeneity $\alpha = 10$ tested here, we believe that the errors will also be acceptable for the less extreme systems usually arising in the biomedical application. Notice that, as with the homogeneous system, the initial peak in $E(t)$ is responsible for the maximum error levels.

6.3. Verification in 2-layer system

We have independently verified the accuracy of our method in the case of 2-layer systems, by comparing against the real-space analytic power series method of Tualle et al. [33,52] (with summation parameters

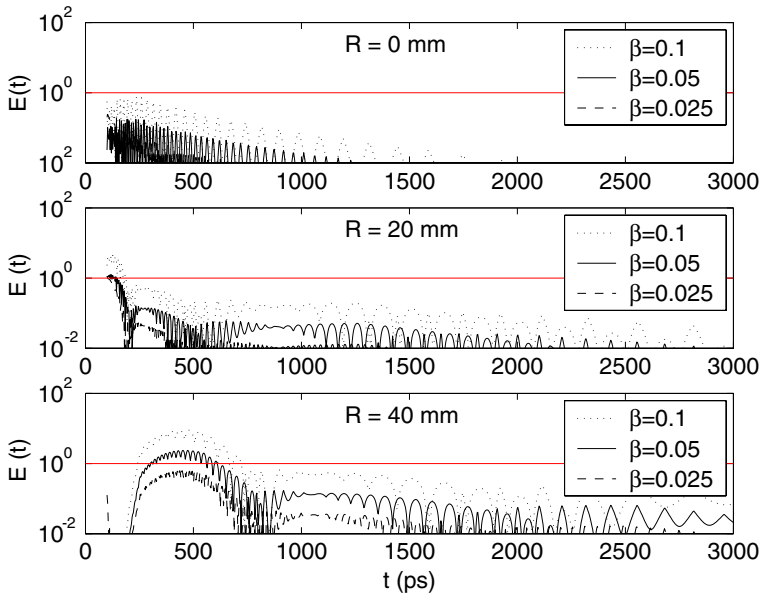


Fig. 9. Convergence of timestepping scheme: signal error measure $E(t)$ as a function of time t, for three timestepping choices of β, at three separations R. The system was homogeneous (SYS1 from Table 1). Other parameters were $h = 0.5$ mm. Error level $E(t)$, which for large signals measures % relative error, is defined in Eq. (42), using the known analytical formula for $I_{\text{exact}}(t)$.

$n = 2$, $l = 20$; see [33]). Fig. 11 illustrates the agreement between the two methods for an example 2-layer system with $\alpha = 4$. The agreement is excellent, typically at the 0.5% level. Notice that the fact that the peak values agree demonstrates the accuracy of the *relative* signals at different separations R (the peaks have not been normalized to have the same values). In terms of the error level $E(t)$, we see that choices of method parameter values h and β derived from the above convergence experiments result in acceptable error. Because the method of Tualle et al. uses the detection of fluence gradient (that is, $A = 0$ in Eq. (4)), our comparison was made using this choice; in any case the variation with A is very slight. Finally we note that our method is stable over much wider variety of system parameters than that of Tualle et al. [52]: we have found that for 2-layer systems where the lower layer is significantly less scattering than the upper, their method can give unusable (nonconvergent) results.

7. Scaling of computational effort

We would like to know how the computational effort scales with the various system and numerical parameters. Nearly all the effort is spent in timestepping the finite-difference method. The cost of increasing the number of desired separations R or interpolating onto more time points t is negligible.

Defining f to be the cost of a single CN step for a single mode m (i.e. a finite-difference system of size n_z), we can estimate total effort with the help of the schematic Fig. 6. Phase X requires $1/\beta$ steps, and phase Y requires $\ln(t_{\min}/t_0)/\beta$ steps, both involving all M modes. In phase Z the number of modes involved at time t is $M\sqrt{t_{\min}/t}$. The total effort is therefore

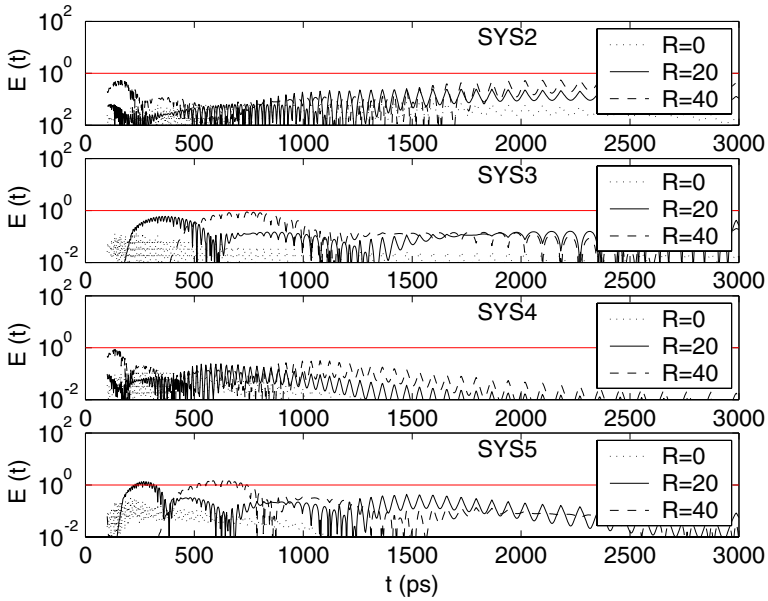


Fig. 10. Verification that timestepping errors are at acceptable levels for the four highly-inhomogeneous systems (SYS2–SYS5) from Table 1, at three source–detector separations R (in mm). Timestepping was fixed at $\beta = 0.04$, with $h = 0.5$ mm. Error measure $E(t)$ is defined in Eq. (42), with $I_{\text{exact}}(t)$ found via uniform CN timestepping at $\Delta t = 0.25$ ps.

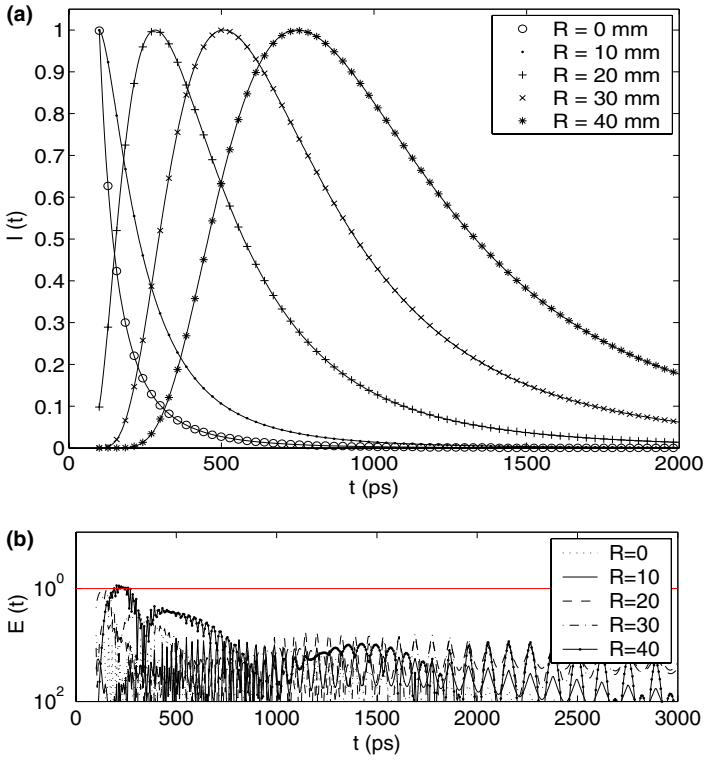


Fig. 11. Verification of method accuracy for the 2-layer system SYS6 (defined in Table 1). (a) Shows computed signals $I(t)$ using our method (symbols) against the signals $I_{\text{exact}}(t)$ computed with the method of Tualle et al. (solid lines). The five source–detector distances R are as labelled. A single multiplicative prefactor was fitted, then used to compare the two sets of signals for all values of R. At each R, $I_{\text{exact}}(t)$ is normalized to have peak value 1 in the time range of interest, and $I(t)$ is normalized by the same factor. (b) Shows error measure $E(t)$, defined in Eq. (42), resulting from comparing the two methods. Timestepping was fixed at $\beta = 0.04$, with $h = 0.5$ mm and $t_{\min} = 100$ ps.

$$F = \frac{Mf}{\beta}\left[1 + \ln\left(\frac{t_{\min}}{t_0}\right)\right] + \int_{t_{\min}}^{t_{\max}} \mathrm{d}t \frac{f}{\beta t} M \sqrt{\frac{t_{\min}}{t}} \tag{43}$$

$$= \frac{Mf}{\beta}\left[3 + \ln\left(\frac{t_{\min}}{t_0}\right) - 2\sqrt{\frac{t_{\min}}{t_{\max}}}\right] \tag{44}$$

$$\approx \frac{qMf}{\beta}, \tag{45}$$

where we have used the fact that $t_{\max} \gg {}_{\min}$. The number q is of order 3–7. The choice of β depends on the required relative accuracy (choice of ϵ); for $\epsilon = 10^{-2}$ appropriate values were discussed in Section 6.2. In the remain required We treat β as a given constant, which is chosen (as discussed in Section 6.2).

7.1. Finite system

We now ask, given the finite cylindrical system of fixed depth a and radius L, how the total effort F scales. We make use of $f = pn_z = pL/h$, where p is the effort per lattice node per timestep in the CN scheme.

Since p is of order a few floating-point operations (flops), this gives F units of flops. Substituting for M using $t = t_{\min}$ in Eqs. (39) and (40), we get

$$F(a,L) \approx \frac{qp\sqrt{\ln(1/\epsilon_{\mathrm{M}})}}{\pi\beta} \cdot \frac{aL}{h\sqrt{c\kappa_0 t_{\min}}}. \tag{46}$$

The first factor is a number of order 10^3–10^4. The second factor is a ratio of two areas: the system's cross-sectional area aL, and the product of lattice size and the heat kernel spreading size by time $t_{\min}$. Furthermore, although we postpone the derivation and tests to future work, the h required for accuracy ϵ_h at early times $t = t_{\min}$ can be estimated (using a homogeneous system) as

$$h \approx \sqrt{\epsilon_h c\kappa_0 t_{\min}}, \tag{47}$$

which follows from the kernel spreading size and the $\mathrm{O}(h^2)$ lattice convergence. We can set $\epsilon_h \sim \epsilon \sim 10^{-2}$. Substituting this gives

$$F(a,L) \approx \frac{qp\sqrt{\ln(1/\epsilon_{\mathrm{M}})}}{\pi\beta\sqrt{\epsilon_h}} \cdot \frac{aL}{c\kappa_0 t_{\min}}, \tag{48}$$

where the first factor is of order 10^4–10^5, and the second factor ratio is the fraction of the cross-sectional area occupied by the spreading kernel by time $t_{\min}$. Notice that $t_{\max}$ has become irrelevant: the limit $t_{\max} \to \infty$ can be taken with almost no extra effort. This contrasts common constant-Δt schemes where the effort scales linearly with $t_{\max}$. This surprising result is due to our logarithmic timestepping and successive 'killing' of irrelevant modes in phase Z.

7.2. Infinite system

In Section 4, we found the a and L necessary for approximation of the infinite system by a finite one. Using Eqs. (13) and (15), assuming $R_{\max} \ll a$ and $\epsilon_{\mathrm{L}} = \epsilon_{\mathrm{a}}$ we can estimate

$$aL \approx c\kappa_{\max} t_{\max} \ln(1/\epsilon_{\mathrm{a}}). \tag{49}$$

Thus area grows linearly in time, as it must in a diffusion process. Combining this with Eq. (48) gives the effort for the semi-infinite problem

$$F \approx \frac{qp[\ln(1/\epsilon_{\mathrm{a}})]^{3/2}}{\pi\beta\epsilon_h^{1/2}} \cdot \frac{\kappa_{\max}}{\kappa_0} \cdot \frac{t_{\max}}{t_{\min}}, \tag{50}$$

where we have assumed $\epsilon_{\mathrm{M}} = \epsilon_{\mathrm{a}}$ for convenience. The first factor is of order 10^5–10^6, the second is a ratio bounded by α, and the third is of order 50 in the biomedical application. This gives a typical $F \sim 10^7$ flops. This is consistent with typical runtimes of 0.1–0.3 s which we observe on a 1 GHz Pentium III processor (whose memory-limited speed is less than 10^8 flops per second), with $t_{\min} = 100$ ps and $t_{\max} = 5000$ ps. Thus we are nearly an order of magnitude faster than the $\sim$1 s run time quoted by Martelli et al. [34,35], and, when a large grid of R and t samples is needed, by Tualle et al. [33].

In summary, with error parameters fixed, our infinite system method scales linearly with $t_{\max}/t_{\min}$, at most linearly with α, and inversely with β.

8. Discussion

We have described a method well-adapted for cylindrical systems of finite radius and depth. We have applied it to the case of infinite radius and depth by using error estimates derived from a homogeneous

medium. Because the diffusion inhomogeneity α is not too large, these estimates are not excessively wrong (Section 4). However, the cross-sectional area of the required finite medium grows linearly in time.

There are two potential improvements in effort scaling for the infinite system: (1) Use of a nonuniform finite-difference lattice, where h increases with depth (this could be extended to be adaptive in time, or furthermore, to have a maximum depth L which increases in time). (2) Non-uniform representation in k-space, in the manner of Greengard and Lin [48]. This latter approach is specifically designed to represent the infinite-space heat kernel efficiently, that is, using as few Fourier modes as possible. These modes are logarithmically spaced in wavenumber. The rigorous convergence proofs in their work would need to be relaxed to handle the inhomogeneous medium case. Since currently a large fraction of effort is spent in high-m modes, the logarithmic spacing would be beneficial. The use of these two changes together could change the effort scaling from $O(t_{\max}/t_{\min})$ to $O(\ln[t_{\max}/t_{\min}])$, although since in the DOT application $t_{\max}/t_{\min}$ is only of order 50 this may be limited to an order of magnitude improvement. In other applications this could be more important.

In general, there is much future analytic work that could be done to improve the error estimates in inhomogeneous systems. For instance, can tighter estimates for a and L be found? (A speed-up factor of up to α is available here.) Can rigorous bounds be placed on the error due to truncation to a given a and L? Can better estimates for M be found? Similarly, we have not analysed in detail how the acceptable h depends on $t_{\min}$ and the optical properties (the scaling Eq. (47) need not be correct for inhomogeneous systems).

It is worth mentioning why a 2D finite-difference approach, that is, a cylindrical lattice in (ρ, z), is likely to be less successful than the method we have presented. The main problem with a 2D system is that higher Fourier modes cannot be 'killed' in the manner of Section 5.4, so efficiency would be lower. The simplicity of the CN implementation and small dataset size (which can fit within a CPU cache) in 1D also tend to favor our choice. However, it would also be interesting to explore the use of our logarithmic timestepping scheme for parabolic 3D finite-difference applications without symmetry.

Currently our method is limited by the fact that our logarithmic timestepping scheme performs poorly when there is a lot of cancellation between mode contributions; this occurs at large R and early times. This is a fundamental problem for the CN scheme that it cannot handle a large range of decay rates accurately (large decay rates are relevant when there is cancellation; small rates are relevant at later times when cancellation disappears). One potential solution is to change λ_{shift} as a function of m in a consistent fashion (e.g. Eq. (51) below). A more radical approach, which may be competitive in speed, is to solve for (some subset of) eigenvalues and eigenvectors of the matrices S_m and perform an exact Method of Lines evolution, bypassing timestepping schemes altogether.

8.1. Consistency and cancellation at early times

Here, we discuss a detail which can be skipped on first reading. By grouping together the M 1D finite-difference problems into a single ordinary differential equation (28), we have simplified our discussion of timestepping. However, initially in our research each 1D evolution problem equation (26) was handled using a separate time-stepping scheme. This was deemed necessary to preserve accuracy: the rapidly-decaying (high-m) finite-difference problems could either be given smaller timesteps Δt or a different λ_{shift} adapted to handling their large decay rates more accurately. A typical resulting signal at large R is shown in Fig. 7(b): here for illustrative purposes we changed the value of Δt_0 during phase X very slightly for each mode m. The evolution of individual mode contributions shows almost no perceptible difference from Fig. 7(a) (their accuracy remains the same), however their sum is radically different. The resulting signal is utterly wrong for the first 700 ps. The accuracy achievable is limited to a constant fraction (here about 10^{-3}) of the *largest terms* in the sum, which here are many orders of magnitude larger than the true signal.

Therefore, there is a consistency condition. It is not how accurately we can evolve each mode but rather whether the scheme is consistent between modes which controls the signal accuracy when there is a high degree of cancellation. Cancellation at large R and small t is present for the basic reason that the Fourier representation of the Gaussian kernel does a poor job of representing the (exponentially-smaller) tails to high *relative* accuracy. It is a pitfall that we cannot avoid in a transverse mode representation. When a scheme is consistent, then the only limit on the size of signal that can be accurately calculated is the machine precision (in Fig. 7(a) if R is further increased, the minimum achievable accurate signal is $\sim 10^{-15}$ of the maximum term in the sum, due to the double precision arithmetic).

We aim to analyze the consistency condition theoretically in future work. We note that consistency requires that a single timestepping scheme is used (that is, all modes m share a common choice of $\Delta t(t)$, as in Section 5.3). We have also found empirically that consistency is preserved by a m-dependent λ_{shift} when it takes the form

$$\lambda_{\text{shift},m} = \lambda_{\text{shift}}(k_m) = c_1 + c_2 k_m^2, \tag{51}$$

where c_1 and c_2 do not depend on m. There is potential to improve the method by using a λ_{shift} of this form.

9. Conclusions

We have described and tested an efficient new method for solution of the time-dependent diffusion equation (6) in stratified media with arbitrary absorption and diffusion coefficient depth profiles. Our method is geared towards the DA to photon transport forward and inverse problems arising in biomedical imaging, in particular, DOT. In this application a typical run time on a 1 GHz CPU is 0.2 s, which is several times faster than quoted for current methods for less-flexible 2- or 3-layer systems. Our algorithm is reliable and includes automatic choice of all numerical parameters (apart from β discussed in Section 6.2); this is important for unsupervised use in iterative inverse problems.

The basic method applies to a finite cylindrical system, using a sum over transverse modes each of which requires solution of a 1D diffusion problem. A novel logarithmic timestepping scheme (one which is generally applicable to sourceless parabolic problems) allows efficient solution of the 1D problems using finite-difference methods. In order to extend the method to the semi-infinite system case, we derived and tested expressions for the finite system size needed to accurately emulate the infinite system.

We have analyzed numerical errors and convergence in depth, using an error measure acceptable for the DOT application. A recurring theme is that because the diffusion inhomogeneity α arising in our application is not too large, we can use convergence estimates derived for a homogeneous system. In particular, the fact that the heat kernel has exponential tails in position space is used for the finite system size error estimates, and the fact that the same kernel has exponential tails in Fourier space is used for truncating the mode sum. Both therefore give exponential convergence. The finite-difference evolution is $O(h^2)$ and $O(\beta^2)$ accurate. Our tests involved a variety of highly-inhomogeneous 2-layer media, although we emphasize that the method handles arbitrary depth-dependent media. The conclusion that the effort scales like $1/t_{\min}$ for the finite system, and $t_{\max}/t_{\min}$ for the infinite system, means that for speed in DOT applications it is important to increase the desired $t_{\min}$ to as large a value as possible.

Within our general framework there is much flexibility. We have presented the case of reflection geometry, however the modification to transmission through a finite-thickness slab is trivial. We expect that major speed gains could result from implementing a nonuniform finite-difference lattice, and a nonuniform representation of the sum over transverse wavenumber. Other than optical tomography, our method can also be of use in impulsive stratified heat conduction problems, and the inverse problem of determining thermal conductivity depth profiles.

Acknowledgement

In completing this work the author has benefited greatly from discussions with Jonathan Goodman, Mike Shelley, Leslie Greengard, Robert Kohn, Marsha Berger, Donna Calhoun, Dwight Barkley, David Boas, Joe Culver, and Ross Barnett. We also thank Jean-Michel Tualle for providing computer code for his 2-layer method, and the comments of the anonymous reviewers. The author is supported by the Courant Institute at New York University.

Appendix A. Parameter choice for initial Crank–Nicolson phase

The logarithmic CN scheme of Eq. (36) relies on starting at a finite time t_0 with all unphysical high decay-rate modes exponentially damped. We do this via an initial phase of CN with constant time-step Δt_0. We call this phase X (see Fig. 6(b)). In order for damping to be reliable, we need to consider two cases separately: t_0 either less than or greater than $t_{\min}$.

We first assume $t_0 \leqslant t_{\min}$. We consider it a good choice to impose $\Delta t_0 = \beta t_0$, in which case $n \approx 1/\beta$ steps will be involved. The highest decay rate of the finite-difference lattice system S defined in Section 5.2 can be estimated as

$$\lambda_h \approx c\kappa_{\max}\left(\frac{4}{h^2} + k_{\mathrm{M}}^2\right) + c\mu_{\mathrm{a,max}}. \tag{A.1}$$

This follows from considering Eq. (23) for the lattice 'optical mode' (alternating values ± 1), and $m = M$. The CN amplitude damping coefficient for this mode is

$$a_h^{\mathrm{CN}}(\Delta t_0) = \frac{1 - \frac{1}{2}\lambda_h \Delta t_0}{1 + \frac{1}{2}\lambda_h \Delta t_0} \approx -1 + \frac{4}{\lambda_h \Delta t_0} - \cdots \tag{A.2}$$

We truncate the series expansion in $1/\lambda_h \Delta t_0$ at the term shown. After $n = 1/\beta$ such steps, we require the damping to have reached the small ratio ϵ_{X}, which we typically set to 10^{-9}. Therefore

$$\epsilon_{\mathrm{X}} = \left(\left|a_h^{\mathrm{CN}}(\Delta t_0)\right|\right)^n \approx \left(1 - \frac{4}{\lambda_h \Delta t_0}\right)^n \approx \mathrm{e}^{-\frac{4n}{\lambda_h \Delta t_0}}. \tag{A.3}$$

Thus choosing $n = 1/\beta$ gives our estimate

$$\Delta t_0 = \frac{4}{\beta \lambda_h \ln(1/\epsilon_{\mathrm{X}})}. \tag{A.4}$$

The total time is then set by $t_0 = \Delta t_0/\beta$.

If the resulting t_0 turns out to be greater than $t_{\min}$, then we cannot guarantee that the damping has been sufficient by the time signals are needed at $t = t_{\min}$. Therefore in this case we fix $t_0 = t_{\min}$, which removes phase Y (see Fig. 6(b)). We then find the Δt_0 needed to perform damping by this time, using $n = t_{\min}/\Delta t_0$ steps, which will differ from $1/\beta$. Substituting this n into Eq. (A.3) gives in this case

$$\Delta t_0 = \sqrt{\frac{4 t_{\min}}{\lambda_h \ln(1/\epsilon_{\mathrm{X}})}}. \tag{A.5}$$

Appendix B. Truncation of heat kernel in Fourier space

We can use a one-dimensional model to bound the error due to truncation in the Fourier domain of the heat kernel at a given time $t > 0$. See Fig. 3(a) for a schematic in the discrete mode case. The error in 2D or

3D is essentially the same. The result is well known (our presentation parallels the discrete wavenumber case presented in Refs. [47,48]). The fluence in a medium with homogenous properties κ, μ_a, with the single spatial coordinate x obeys

$$\frac{1}{c}\frac{\partial\phi}{\partial t} = \kappa\frac{\partial^2\phi}{\partial x^2} - \mu_a\phi, \tag{B.1}$$

which has the fundamental solution

$$\phi(x,t) = \frac{1}{\sqrt{4\pi\kappa ct}}\mathrm{e}^{-\frac{x^2}{4\kappa ct}-\mu_a ct}. \tag{B.2}$$

This has spatial Fourier transform

$$\tilde{\phi}(k,t) \equiv \int_{-\infty}^{\infty}\phi(x,t)\mathrm{e}^{-\mathrm{i}kx}\mathrm{d}x = \mathrm{e}^{-(k^2\kappa+\mu_a)ct}. \tag{B.3}$$

We call $\phi_{\mathrm{trunc}}(x,t)$ the inverse Fourier transform of $\tilde{\phi}(k,t)$ using a truncated wavenumber range $k \in [-k_{\max}, k_{\max}]$. The resulting error is

$$\begin{aligned}\Delta\phi(x,t) \equiv |\phi_{\mathrm{trunc}}(x,t) - \phi(x,t)| &= \left|\frac{1}{2\pi}\left(\int_{-\infty}^{-k_{\max}} + \int_{k_{\max}}^{\infty}\right)\mathrm{e}^{-(k^2\kappa+\mu_a)ct}\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k\right| \\ &\leqslant \frac{1}{2\pi}\left(\int_{-\infty}^{-k_{\max}} + \int_{k_{\max}}^{\infty}\right)\mathrm{e}^{-(k^2\kappa+\mu_a)ct}\,\mathrm{d}k \\ &\leqslant \mathrm{e}^{-k_{\max}^2\kappa_0 ct}\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{e}^{-(k^2\kappa+\mu_a)ct}\,\mathrm{d}k = \mathrm{e}^{-k_{\max}^2\kappa ct}\phi(0,t).\end{aligned} \tag{B.4}$$

The second inequality can be shown by shifting the integration variable by $\pm k_{\max}$ as appropriate, and dropping the cross term from the resulting $(k \pm k_{\max})^2$ factor in the exponent. Therefore, the relative error $\Delta\phi(x,t)/\phi(0,t)$ (error relative to the spatial peak) at time t is bounded by $\mathrm{e}^{-k_{\max}^2\kappa ct_{\min}}$. Setting this relative error equal to ϵ_M, and using the top-layer diffusion constant $\kappa = \kappa_0$, gives the estimate Eq. (39).

References

[1] D.A. Boas, D.H. Brooks, E.L. Miller, C.A. DiMarzio, M. Kilmer, R.J. Gaudette, Q. Zhang, Imaging the body with diffuse optical tomography, IEEE Signal Processing Magazine 18 (6) (2001) 57–75.

[2] A. Villringer, B. Chance, Non-invasive optical spectroscopy and imaging of human brain function, Trends Neurosci. 20 (1997) 435–442.

[3] S.R. Hintz, W.-F. Cheong, J.P. van Houten, D.K. Stevenson, D.A. Benaron, Bedside imaging of intracranial hemorrhage in the neonate using light: comparison with ultrasound, computed tomography, and magnetic resonance imaging, Pediatr. Res. 45 (1999) 54–59.

[4] D. Grosenick, K.T. Moesta, H. Wabnitz, J. Mucke, C. Stroszczynski, R. Macdonald, P.M. Schlag, H. Rinneberg, Time-domain optical mammography: initial clinical results on detection and characterization of breast tumors, Appl. Opt. 42 (2003) 3170–3186.

[5] S. Fantini, M.A. Franceschini-Fantini, J.S. Maier, S.A. Walker, B. Barbieri, E. Gratton, Frequency-domain multichannel optical detector for noninvasive tissue spectroscopy and oximetry, Opt. Eng. 34 (1995) 32–42.

[6] A. Klose, A.H. Hielscher, K.M. Hanson, J. Beuthan, Three-dimensional optical tomography of a finger joint model for diagnostic of rheumatoid arthritis, in: D.A. Benaron, B. Chance, M. Ferrari, M. Kohl (Eds.), Photon Propagation in Tissue IV, Proc. SPIE 3566, 1998, pp. 151–160.

[7] A. Torricelli, A. Pifferi, P. Taroni, E. Giambattistelli, R. Cubeddu, *In vivo* optical characterization of human tissues from 610 to 1010 nm by time-resolved reflectance spectroscopy, Phys. Med. Biol. 46 (2001) 2227–2237.

[8] F.E. Schmidt, M.E. Fry, E.M.C. Hillman, J.C. Hebden, D.T. Delpy, A 32-channel time-resolved instrument for medical optical tomography, Rev. Sci. Instruments 71 (2000) 256–265.

[9] S.R. Arridge, Optical tomography in medical imaging, Inverse Problems 15 (1999) R41–R93.

[10] K.M. Hanson, G.S. Cunningham, S.S. Saquib, Inversion based on computational simulations, in: G.J. Erickson, J.T. Rychert, C.R. Smith (Eds.), Maximum Entropy and Bayesian Methods, Kluwer Academic, Dordrecht, 1998, pp. 121–135.
[11] A. IshimaruWave Propagation and Scattering in Random Media, vol. 1, Academic, 1978.
[12] J.D. Moulton, Diffusion modelling of picosecond laser pulse propagation in turbid media, Masters thesis, McMaster University, 1990.
[13] S.S. Saquib, K.M. Hanson, G.S. Cunningham, Model-based image reconstruction from time-resolved diffusion data, in medical imaging: image processing, Proc. SPIE 3034 (1997) 369–380.
[14] A.H. Barnett, J.P. Culver, A.G. Sorensen, A. Dale, D.A. Boas, Robust inference of baseline optical properties of the human head with 3D segmentation from magnetic resonance imaging, Appl. Opt. 42 (2003) 3095–3108.
[15] S. Oh, A.B. Milstein, R.P. Millane, C.A. Bouman, K.J. Webb, Source–detector calibration in three-dimensional Bayesian optical diffusion tomography, J. Optical Soc. Am. A 19 (2002) 1983–1993.
[16] A.H. Hielscher, R.E. Alcouffe, R.L. Barbour, Comparison of finite-difference transport and diffusion calculations for photon migration in homogeneous and heterogeneous tissues, Phys. Med. Biol. 43 (1998) 1285–1302.
[17] A.H. Hielscher, S.L. Jacques, L. Wang, F.K. Tittel, The influence of boundary conditions on the accuracy of diffusion theory in time-resolved reflectance spectroscopy of biological tissues, Phys. Med. Biol. 40 (1995) 1957–1975.
[18] L. Wang, S.L. Jacques, L. Zheng, MCML-Monte Carlo modeling of light transport in multi-layered tissues, Comput. Methods Programs Biomed. 47 (1995) 131–146.
[19] A.H. Hielscher, H.L. Liu, B. Chance, F.K. Tittel, S.L. Jacques, Time-resolved photon emission from layered turbid media, Appl. Opt. 35 (1996) 719–728.
[20] A.H. Hielscher, A.D. Klose, K.M. Hanson, Gradient-based iterative image reconstruction scheme for time-resolved optical tomography, IEEE Trans. Medical Imaging 18 (1999) 262–271.
[21] M. Firbank, S.R. Arridge, M. Schweiger, D.T. Delpy, An investigation of light transport through scattering bodies with non-scattering regions, Phys. Med. Biol. 41 (1996) 767–783.
[22] E. Okada, D.T. Delpy, Near-infrared light propagation in an adult head model. I. Modeling of low-level scattering in the cerebrospinal fluid layer, Appl. Opt. 42 (2003) 2906–2914.
[23] V. Ntziachristos, B. Chance, Accuracy limits in the determination of absolute optical properties using time-resolved NIR spectroscopy, Med. Phys. 28 (2001) 1115–1124.
[24] V. Ntziachristos, A.H. Hielscher, A.G. Yodh, B. Chance, Diffuse optical tomography of highly heterogeneous media, IEEE Trans. Med. Imaging 20 (2001) 470–478.
[25] R.C. Haskell, L.O. Svaasand, T.-T. Tsay, T.-C. Feng, M.S. McAdams, B.J. Tromberg, Boundary conditions for the diffusion equation in radiative transfer, J. Opt. Soc. Am. A 11 (1994) 2727–2741.
[26] R.J. LeVeque, Finite Difference Methods for Differential Equations, notes available at http://www.amath.washington.edu/~rjl/booksnotes.html.
[27] R.D. Richtmeyer, K.W. Morton, Difference Methods for Initial-Value Problems, John Wiley & Sons, NY, 1967.
[28] S. Takatani, M.D. Graham, Theoretical analysis of diffuse reflectance from a two-layer tissue model, IEEE Trans. Biomed. Eng. 26 (1979) 656–664.
[29] J.M. Schmitt, G.X. Zhou, E.C. Walker, R.T. Wall, Multilayer model of photon diffusion in skin, J. Opt. Soc. Am. A 7 (1990) 2141–2153.
[30] I. Dayan, S. Havlin, G.H. Weiss, Photon migration in a two-layer turbid medium. A diffusion analysis, J. Mod. Opt. 39 (1992) 1567–1582.
[31] A. Kienle, M.S. Patterson, Improved solutions of the steady-state and the time-resolved diffusion equations for reflectance from a semi-infinite turbid medium, J. Opt. Soc. Am. A 14 (1997) 246–254.
[32] A. Kienle, M.S. Patterson, N. Dögnitz, R. Bays, G. Wagnières, H. van den Bergh, Noninvasive determination of the optical properties of two-layered media, Appl. Opt. 37 (1998) 779–791.
[33] J.-M. Tualle, J. Prat, E. Tinet, S. Avrillier, Real-space Green's function calculation for the solution of the diffusion equation in stratified turbid media, J. Opt. Soc. Am. A 17 (2000) 2046–2055.
[34] F. Martelli, A. Sassaroli, Y. Yamada, G. Zaccanti, Analytical approximate solutions of the time-domain diffusion equation in layered slabs, J. Opt. Soc. Am. A 19 (2002) 71–80.
[35] F. Martelli, A. Sassaroli, S. Del Bianco, Y. Yamada, G. Zaccanti, Solution of the time-dependent diffusion equation for layered diffusive media by the eigenfunction method, Phys. Rev. E 67 (2003) 056623.
[36] F. Martelli, S. Del Bianco, G. Zaccanti, Procedure for retrieving the optical properties of a two-layered medium from time-resolved reflectance measurements, Opt. Lett. 28 (2003) 1236–1238.
[37] F. Martelli, A. Sassaroli, G. Zaccanti, Y. Yamada, Properties of the light emerging from a diffusive medium: angular dependence and flux at the external boundary, Phys. Med. Biol. 44 (1999) 1257–1275.
[38] M.S. Patterson, B. Chance, B.C. Wilson, Time-resolved reflectance and transmittance for the non-invasive measurement of tissue optical properties, Appl. Opt. 28 (1989) 2331–2336.

[39] K.M. Yoo, F. Liu, R.R. Alfano, When does the diffusion approximation fail to describe photon transport in random media?, Phys. Rev. Lett. 64 (1990) 2647.
[40] G. Bal, Transport through diffusive and non-diffusive regions, embedded objects, and clear layers, SIAM J. Appl. Math. 62 (5) (2002) 1677–1697.
[41] G.J. Tearney, M.E. Brezinski, J.F. Southern, B.E. Bouma, M.R. Hee, J.G. Fujimoto, Determination of the refractive index of highly scattering human tissue by optical coherence tomography, Opt. Lett. 20 (1995) 2258–2260.
[42] F. Bevilacqua, D. Piguet, P. Marquet, J.D. Gross, B.J. Tromberg, C. Depeursinge, In vivo local determination of tissue optical properties: applications to human brain, Appl. Opt. 38 (1999) 4939–4950.
[43] In this work, we ignore complications arising from convolution by the instrument response function (IRF). We have found that even if IRF convolution is included in the forward model, this does not affect our choice of t_{min}.
[44] We exclude the in vitro white matter value of $\mu'_s = 9.1\ \mathrm{mm}^{-1}$ used by Okada [22] since it is a whole order of magnitude larger than any in vivo literature value. Disagreements remain about in vivo baseline properties of the scalp, skull, grey and white matter [16,42,14,22].
[45] This simple picture is complicated by the fact that time-correlation in the forward signals I(R,t) can effectively average statistically-independent measurement errors, forcing us to make the forward model more accurate. Ultimately we plan an error analysis of the full (linearized) inverse problem in order to determine what types and sizes of forward errors have negligible effect on answers to the inverse problem.
[46] J.D. Jackson, Classical Electrodynamics, Wiley, New York, 1975.
[47] L. Greengard, J.A. Strain, A fast algorithm for evaluating heat potentials, Comm. Pure Appl. Math XLIII (1990) 949–963.
[48] L. Greengard, P. Lin, Spectral approximation of the free-space heat kernel, Appl. Computat. Harmonic Anal. 9 (2000) 83–97.
[49] J.R. Kuttler, V.G. Sigillito, Eigenvalues of the laplacian in two dimensions, SIAM Rev. 26 (1984) 163–193.
[50] LAPACK, available at http://www.netlib.org/lapack.
[51] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C: The Art of Scientific Computing (Cambridge University Press, 1992), p. 848, available at http://lib-www.lanl.gov/numerical/bookcpdf.html.
[52] J.-M. Tualle, private communication.

Robust inference of baseline optical properties of the human head with three-dimensional segmentation from magnetic resonance imaging

Alex H. Barnett, Joseph P. Culver, A. Gregory Sorensen, Anders Dale, and David A. Boas

We model the capability of a small (6-optode) time-resolved diffuse optical tomography (DOT) system to infer baseline absorption and reduced scattering coefficients of the tissues of the human head (scalp, skull, and brain). Our heterogeneous three-dimensional diffusion forward model uses tissue geometry from segmented magnetic resonance (MR) data. Handling the inverse problem by use of Bayesian inference and introducing a realistic noise model, we predict coefficient error bars in terms of detected photon number and assumed model error. We demonstrate the large improvement that a MR-segmented model can provide: 2–10% error in brain coefficients (for 2×10^6 photons, 5% model error). We sample from the exact posterior and show robustness to numerical model error. This opens up the possibility of simultaneous DOT and MR for quantitative cortically constrained functional neuroimaging.
© 2003 Optical Society of America
OCIS codes: 000.5490, 100.3190, 170.3010, 170.6920, 170.7050.

1. Introduction

Diffuse optical tomography (DOT)[1] is arousing growing interest as a noninvasive tool for neuroimaging, both clinical[2] and functional,[3,4] as well as other clinical applications such as breast tumor detection,[5–8] tracking muscular oxygenation,[9] and arthritic joint imaging.[10] By use of spectroscopy, this modality can couple directly to hemodynamic quantities of interest, measuring both total hemoglobin concentration and oxygenation. Optical contrast can also arise from cell-based mechanisms such as the redox state of cytochrome-*c*-oxidase and *in vivo* contrast agents such as fluorescent and voltage-sensitive dyes.[4] By relying on photon transport through tissue, DOT also accesses spatial information to a depth of several centimeters. DOT shows great promise for functional neuroimaging: The advantages of functional DOT (fDOT) include fast temporal resolution of the order of 10 ms and moderate spatial resolution of the order of 1–2 cm, placing it in these two categories somewhere between both MEG (magnetoencephalography) and EEG (electroencephalography), and functional magnetic resonance imaging. In addition to sensitivity to absorbing chromophores, DOT can be sensitive to cellular scattering changes during neuronal activation.[4,11,12] DOT is noninvasive and nonionizing. The apparatus is relatively inexpensive, is compatible with (and complementary to) other modalities allowing simultaneous imaging,[13,14] is portable, is capable of continuous monitoring, and does not require an immobile subject.

The key challenge of DOT is that of extracting spatial maps of the optical properties (absorption coefficient μ_a and reduced scattering coefficient μ_s') within a highly scattering tissue volume by coupling multiple light sources and multiple detectors to the surface of the skin. The available signals comprise light intensities at each detector that are due to each source. For each of these source–detector pairs, it is possible to measure a dc intensity (continuous-wave systems), or intensity amplitude and phase (rf-modulated systems), or obtain the intensity distribution as a function of time of flight (time-domain systems). Because tissue dimensions of interest are much larger than the photon mean free path, the path taken by photons from source to detector is dif-

The authors are with the Nuclear Magnetic Resonance Center, Massachusetts General Hospital, Harvard Medical School, Charlestown, Massachusetts 02129. The e-mail address for A. H. Barnett is barnett@cims.nyu.edu.

Received 7 October 2002; revised manuscript received 10 January 2003.
0003-6935/03/163095-14$15.00/0
© 2003 Optical Society of America

fusive rather than straight; this is what limits the spatial resolution of DOT. Recovery of $\mu_a(\mathbf{r})$ and $\mu_s'(\mathbf{r})$ from measured signals requires the solving of an inverse problem, nonlinear in the optical parameters, and known to be ill-posed.[15,16] Usually $\mu_a(\mathbf{r})$ is measured at several wavelengths within the near-infrared low-tissue-absorption window of 650–900 nm. From this, the known spectral absorption curves[17] of the chromophores HbR and HbO_2 allow spatial maps of [HbR] and [HbO_2] (and therefore of total hemoglobin concentration) to be constructed.

Our goal in this paper is to explore the limits of accuracy in the measurement of baseline optical properties of the human head using a time-domain apparatus when high-resolution anatomical information from magnetic resonance (MR) is used to constrain the optical model. Absolute cerebral oximetry and blood volume measurement, which require baseline (absolute) quantification of cortical μ_a, are important for the study of stroke,[18] head trauma, migraine, neonatal ischemia,[19] and brain development.[20] Baseline properties are important for another reason. For DOT it is common to use a perturbation model where signal changes are taken to be a linear function of optical parameter changes. This allows rapid imaging by matrix inverse methods.[15,21–23] However, one cannot[24] compute the required sensitivity functions without knowing both baseline (absolute) μ_a and μ_s'. We envisage using simultaneous MR and DOT to perform accurate baseline measurements, followed by cortically constrained fDOT imaging and simultaneous and BOLD (blood oxygenation level dependent) fMRI.[3,14]

In our model the scalp, skull, brain, and cerebrospinal fluid (CSF) are assigned separate optical properties, assumed uniform within each tissue type, and their three-dimensional (3-D) geometry is taken from automatically segmented MR images. This greatly reduces N the number of unknowns (we use $N = 6$, compared with $N \sim 10^2$–10^4 that is common in pixel- or voxel-based representations), making the inverse problem much more tractable. This idea of MR-constrained reconstruction is not new to DOT,[25] certainly in two dimensions,[26,27] nor to other medical inverse problems.[28,29] Note that our approach differs from the class of segmentation methods in which tissue boundaries are themselves parameterized by unknowns.[30,31]

Time-resolved intensity measurement is popular to extract baseline properties in simple homogeneous[32–34] and layered[35–38] slab tissue models, in both transmission and reflection geometries. Time-of-flight information resolves otherwise indistinguishable effects of absorption and scattering changes.[15] For example, Kienle and Glanzmann[39] use a two-layer analytic diffusion model to fit for optical coefficients of *in vivo* arm muscle, with the superficial fat layer thickness constrained by another imaging modality (in this case ultrasound). For the adult head, *in vivo* measurements also have usually been analyzed with simple analytical diffusion models, even though the internal geometry may be poorly represented as a slab or layered slab.

To overcome these limitations, we use a numerical forward model that can handle arbitrary 3-D tissue geometries and optode locations in the diffusion approximation. We use this to perform model-based fitting by nonlinear optimization.[15,40,41] We use the Bayesian paradigm for inference, giving us the full posterior probability density function (PDF) of the parameters (which in turn could be used to obtain the full PDF of baseline [HbR] and [HbO_2]). Bayesian inference has been used with success in related medical inverse problems.[29,42–44] However, fully Bayesian[41,45,46] or Bayes-influenced[47] applications in DOT are less common. In contrast to most previous research, we use the posterior PDF to study expected errors (i.e., error bars) and the detailed distribution in the parameters rather than presenting a single best solution. We also introduce a novel realistic noise model that captures our belief in signal reliability at both small and large numbers of detected photons. This accounts for photon detection noise and forward modeling errors; the latter have been ignored in the Bayesian DOT literature thus far.

This paper is organized as follows. In Section 2 we outline the application of Bayesian inference to our problem and introduce the general form of the likelihood function. In Section 3 we present the head system and the forward and noise models used for inference and simulation of noisy experimental signals. In Section 4 we give results showing the accuracy of inference achievable using a homogeneous head model and a segmented head model. We present achievable error bars as a function of total collected photons and assumed model errors. We investigate in detail the posterior distribution in the optical parameters using a PDF sampling method and validate our choice of posterior approximation. We also demonstrate robustness to forward model errors. We conclude and suggest future directions in Section 5. Two appendices contain methodological details that would otherwise impede the flow of the main text.

2. Inference Framework

The Bayesian approach treats an inverse problem as an inference problem: Our lack of knowledge about model parameters is represented by a PDF over those parameters. A model represented by $\mathcal{H}$ (for hypothesis) contains unknown parameters given by the vector $\mathbf{x} \equiv \{x_1 \ldots x_N\}$ and includes a prior PDF on those parameters $p(\mathbf{x}|\mathcal{H})$. We consider how this PDF is modified by the arrival of the experimental data (measured signal) vector $\mathbf{y} \equiv \{y_1 \ldots y_M\}$ to give a posterior PDF:

$$p(\mathbf{x}|\mathbf{y}, \mathcal{H}) \propto p(\mathbf{y}|\mathbf{x}, \mathcal{H}) p(\mathbf{x}|\mathcal{H}), \tag{1}$$

where $p(\mathbf{y}|\mathbf{x}, \mathcal{H})$ is the likelihood defined by model $\mathcal{H}$, and the constant of proportionality depends only on $\mathbf{y}$. The posterior encodes everything we now know about $\mathbf{x}$. When we are not distinguishing between different models, we drop the conditionality on $\mathcal{H}$ in our notation.

The relevant advantages of this framework include the following: (1) Explicit formulation of all assumptions that are present in all approaches to inverse problems but are not always expressed. (2) The ill-posed nature of the inverse problem, implying uncertainty in some parameter directions, is embraced and handled probabilistically. This contrasts with traditional approaches where this is viewed as 'instability' and is therefore removed by *ad hoc* regularization methods. (3) The posterior PDF contains all confidence intervals and error correlations of parameters. In the case of hemoglobin concentration parameters, their posterior could be used to infer neural activation (with fitting models of neural-hemodynamic coupling[48]) with more reliability than best-fit parameter values alone. (4) Probabilistic prior information, for example, from another imaging method, can be consistently made use of, allowing optimal use of multimodal imaging information. (5) The Bayesian method is optimal in the sense of giving unbiased minimum-variance estimators when the likelihood and prior are correct. (6) Complex models (with many additional, or hidden, parameters such as amplitude calibrations[49]) can be handled without the danger of overfitting that can occur with simple cost-function minimization. (7) Competing models $\mathcal{H}_i$ can be ranked by their ability to explain the data, even if they have different structure and numbers of parameters. Offsetting these advantages is the main obstacle to use of Bayesian methods: They can be computationally intensive compared with more *ad hoc* approaches. For an introduction, see the informal tutorial by MacKay[50] or other more detailed reviews.[51–55]

We take a forward model $\mathbf{f}(\mathbf{x})$ that encapsulates our physical model of the DOT system. Given an optical parameter vector $\mathbf{x}$, it returns the expected signal vector $\mathbf{f} \equiv \{f_1 \ldots f_M\}$. We also need a noise model, giving the probability that a signal $\mathbf{y}$ could be generated by the addition of noise to the noise-free signal $\mathbf{f}$. We use an independent Gaussian noise model, giving the likelihood function

$$p(\mathbf{y}|\mathbf{x}) \propto \exp[-L(\mathbf{y};\, \mathbf{x})], \tag{2}$$

with the negative log likelihood

$$L(\mathbf{y};\, \mathbf{x}) = \frac{1}{2}\sum_{m=1}^{M} \ln \sigma_m{}^2 + \frac{1}{2}\sum_{m=1}^{M} \frac{[f_m(\mathbf{x}) - y_m]^2}{\sigma_m{}^2}. \tag{3}$$

Each noise standard deviation $\sigma_m \equiv \sigma[f_m(\mathbf{x})]$ is given by the same function applied to the corresponding signal level $f_m(\mathbf{x})$. The form of $\sigma(f)$ is discussed in Subsection 3.D. The second term is analogous to a χ^2 cost function. The constant of proportionality and the log term arise from the normalization requirement $\int d^M\mathbf{y} p(\mathbf{y}|\mathbf{x}) = 1$. Our inference procedure is shown schematically in Fig. 1.

For this study we used a constant prior $p(\mathbf{x})$ within certain biologically motivated bounds on each parameter $x_n \in [x_{n,\min}, x_{n,\max}]$ and zero outside (in other words, a standard minimum–maximum prior). This reflects a lack of preference over $\mathbf{x}$ within these bounds and zero belief in values outside the bounds. Therefore the posterior is

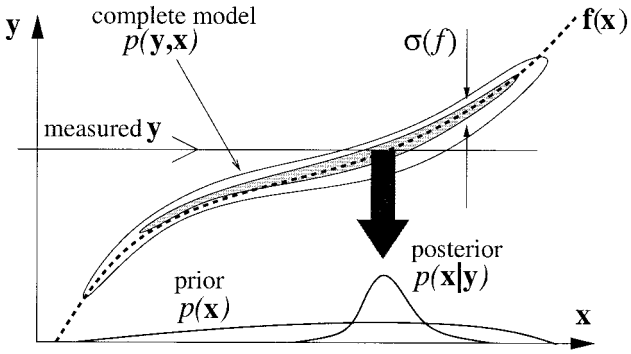


Fig. 1. Illustration of Bayesian inference of unknown parameters $\mathbf{x}$ given the measured signal vector $\mathbf{y}$. The complete model comprises a forward model $\mathbf{f}(\mathbf{x})$ (dotted curve) and an inference noise model, giving the joint PDF $p(\mathbf{y}, \mathbf{x})$ that can be written as a prior $p(\mathbf{x})$ multiplied by a likelihood $p(\mathbf{y}|\mathbf{x})$. The inference noise model describes all measurement and model errors; here we use independent Gaussian noise with signal-dependent width $\sigma(f)$.

$$p(\mathbf{x}|\mathbf{y}) \propto \begin{cases} \exp[-L(\mathbf{y};\, \mathbf{x})] & x_n \in (x_{n,\min}, x_{n,\max})\ \forall\, n \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where we do not need to know the constant of proportionality. It would be simple to encode more specific prior beliefs than this simple constraint.

For N more that 2 or 3, the high dimensionality and the fact that each evaluation of $\mathbf{f}(\mathbf{x})$ requires a time-consuming solution of a PDF make it impossible in general to characterize the posterior over all $\mathbf{x}$. However, for differentiable $\mathbf{f}(\mathbf{x})$, there will always be a sufficiently small choice of $\sigma(f)$ that the bulk of the posterior probability mass will be close to the maximum *a posteriori* (MAP) value $\mathbf{x}_{\mathrm{MAP}}$ and will be well approximated by a multivariate Gaussian with covariance matrix Σ_{MAP} and mean $\mathbf{x}_{\mathrm{MAP}}$.[52] The inference task is then to locate $\mathbf{x}_{\mathrm{MAP}}$ and measure Σ_{MAP}, from which confidence intervals on each x_n can be computed. In all but Subsection 4.D we use this approach; details are presented in Appendix A.

We note that this search for $\mathbf{x}_{\mathrm{MAP}}$ is analogous to model-based approaches where an objective function (cost function) $L(\mathbf{y};\, \mathbf{x})$, in our case differing from the weighted least-squares form only by an additional log term, is minimized by an iterative approach.[15,56] Indeed much of the machinery is in common. However, the interpretation is different: Those who use model-based approaches are generally interested in the single best solution[41,57,58] rather than properties of the full posterior PDF. (Often this restriction is a practical one because of large N.)

When the posterior is significant over a region where $\mathbf{f}(\mathbf{x})$ is no longer linearizable, the Gaussian approximation becomes bad. In this case we use Markov chain Monte Carlo (MCMC) sampling[55] to

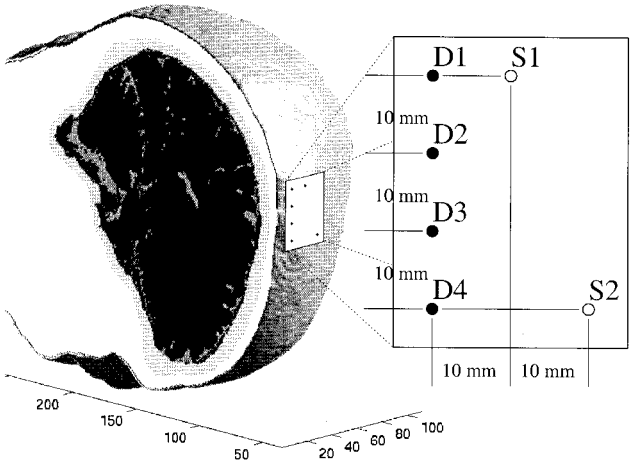


Fig. 2. Simulated optode arrangement and placement on the head (the face, pointing downward, is mostly hidden). 3-D MR-segmented head geometry is exposed by a sagittal slice. The tissue-type color coding on this slice is, from lightest to darkest, scalp, skull, CSF, brain.

generate a set of uncorrelated samples from the exact posterior, providing a (noisy) cloudlike impression of the full PDF. We demonstrate this more general but computationally intensive method in Subsection 4.D.

3. Forward and Noise Model

A. Head System

We start with a human head segmentation geometry obtained at a 1 mm by 1 mm by 1.5 mm voxel resolution from a MR structural image using the technique of Dale *et al.*[59] This data consist of an assignment of each voxel to one of the five categories: scalp, skull, CSF, brain, or not tissue (air). (We did not distinguish between gray and white matter; gray matter will dominate the optical properties because of the rapid decay of sensitivity with depth.) For the subject data used, the scalp and skull thicknesses were typical, being 5–7 and 7–8 mm, respectively, in the relevant region of the head. Figure 2 shows our head geometry. Unless otherwise stated, we use as our standard optical properties for the four tissue types the values given in Table 1. These are believed to be typical, falling within the quite wide variation of published values.[33,60,61]

The CSF falls into a special category. It is presumed to be approximately 10^2 times less scattering and less absorbing than other tissue types and occupies a much smaller volume. Its structure is convoluted and uneven, varying from a thin layer of approximately 1–2 mm thickness hugging the dura, to folds and pockets of several millimeters in thickness and size, following the folds of the brain surface.[62] Once the CSF geometry is held fixed, realistic biological variation in either CSF μ_a or μ_s' causes negligible changes in the photon tranport, and hence detected signals. For μ_a, this is because the values and volume are too small to cause any significant absorption of light compared with that caused by surrounding tissues.

For μ_s', the reason is subtly different: The extremely long reduced scattering length of $1/\mu_s' \approx 100$ mm is much longer (by a factor of the order of 5) than maximal line-of-sight distances within the CSF void region. This is mainly due to the folded and irregular geometry and is most likely enhanced by surface roughness[63] and the presence of vasculature in the CSF below the MR resolution. Thus typical free transport distances in the CSF are dominated by the length scale of the irregular folds and vessels. We believe that photons in the CSF pass into other highly scattering tissue types long before the particular value of μ_s' becomes relevant.

Therefore we do not try to infer CSF properties in our segmented model, and our unknowns will consist of μ_a and μ_s' for the remaining three tissues. We compare two models:

- $\mathcal{H}_{\mathrm{hom}}$: The interior of the head is assumed homogeneous, with $N = 2$ unknown parameters $\mu_{a,\mathrm{hom}}$ and $\mu_{s,\mathrm{hom}}'$ describing a single tissue type.
- $\mathcal{H}_{\mathrm{seg}}$: The interior of the head is segmented as described above, with $N = 6$ unknown parameters $\mu_{a,\mathrm{scalp}}$, $\mu_{s,\mathrm{scalp}}'$, $\mu_{a,\mathrm{skull}}$, $\mu_{s,\mathrm{skull}}'$, $\mu_{a,\mathrm{brain}}$, and $\mu_{s,\mathrm{brain}}'$. Note that the CSF is present in this model but has fixed parameters.

The unknown (or fitting) parameters were given bounds $[x_{n,\min}, x_{n,\max}]$ of [0.004, 0.04] mm^{-1} for all absorption and [0.4, 4.0] mm^{-1} for all reduced scattering coefficients.

B. Optodes and Detection Model

We chose to model a two-source ($N_s = 2$) and four-detector ($N_d = 4$) fiber-coupled time-resolved system contacting a small region of the scalp at the top of the head. The arrangement, shown in Fig. 2, was chosen to cover a range of source–detector distances of 10–36 mm. We assume that the system counts photons and bins these counts according to their arrival time, building up a histogram of the temporal point-spread function (TPSF). We assume no convolution of the TPSF because of the measurement system (i.e., instrument response function), but this would be easy to include in the forward model.

For simplicity, we assume that optode locations are known accurately and that the system has been calibrated to provide absolute measurements of amplitude and of time of flight. (We do not model the

Table 1. Standard Set of Optical Properties of Human Head Tissue Types Used in this Study[a]

Tissue	μ_a (mm^{-1})	μ_a' (mm^{-1})
Scalp	0.0149	0.8
Skull	0.01	1.0
CSF	0.0004	0.01
Brain	0.0178	1.25

[a]The tissue type brain includes gray and white matter but is optically dominated by gray matter.

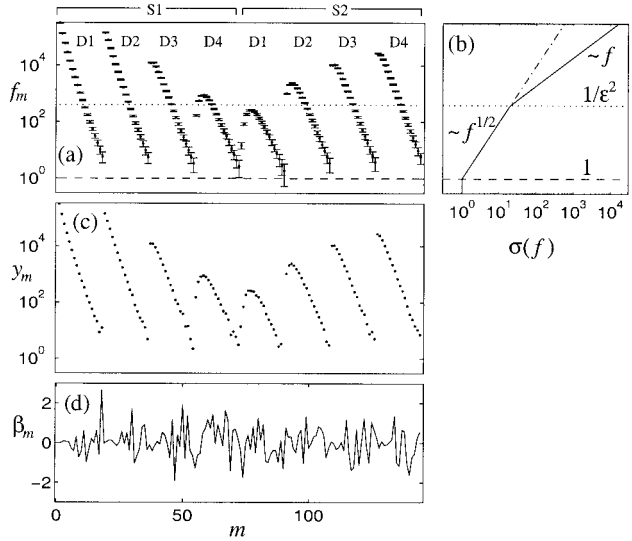


Fig. 3. Signals and noise models. (a) Typical signal expectation vector $\mathbf{f} \equiv \{f_m\}$ for $m = 1 \ldots M$, corresponding to $N_p = 1.1 \times 10^6$ detected photons, with error bars representing the noise standard deviation σ used for inference. The eight time courses correspond to the interval 0.2–2.0 ns and are labeled according to source and detector number. The 1 photon level is shown by the dashed line; the level for crossover to fractional-error-dominated noise at $1/\epsilon^2$ is shown by the dotted line for $\epsilon = 0.05$. (b) The solid line is the inference noise $\sigma(f)$ given in Eq. (6), and the dashed–dotted line is the simulated experimental noise $\sigma_{\mathrm{sim}}(f)$ given in Eq. (7), plotted horizontally to share the same vertical scale as (a). (c) A simulated noisy signal vector $\mathbf{y}$ generated by the addition of Gaussian noise of size $\sigma_{\mathrm{sim}}(f)$ to the expectation $\mathbf{f}$ on same horizontal axis as (a). (d) The residual β_m [in standard deviation units, see Eq. A(3)] that would result from the $\mathbf{y}$ and $\mathbf{f}$ shown on same horizontal axis as (a).

variable attenuation that is often used[7,64] to prevent detector saturation at small optode separations.)

One TPSF is available for each of the $N_s \times N_d = 8$ source–detector pairs. We represent each TPSF by 18 numbers, giving the photons collected in each of the 100-ps-wide time gates in the time-of-flight range 0.2–2.0 ns. We found that for time of flights below 0.2 ns, diffusion forward model errors were too large to be useful.[65] Beyond 2.0 ns, the signals generally fall below the 1-photon level. An example simulated signal vector $\mathbf{y}$ is shown in Fig. 3. Note that our signals are presented in units of detected photons.

C. Diffusion Forward Model

Photon transport in scattering media at macroscopic scales, where interference effects can be ignored, is described by the transport equation.[15,66] In media where $\mu_a \ll \mu_s'$ and length scales of interest are much larger than $1/\mu_s'$, transport can be well described[15] by the diffusion approximation (DA):

$$\frac{1}{\nu}\frac{\partial}{\partial t}\phi = \nabla \cdot (\kappa \nabla \phi) - \mu_a \phi + q, \tag{5}$$

where $\phi(\mathbf{r}, t)$ is the fluence, $q(\mathbf{r}, t)$ is the source term, $\kappa(\mathbf{r})$ is the local diffusion constant, $\mu(\mathbf{r})$ is the local absorption, and $\nu(\mathbf{r})$ is the local speed of light in the medium. We use the good approximation[67,68] $\kappa = 1/(3\mu_s')$ and assume $\nu(\mathbf{r}) = c$ everywhere corresponding to unity refractive index. Our segmented map of μ_s' is given when we assign $\mu_s'(\mathbf{r}) = \mu_{s,\mathrm{tissue}(\mathbf{r})}'$ where tissue($\mathbf{r}$) represents the tissue type of the voxel within which $\mathbf{r}$ falls. Note that the resulting discontinuities in optical properties can cause discontinuities in $\nabla\phi$, but not in the value of ϕ itself (this would not be the case had we allowed refractive-index changes). For tissue–air boundary conditions we found the Dirichlet ($\phi = 0$) approximation to be sufficiently close to the more accurate Robin boundary condition[15] because κ is much smaller than the thicknesses of interest.

We use a finite-difference time-domain (FDTD)[69] method to compute the evolution of Eq. (5), from which the signal expectation vector $\mathbf{f}(\mathbf{x})$ is extracted by use of the detector locations. Details are given in Appendix B. The method represents the smooth function $\phi(\mathbf{r}, t)$ as discrete values on a cubical lattice of nodes, at a sequence of time steps. We resample the segmentation voxel map to the required lattice resolution, in this study either 2 or 1 mm. For a 2-mm voxel size, with 3.3×10^4 nodes in the required head volume, our current FDTD method takes approximately 8 s/source (on a 1-GHz CPU) to simulate 2 ns of propagation time. We did not optimize the forward model, and there exist refinements of FDTD that can be much faster (see Appendix B). Typical error sizes at 2 mm can be judged from Fig. 9 (keep in mind that these signals span close to 6 orders of magnitude).

A note is necessary to explain how we treat the CSF. In recent years, research has been done that shows the large fluence errors that can result when the extremely low μ_s' value in voidlike regions is fed directly into the DA.[67] (These errors were measured by comparison with transport equation solutions.) We chose a different tactic, similar to that of Ripoll *et al.*[63] By giving the CSF an effective μ_s' for use within the DA, we were able to approximate the physics much better than possible using the true μ_s'. As discussed in Subsection 3.A, line-of-sight distances in the CSF are small, we believe of the order of $l \sim 10$ mm because of highly irregular geometry and vasculature. We believe that the optimal DA choice (which may vary from subject to subject) is a μ_s' of the order of l^{-1}. Our preliminary results suggest that the fluence field is not sensitive to the exact μ_s' chosen when the full 3-D MR head geometry is modeled. In contrast, most previous comparisons have used idealized two-dimensional CSF geometries with long lines of sight[60,67,70] or two-dimensional models taken from a single MR slice.[67] In this paper we fix $\mu_{s,\mathrm{CSF}}' = 0.4\ \mathrm{mm}^{-1}$. The largeness of this choice is in part influenced by numerical efficiency: The CPU time for our current FDTD scales inversely with the smallest μ_s' in the system, which is always $\mu_{s,\mathrm{CSF}}'$ in our case.

D. Noise Model for Inference

The forward model provides the signal expectation value vector $\mathbf{f}(\mathbf{x})$. Bayesian inverse methods always require a noise model, which captures our uncertainties about the difference between the numerically generated $\mathbf{f}(\mathbf{x})$ and the real-world signals $\mathbf{y}$, given that the parameters $\mathbf{x}$ were actually correct.

Recall that in Eq. (3) we assigned Gaussian distributions to each of the components $m = 1 \ldots M$ independently, each with a mean f_m and a standard deviation $\sigma(f_m)$. The Gaussian approximation to the Poisson photon detection statistics[71] would yield $\sigma^2(f) = f$. Remember that signal units are in (effective) photons per time gate. This approximation is good for $f \gg 1$, becoming bad (especially in the tails) as f approaches 1. To handle sensibly the case $f < 1$, we place a lower limit of 1 on σ. (This limit could be increased to account for additive electronic noise.)

However, there is a problem at large signals: If some f_m is, say, 10^6, giving $\sigma_m = 10^3$, then we cannot be expected to believe that the fractional error σ_m/f_m, and therefore our trust in our forward model's closeness to reality, is 0.1%. Rather we choose a fraction ϵ that is a lower limit on our fractional error and takes account of

(1) modeling of physics of photon migration in a given tissue geometry (e.g., transport corrections to the DA);
(2) deviations from the assumption of uniform optical properties within a given tissue type;
(3) incorrect tissue-type assignments from automatic MR segmentation; and
(4) the accuracy with which absolute amplitude (and time offset) calibration can be carried out, including effects of optode–skin coupling variations that are due to, e.g., skin pigmentation.

We typically choose $\epsilon = 0.05$–0.2, corresponding to a 5%–20% model and calibration error.

Combining the above, the noise level $\sigma(f)$ is a piecewise power law:

$$\sigma(f) = \begin{cases} 1, & f \le 1 \\ f^{1/2}, & 1 < f \le 1/\epsilon^2 \\ \epsilon f, & f > 1/\epsilon^2 \end{cases}. \tag{6}$$

This is shown by the solid line in Fig. 3(b). Typically the peaks in f are fractional-error dominated, whereas the tails are Poisson statistics dominated.

E. Simulating Experimental Signals

In this proof-of-concept study, we generate simulated experimental signals using numerical forward models of the same type as used for the inference (inverse) problem. Given a true optical parameter vector $\mathbf{x}^{(0)}$, we generate a noise-free signal expectation vector $\mathbf{f}[\mathbf{x}^{(0)}]$, to which we apply a simulated noise model.

Our simulated noise is identical to our inference noise (as described in Subsection 3.D), except it includes only the detection noise component. We do not explicitly add noise to simulate model error; rather in Subsection 4.E we investigate the effects of model error by changing the forward model used to calculate $\mathbf{f}[\mathbf{x}^{(0)}]$. This yields the simulated noise level

$$\sigma_{\text{sim}}(f) = \begin{cases} 1, & f \le 1 \\ f^{1/2}, & f > 1 \end{cases}. \tag{7}$$

This is shown by the dashed–dotted line in Fig. 3(b). Note that the difference between simulated and inference noise models manifests itself in Fig. 3(d) as a residual β_m [see Eq. (A3)], which has a variance $\ll 1$ for the m values where $f_m \gg 1/\epsilon^2$.

The recipe to simulate experimental signals is

$$y_m = \max[f_m(\mathbf{x}_0) + n_m, 0], \qquad m = 1 \ldots M, \tag{8}$$

where each n_m is independently sampled from a univariate Gaussian distribution with zero mean and variance $\sigma_{\text{sim}}{}^2[f_m(\mathbf{x}_0)]$. The maximum value operation removes unphysical negative signals.

4. Results and Discussion

A. Homogeneous versus Segmented Head Fit

We are interested in finding out what improvement is possible in fitting baseline optical parameters when anatomical segmentation information becomes available. Therefore in this subsection we compare the models $\mathcal{H}_{\text{hom}}$ and $\mathcal{H}_{\text{seg}}$ in their ability to infer these parameters, using simulated data also generated from $\mathcal{H}_{\text{hom}}$ or $\mathcal{H}_{\text{seg}}$. We choose a total detected photon number (which includes both sources)

$$N_p \equiv \sum_{m=1}^{M} y_m \tag{9}$$

at approximately 2×10^6, typical for photon-counting DOT systems. As we vary the head optical properties, we decided to hold the collection time constant rather than fix N_p. Note that some advantage, in terms of smaller N_p values, could be gained if variable detector attenuation[64] were used. (Figure captions 4–7 and 10 give the N_p range for each experiment.) In this subsection we fix $\epsilon = 0.05$ and use a 2-mm forward model lattice. Recall that all simulated signals include realistic detection noise.

1. *Homogeneous–Homogeneous Fit*

Using signals from $\mathcal{H}_{\text{hom}}$, we fit using $\mathcal{H}_{\text{hom}}$ and sweep the true optical parameters. The results are presented in Fig. 4. This shows that if the head were indeed homogeneous, then accurate inference of its baseline optical parameters would be possible. The error bars (representing the expected measurement error of the parameters) are 0.5% or less. The true values fall within, or just outside, the error bars.

2. *Segmented–Homogeneous Fit*

Using signals from the more realistic $\mathcal{H}_{\text{seg}}$, we fit using $\mathcal{H}_{\text{hom}}$ (see Fig. 5). This shows that the homogeneous model does a poor job of measuring the brain's properties: Changes in absorption are only

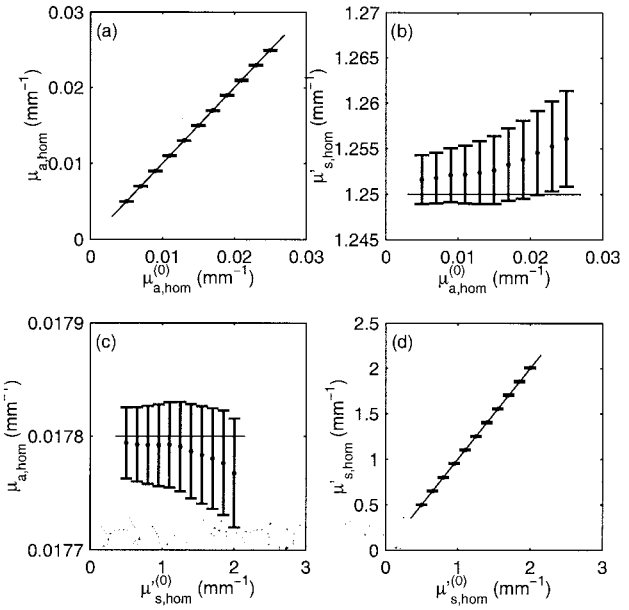


Fig. 4. Tracking of inferred optical parameters of a homogeneous head model by use of noisy signals derived from the same model. (a) and (b) The effect that a change in the true absorption $\mu_{a,\mathrm{hom}}^{(0)}$ has on inferred $\mu_{a,\mathrm{hom}}$ and $\mu_{s,\mathrm{hom}}'$, respectively. The range of total detected photons N_p is from 3.4×10^6 at the smallest $\mu_{a,\mathrm{hom}}^{(0)}$ to 3.6×10^5 at the largest. (c) and (d) Same as (a) and (b) except for a change of the true reduced scattering $\mu_{s,\mathrm{hom}}'^{(0)}$ over which N_p varies from 1.8×10^6 to 3.5×10^5. The error bars show $\pm 1\sigma$ about the MAP (best-guess) value for the (marginal) posterior distribution of inferred values, computed by the methods of Appendix A. For comparison, the thin lines show the true values. The fractional model error is $\epsilon = 0.05$.

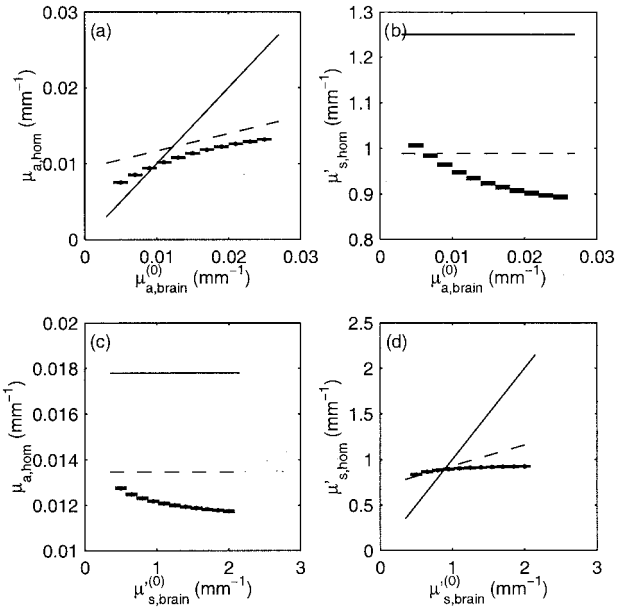


Fig. 5. Same as Fig. 4, except the noisy signals are generated by use of the segmented forward model, sweeping over a range of brain parameters. (The scalp and skull are fixed at the standard parameters given in Table 1.) The inference is still performed by the homogeneous model. The dashed lines show true parameters volume averaged over a depth of 17 mm. (We chose a constant effective thickness of 4 mm for the brain's volume contribution, a typical time-independent fluence decay length in a semi-infinite brain.) N_p spans 2.4×10^6 to 2.1×10^6 over the range of $\mu_{a,\mathrm{brain}}^{(0)}$ and 2.1×10^6 to 2.2×10^6 over the range of $\mu_{s,\mathrm{brain}}'^{(0)}$. The fractional model error is still $\epsilon = 0.05$.

20% of the true change and for scattering are only 5%. This is due to detected photons spending a large fraction of time in the scalp and skull layers. To take this into account we also compare against local volume-averaged properties (dashed lines in Fig. 5); as expected, the homogeneous model reflects these better than the brain's properties. However, there are two major failures of the homogeneous model: cross talk and nonlinearity. Neither of these failures can be compensated for by any (linear) volume-averaging correction. The absorption and scattering cross talk dominates the fitted $\mu_{s,\mathrm{hom}}'$. (There is of course large cross talk with scalp and skull parameters, too, which we do not plot.) The tracking is highly nonlinear, meaning that even relative changes in brain parameters could not be quantitatively assessed. Note that $\mathcal{H}_{\mathrm{hom}}$ is a more physically accurate model than the homogeneous semi-infinite slab model commonly used to fit background properties because it incorporates the correct head surface shape. The fact that the true values are hundreds of standard deviations outside the error bars implies that the inference noise model grossly underrepresents the large true model errors.

3. *Segmented–Segmented Model*

We now fit these same $\mathcal{H}_{\mathrm{seg}}$ signals with the $\mathcal{H}_{\mathrm{seg}}$ model (see Fig. 6). Measurement errors are 2–2.5%

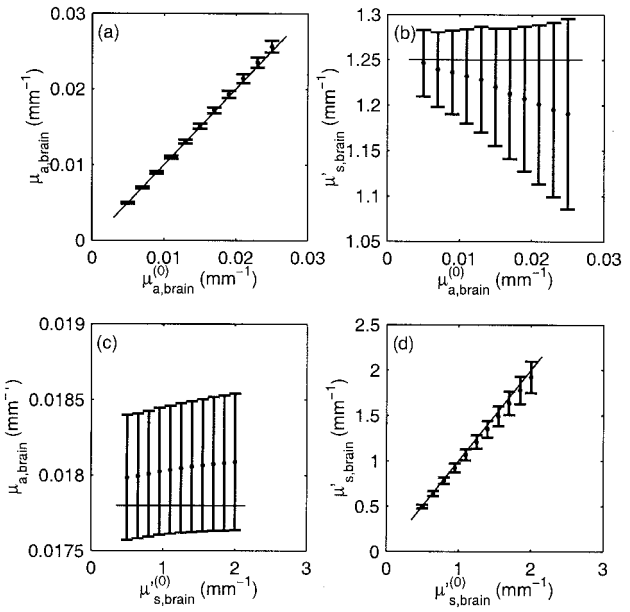


Fig. 6. Tracking of inferred brain optical parameters in a segmented model by use of noisy signals also derived from the segmented model. (The scalp and skull are fixed at the standard parameters given in Table 1, and the inferred scalp and skull parameters are not shown.) N_p and ϵ are the same as for Fig. 5.

in $\mu_{a,\mathrm{brain}}$ and 5–10% in $\mu_{s,\mathrm{brain}}'$ across the complete range of biologically relevant true parameter values. The linearity is excellent, and cross talk causes a 0.5% change in $\mu_{a,\mathrm{brain}}$ and a 5% change in $\mu_{s,\mathrm{brain}}'$ across the parameter range. Importantly, although error bars are larger than with $\mathcal{H}_{\mathrm{hom}}$, true values always fall within the error bars. This shows that we believe reliable measurements of the brain's baseline properties can be made when anatomical information is provided.

Recalling that we are simultaneously fitting for scalp, skull, and brain parameters, we can ask how accurately the scalp and skull have been characterized. At the standard values of Table 1, expected errors are approximately 8% for $\mu_{a,\mathrm{scalp}}$, 3% for $\mu_{s,\mathrm{scalp}}'$, 4% for $\mu_{a,\mathrm{skull}}$, and 2% for $\mu_{s,\mathrm{skull}}'$.

We believe that in this fitting (nonimaging) example, the total photon number N_p will be more important to determine accuracy than the number of sources and detectors. Only when N is increased, as in an imaging application, or when unknown optode calibration parameters are included, do we expect that an increase in the number of sources and detectors will significantly improve accuracy. Note that generally to optimize the optode location and number, and TPSF parametrization, it would be important to know which components m of the signal $\mathbf{y}$ are most important to determine error-bar size for $\mu_{a,\mathrm{brain}}$ and $\mu_{s,\mathrm{brain}}'$. Within the Gaussian posterior approximation, all this information is contained in the singular value decomposition of $\bar{J}$ [defined in Eq. (A6)]. We will present these results in future publications.

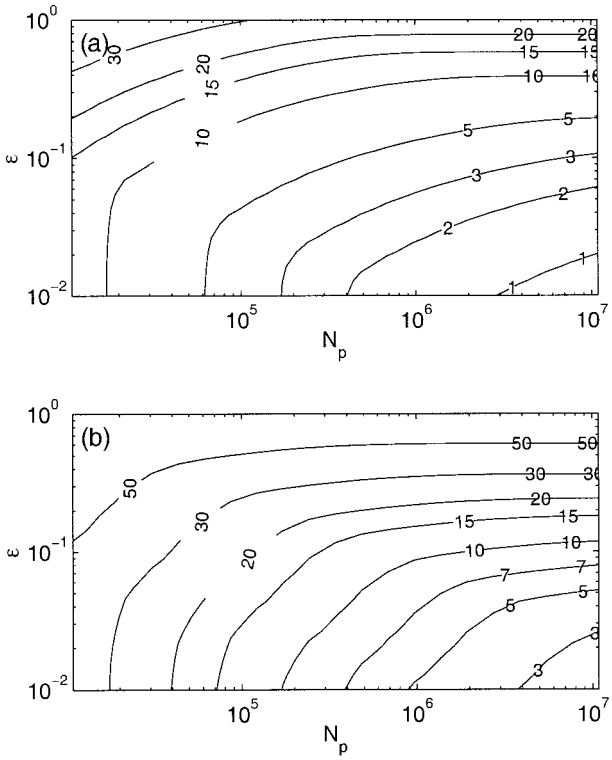


Fig. 7. Expected percentage marginal error as a function of noise model N_p (total detected photons) and ϵ (fractional model error) for parameters (a) $\mu_{a,\mathrm{brain}}$ and (b) $\mu_{s,\mathrm{brain}}'$. The standard set of optical properties are used for inference with $\mathcal{H}_{\mathrm{seg}}$ and simulated noisy signals from the same model. Note that the contour lines bend quite sharply at the transition from Poisson-statistics-limited error (lower left) to model-limited error (upper right).

B. Systematic Deviations

In this subsection we discuss a detail of the inference procedure. (This can be skipped on first reading.)

The reason why there are systematic deviations between MAP and true parameter values in Figs. 4 and 6 is not immediately intuitive. One might suspect that, because exactly the same models are used for simulation as for inference, zero error would result, averaging over the simulated noise. This effect is worth discussing further. In essence the systematic difference comes from the mismatch between the simulated and the inference noise models: $\sigma_{\mathrm{sim}}(f) \neq \sigma(f)$ for large f [see Eqs. (7) and (6)]. If the inference noise model is correct (i.e., reflects the population from which noisy signals are sampled), Bayesian inference provides unbiased estimators[52] (in this case, $\mathbf{x}_{\mathrm{MAP}}$ values). We have checked this: If $\sigma_{\mathrm{sim}}(f) = \sigma(f)$ for all f, then the deviations change from systematic to random. However, model error is not random; we prefer to investigate its effect in particular cases of realistic model error (see Subsection 4.E).

Mathematically, the systematic deviation can be seen to arise from the weak $\mathbf{x}$ dependence of $\sigma[f(\mathbf{x})]$ [giving σ' terms in Eq. (A2)]. It would be misleading to attempt to correct this deviation as this would be tantamount to benefiting from more information about the system than we have formalized in the noise model of Eqs. (3) and (6). The deviation is always of a size consistent with the error bars; the correct way to reduce this deviation is to build more accurate forward models whose smaller known model error we can then use for inference.

C. Noise Model Effect on Brain Parameter Error Bars

It is natural to wonder what the effects of the number of detected photons N_p and the fractional model error ϵ are on the measurement accuracy. This allows the experiment designer to know in advance the detection time needed to reach a certain expected error in brain optical properties. In Fig. 7 we plot contours of expected percentage error in the two brain parameters across the N_p–ϵ plane. We chose the standard set of optical parameters using $\mathcal{H}_{\mathrm{seg}}$ for inference on a noisy signal from the same segmented model. The results show that, even if we assume a fractional model error as large as $\epsilon = 0.2$, expected errors on brain parameters can be 5% for absorption and 15–20% for reduced scattering.

To generate Fig. 7, it would have been a costly operation to find a solution for Σ_{MAP} if we located $\mathbf{x}_{\mathrm{MAP}}$ for each point in the plane. Instead we observe that, because $\mathbf{f}(\mathbf{x})$ is locally linear, its Jacobean J at $\mathbf{x}_{\mathrm{MAP}}$ will be similar to that at $\mathbf{x}^{(0)}$, when $\mathbf{x}_{\mathrm{MAP}}$ is close to $\mathbf{x}^{(0)}$. [In our case, it becomes a bad approximation

only when marginal expected errors are greater than roughly 20%, when nonlinearity of $\mathbf{f}(\mathbf{x})$ becomes relevant.] Therefore we can use J measured once at $\mathbf{x}^{(0)}$ to approximate Σ_{MAP} for any noise model parameters, according to approximation (A5) and Eqs. (A6) and (A7). This procedure makes investigation of the noise model a rapid process.

D. Visualization of Full Posterior Distribution

In Subsection 4.A, for inference with $\mathcal{H}_{\mathrm{seg}}$, the error bars (marginal expected errors) in the six parameters were of the same order of magnitude, 2–10%. However, the hyperellipsoid defining the full posterior is actually narrow in some directions; these directions are not aligned with the parameter axes. The ratio of the largest to smallest eigenvalue of Σ_{MAP} is 3×10^3, implying that the shortest principal axis is only approximately 2% of the longest. The distribution is in fact pancakelike. This phenomenon of both well-constrained and poorly constrained parameter directions is a feature of ill-posed inverse problems.

Combined with the nonlinear nature of $\mathbf{f}(\mathbf{x})$, this gives us some motivation to be suspicious of a Gaussian approximation to the posterior. Figure 8 shows that, even when our posterior is much wider (5% in $\mu_{a,\mathrm{brain}}$ and 14% in $\mu_{s,\mathrm{brain}}'$), the Gaussian approximations to the marginal distributions are adequate and would give adequate confidence intervals on each x_n. However, the non-Gaussian nature has become relevant in the tails of the distribution [e.g., at large $\mu_{s,\mathrm{brain}}'$, Fig. 8(i)]. We generated the samples from the true posterior by MCMC using the Metropolis method[55] with the stepping distribution chosen to be uniform within a hyperellipsoid of shape given by the Hessian matrix [approximation (A5)]. By optimizing the hyperellipsoid size, we could extract one independent sample from the Markov chain roughly every 40 forward model evaluations. The total cloud took 20 h of CPU time to generate. Faster MCMC methods exist.[55]

Note that in Fig. 8 the *conditional* distributions (i.e., slices through the posterior; conditional distributions are much easier to generate than true marginal distributions) do a poor job of telling us the width of the marginal distributions. This results from the pancakelike nature of the PDF.

E. Robustness to Forward Model Error

If we run our forward model at a higher lattice resolution, we can generate simulated signals $\mathbf{y}$ that are more accurate, given the DA physics and the tissue segmentation model. In this way we can see how the inference procedure (based on a lower lattice resolution) handles realistic model error, which is highly correlated. In other words, we avoid committing the inverse crime of using identical forward and inverse models, of which we were guilty of in Subsection 4.A.

In Fig. 9(a) we show that a change from a 2-mm to a 1-mm lattice voxel size in the forward model causes changes of roughly 20%. Most are smaller, and some are much larger, approaching 50% at early times. This latter early-time error is due to the non-

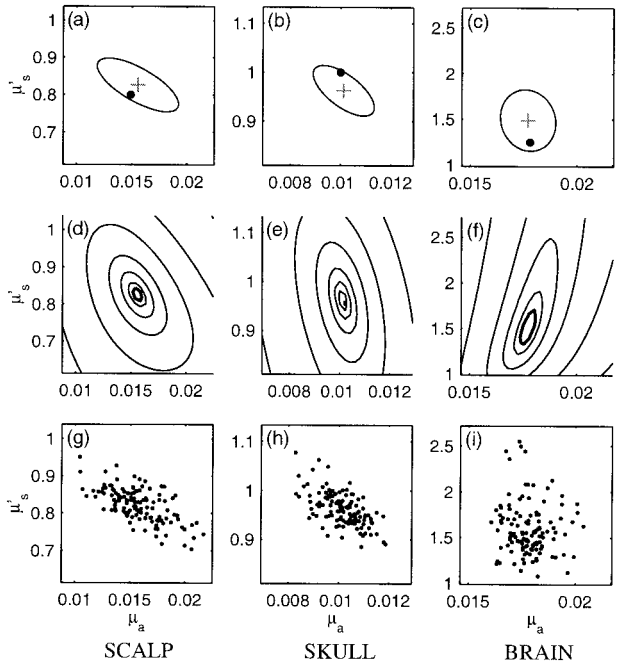


Fig. 8. Views of the posterior distribution $p(\mathbf{x}|\mathbf{y}, \mathcal{H}_{\mathrm{seg}})$ with signals $\mathbf{y}$ from the same forward model (2-mm lattice) by use of our standard optical parameters given in Table 1. Only 3.2×10^5 photons were collected, and $\epsilon = 0.1$. The three columns of graphs show the μ_a–μ_s' plane separately for each tissue type. (a), (b), and (c) show true $\mathbf{x}_0$ (dot), $\mathbf{x}_{\mathrm{MAP}}$ (cross), and the marginal PDF as an elliptical contour enclosing 63% of the probability mass in the Gaussian posterior approximation. The contour is at e^{-1} times the peak density. (d), (e), and (f) show for comparison, on the same axes, conditional distributions (slices through the PDF with other components of $\mathbf{x}$ fixed at the $\mathbf{x}_{\mathrm{MAP}}$ values) at contours of e^{-1} (shown with thicker line), e^{-3}, e^{-10}, e^{-30}, e^{-100}, e^{-300}, and e^{-1000} times the peak density. (g), (h), and (i) show 123 independent samples from the posterior obtained by MCMC. This displays the true marginal posterior PDF as a density cloud.

zero depth below the surface at which the optode couples to the fluence gradient (see Appendix B). The smaller late-time error is due to the voxelated (rough) surface representation local to each optode.

We perform inference using the usual 2-mm lattice model $\mathcal{H}_{\mathrm{seg}}$ and the choice of $\epsilon = 0.2$ in the inference noise model, motivated by the 20% observed model error. We increase the collection time to yield roughly 1.2×10^7 photons. We find [Fig. 9(b)] that the variance of the resulting normalized residuals β_m [see Eq. (A3)] is close to 1, indicating a rough match of the noise model to true forward model errors.

Figure 10 shows the results of sweeping the true brain optical parameters. The accuracy of measurement of $\mu_{a,\mathrm{brain}}$ is approximately 4% and for $\mu_{s,\mathrm{brain}}'$ it is 10–20%. The linearity is good, and maximum cross talk is 6% (only significantly affecting $\mu_{s,\mathrm{brain}}'$). The systematic errors are certainly larger than those achieved in the more artificial situation shown in Fig. 6; however, the error bar still does a good job of indicating the size of this error. Therefore model error is being handled in a robust fashion.

Clearly, incorporating the fact that model errors

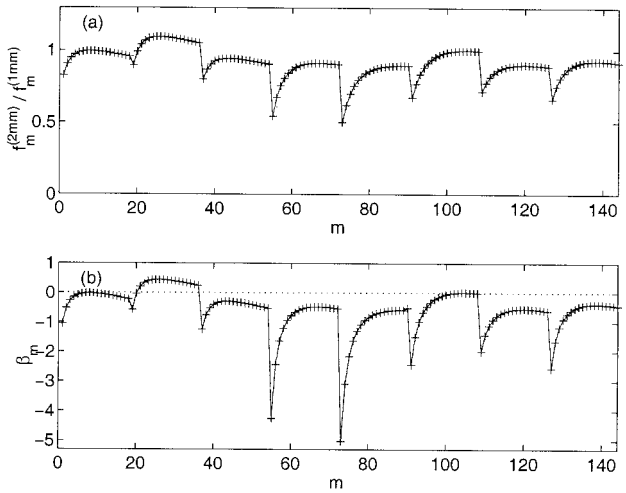


Fig. 9. (a) Approximate error of the 2-mm lattice forward model signal expectation, expressed as a ratio against its (more accurate) 1-mm lattice equivalent. The standard set of optical properties is used for the segmented head model. (b) The normalized residual β_m, which results when we compare the 1-mm lattice signal with the 2-mm lattice signal using noise model parameters $y_p = 10^{-4}$ and $\epsilon = 0.2$.

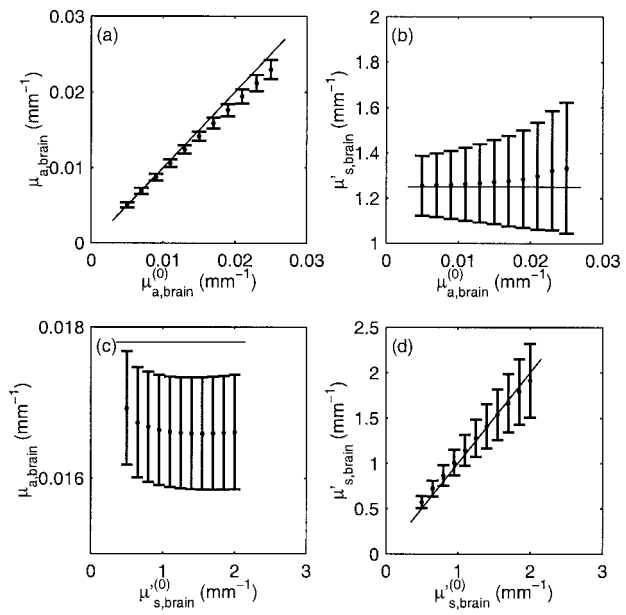


Fig. 10. Inference by use of the 2-mm lattice forward model $\mathcal{H}_{\text{seg}}$ on simulated noisy signals generated from a 1-mm lattice forward model. The detection time was larger than in Fig. 6, with N_p spanning 1.18×10^7 to 1.35×10^7 over the range of $\mu_{a,\text{brain}}^{(0)}$ and $\mu_{s,\text{brain}}'^{(0)}$. The fractional model error of $\epsilon = 0.2$ was chosen to reflect our knowledge of the 2-mm lattice model errors.

are large only at early times would be advantageous. We discuss ideas for model error improvement in Section 5.

5. Conclusion

We have demonstrated that an accurate determination of the baseline optical properties (both absorption μ_a and scattering μ_s') of the human head could be made using a small time-resolved DOT apparatus with high-resolution 3-D anatomical information from MR imaging. We use an automatic MR segmentation into the basic tissue types (scalp, skull, CSF, and brain) and fit for μ_a and μ_s' of scalp, skull, and brain (holding the CSF fixed), assuming each tissue is optically homogeneous. We perform simultaneous nonlinear model-based fitting (inference) of the six optical parameters and demonstrate the superiority of this approach over a homogeneous head model in the biologically relevant parameter ranges $0.005\ \text{mm}^{-1} < \mu_{a,\text{brain}} < 0.025\ \text{mm}^{-1}$ and $0.5\ \text{mm}^{-1} < \mu_{s,\text{brain}}' < 2.0\ \text{mm}^{-1}$. We emphasize that even the homogeneous head model is more realistic than the commonly used homogeneous semi-infinite slab model. For our research we use numerically simulated noisy measured signals. Our finite-difference forward model can simulate the time-domain diffusion approximation in arbitrary 3-D head geometries; we use approximately 3×10^4 nodes. We believe that this approximation is adequate given the convoluted CSF geometry, although this an important area of active research for us and others in the field. Fitting takes approximately 10 min on a standard 1-GHz CPU.

We employ a realistic but novel Gaussian noise model incorporating both shot noise, parametrized by the total detected photons N_p, and model error (including calibration) parametrized by a fraction ϵ. In the Bayesian inference framework, with a flat prior, we find the baseline measurement accuracy (error bars) by examining the marginal width of the posterior distribution for each parameter. We note that all previous demonstrations of Bayesian DOT have either been in two dimensions or simple cuboid geometries and have not considered more than the single best-fit (MAP) solution. We find that 2×10^6 detected photons yield a 2.5% error bar in $\mu_{a,\text{brain}}$ and 5–10% in $\mu_{s,\text{brain}}'$, assuming model errors of 5%. The cross talk between surface layers and the cortex, and between μ_a and μ_s', present with the homogeneous model is eliminated.

By using a finer computational lattice, we tested a realistic forward model error and found that, even with a 20% model error, the error bars can still be adequate: 4% in $\mu_{a,\text{brain}}$ and 10–20% in $\mu_{s,\text{brain}}'$ for $N_p \approx 10^7$. (All N_p values could be reduced experimentally by use of per-detector variable attenuation.) Robustness is demonstrated by the fact that error bars always reflect the size of the true deviations. We have also demonstrated the power of MCMC for posterior sampling when the Gaussian MAP approximation is no longer good.

The Bayesian framework allows, without fuss, use of improved noise models. We believe it will be important to encode into such models both measurement properties of the DOT apparatus and physical and numerical human head forward model errors. Noise models should also provide robustness to measurement outliers by replacing the Gaussian with a heavier-tailed distribution. In future studies we plan to include unknown optode amplitude calibra-

tion parameters,[49] which can then be marginalized over naturally within the Bayesian framework. (We note that Oh *et al.*[45] have recently simulated Bayesian optode calibration in a 3-D cubical geometry, at least in terms of a single MAP solution, with success.) By increasing the number of parameters from 6 to 10^2 or more, our Bayesian approach could give error bars and correlations of voxel parameters in a true imaging (tomography) context, something we intend to address in future research. In general the value of the Bayesian paradigm in neuroimaging lies with its statistical roots in what is becoming an increasingly statistical practice. On a lower level, information from other modalities can optimally be included as a prior. On a higher level, the full posterior PDF, containing all error correlations, could be fed to dynamical models of neural activation.

Our study provides realistic expectations for achievable baseline accuracy in human hemodynamic parameters. An experimental implementation could have rapid clinical benefit in the stroke and neonatal fields. Our research also expands the possibility of simultaneous DOT and MR for functional neuroimaging: Quantitative optical imaging of hemodynamic changes will require accurate baseline properties, which can best be achieved through the type of anatomical modeling presented here.

Appendix A: Finding the Gaussian Approximation to the Posterior

Maximizing relation (4) is equivalent to minimizing $L(\mathbf{y}; \mathbf{x})$ over $\mathbf{x}$ within the given bounds. We rescale the components of $\mathbf{x}$ to have the same typical size by multiplying absorption coefficients by a constant $C = 10^2$, given, for example, in the fully segmented case $\mathbf{x} \equiv \{C\mu_{a,\mathrm{scalp}}, \mu_{s,\mathrm{scalp}}', C\mu_{a,\mathrm{skull}}, C\mu_{a,\mathrm{brain}}, \mu_{s,\mathrm{brain}}'\}$. This elementary preconditioning enhances the performance of the inference methods presented below.

Because of the nature of the numerical solution of the forward model (Appendix B), the dependence of L on $\mathbf{x}$ is not completely smooth or differentiable. This jitter in L, although estimated to be on a scale of less than 10^{-3}, means that care has to be taken in the choice of an optimization algorithm. We found Newton's method[72] most successful, the ith iteration step being

$$\mathbf{x}_{i+1} = \mathbf{x}_i - H^{-1}(\mathbf{x}_i) \cdot \nabla L(\mathbf{x}_i), \tag{A1}$$

where the exact gradient of L has components

$$[\nabla L(\mathbf{x})]_n = \sum_{m=1}^{M} \frac{J_{mn}}{\sigma_m} [\beta_m + (1 - \beta_m{}^2)\sigma_m{}'], \tag{A2}$$

which follows from Eq. (3). Here the normalized residual is

$$\beta_m \equiv \frac{f_m(\mathbf{x}) - y_m}{\sigma_m}, \tag{A3}$$

and the Jacobean (Frechet) derivative of the forward model is $J_{mn} \equiv \partial f_m / \partial x_n$. The symbol $\sigma_m{}'$ is an abbreviation for $\mathrm{d}\sigma/\mathrm{d}f|_{f_m}$, which in our case we can find in closed form by differentiating Eq. (6). All quantities are evaluated at $\mathbf{x}$. Because the number of unknowns N is small, we evaluate J using crude finite differencing along the axes in $\mathbf{x}$ space:

$$J_{mn}(\mathbf{x}) \approx \frac{1}{\delta x} [f_m(\mathbf{x} + \delta x \mathbf{e}_n) - f_m(\mathbf{x})], \tag{A4}$$

where $\mathbf{e}_n$ is the nth unit vector, requiring $N + 1$ forward evaluations. With adjoint differentiation it would be possible to obtain J with an effort roughly equal to one forward evaluation[40,41,57]; this would be worthwhile only if N were larger.

It is simple and efficient to use an approximate Hessian matrix,

$$H \approx H_{\mathrm{approx}} \equiv \bar{J}^T \bar{J}, \tag{A5}$$

where

$$\bar{J}_{mn} \equiv \frac{\partial \beta_m}{\partial x_n} = \frac{J_{mn}}{\sigma_m} (1 - \beta_m \sigma_m{}') \tag{A6}$$

is the Jacobean of the normalized residual. This requires no extra function evaluations to be performed. The stopping criterion for locating $\mathbf{x}_{\mathrm{MAP}}$ is $|\mathbf{x}_{i+1} - \mathbf{x}_i| < \epsilon$ where ϵ is of the order of 10^{-3}. The initial value $\mathbf{x}_1$ is chosen randomly in the optical parameter range of interest and has no effect on the $\mathbf{x}_{\mathrm{MAP}}$ found. (Occasionally if $\mathbf{x}_{\mathrm{MAP}}$ is very far from $\mathbf{x}_1$, then an initial rough Nelder–Mead simplex optimization[72] is required.) Typically 3–8-newton iterations are required. Given that a 2-mm lattice evaluation of $\mathbf{f}(\mathbf{x})$ takes approximately 16 s, this means 5–15 min are required to find $\mathbf{x}_{\mathrm{MAP}}$ and Σ_{MAP}.

Because the derivatives of log terms from Eq. (3) are small, the above form of the Hessian usually agrees to within a few percent with a numerically estimated Hessian found by means of finite differencing $N(N + 1)/2$ samples of L in $\mathbf{x}$ space. We have experimented with procedures for this estimation, but the combination of jitter on L, the fact that L is orders of magnitude more sensitive in some directions than others (i.e., the Hessian is close to singular), and the nonlinearity of $\mathbf{f}(\mathbf{x})$ and hence $L(\mathbf{x})$ causes a tendency for large errors and nonpositive definiteness. Nothing as reliable as the above (guaranteed positive definite) approximate Hessian form has been found. Therefore our multivariate Gaussian covariance matrix estimate is

$$\Sigma_{\mathrm{MAP}} = H_{\mathrm{approx}}{}^{-1}. \tag{A7}$$

To calculate confidence intervals on single parameters, and the lower-dimensional marginal Gaussian distributions shown in Figs. 8(a)–8(c), we need the recipe for marginalizing a multivariate Gaussian onto a subspace. We split $\mathbf{x}$ into subspaces $\mathbf{a}$ and $\mathbf{b}$. By reordering $\{x_n\}$, without loss of generality, this can be written as $\mathbf{x}^T \equiv [\mathbf{a}^T, \mathbf{b}^T]$. Any covariance Σ in $\mathbf{x}$ space can then be written in block form:

$$\Sigma \equiv \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ab}{}^T & \Sigma_{bb} \end{bmatrix}. \tag{A8}$$

By analytically integrating over **b**, the inverse covariance of the marginal distribution in subspace **a** can be shown to be[73]

$$\Sigma_{(a)}{}^{-1} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}{}^{-1}\Sigma_{ab}{}^{T}. \tag{A9}$$

In the case where **a** has a single dimension $a \equiv x_n$, then $\Sigma_{(a)} \equiv \sigma_{x_n}$ is a scalar giving the marginal standard deviation of parameter x_n.

Appendix B: Finite-Difference Method for the Diffusion Equation

Here we provide details of the method used to approximate the time evolution of the 3-D parabolic of Eq. (5) with constant $\nu(\mathbf{r}) = c$. There are many issues, most of which we discuss only cursorily.

Because $q(\mathbf{r}, t)$ is nonzero only at $t = 0$, this is a initial value problem, and we use a FDTD method with a regular cubical lattice of nodes with spacing Δx and a time step Δt. We chose this over finite-element methods for simplicity and the fact that our segmentation map is also based on a cuboid lattice. We chose a node-based representation of fluence ϕ and discretized the spatial derivative at node location $\mathbf{r}_{i,j,k}$ in the standard fashion[72,74,75]:

$$\left.\frac{\partial}{\partial x}\,\kappa\,\frac{\partial\phi}{\partial x}\right|_{\mathbf{r}_{i,j,k}} \approx \frac{\kappa_{i+\frac{1}{2},j,k}(\phi_{i+1,j,k} - \phi_{i,j,k}) - \kappa_{i-\frac{1}{2},j,k}(\phi_{i,j,k} - \phi_{i-1,j,k})}{\Delta x^2}, \tag{B1}$$

with equivalent expressions in the y and z directions, where i, j, k are integer node labels in x, y, z axes, respectively. This is second-order accurate in Δx. The material property κ is voxel based. We locally average four voxels to give edge-based values $\kappa_{i\pm 1/2,j,k}$.

The initial fluence $q(\mathbf{r}, 0)$ is discontinuous, comprising a delta function (or convolved delta function) near the source; we found that Crank–Nicolson-type $O(\Delta t^2)$ methods do not handle the nonsmooth fluence well because of the slow decay of nonphysical oscillatory modes excited by the discontinuity. We note that these methods (specifically the alternating-direction implicit method) are commonly used[15,40,41,57] for simulation of Eq. (5) without mention of this issue. Instead we use the explicit Euler $O(\Delta t)$ method, which is less efficient and requires $\Delta t < \Delta x^2/(6\kappa c)$ at every point in the lattice for stability.[75] Evolution is performed by repeated sparse matrix multiplication. Our computational effort is proportional to the number of nodes times the number of time steps: This scales like $1/(\Delta x^3 \Delta t)$, which is proportional to $1/(\Delta x^5 \mu_{s,\min}')$ where $\mu_{s,\min}'$ is the smallest value of μ_s' in the system. Changing Δx from 2 to 1 mm therefore increases the effort by a factor of 32. Implicit methods would be more efficient; we are developing an optimal way to combine explicit and Douglas–Gunn[75] evolution to handle the discontinuous initial condition.

We emulated Dirichlet boundary conditions by removing degrees of freedom $\phi_{i,j,k}$ for nodes on or outside the tissue–air boundary. (The issue of FDTD stability with the more correct Robin boundary conditions is an area for research; we found resulting differences to be slight in our simulations for reasons given in Subsection 3.C.) We simulate only a fraction of the human head volume, corresponding to approximately 3×10^4 nodes for $\Delta x = 2$ mm. Detector signal relative changes were less than 1% compared with the full head simulation of approximately 6×10^5 nodes.

To inject and detect fluence signals, we developed lattice versions of standard optode models that couple to the fluence normal gradient $\hat{\mathbf{n}} \cdot \nabla\phi$ at the tissue–air boundary.[15] To reduce dependence on the detailed local voxel representation of the head surface required that each optode be coupled to the fluence derivative in the beam direction, averaged over a beam of 2-mm radius at the rather large fixed depth of 1.5 voxels below the local surface along the beam direction. This large depth is responsible for the large early-time errors in Fig. 9, but is necessary to maximize immunity to voxelated surface roughness. The source optodes give $q(\mathbf{r}, 0)$; the time-dependent detector signals are then integrated over the time gates to give the components of the signal expectation vector $\mathbf{f}(\mathbf{x})$. Note that this gradient-based source model is equivalent to more usual point-source models, but is more immune to voxelated surface roughness. Exploration of other optode models is an area for future research.

The procedure was validated against known analytic semi-infinite slab solutions (<5% errors beyond 200 ps) and against a Monte Carlo calculation of a heterogeneous cuboid containing an absorbing and scattering inclusion whose nearest approach to the surface was as close as 6 mm (~20% errors). We reserve validation against Monte Carlo in the full head model for a future publication.

We have benefitted from discussions with Jon Stott, Quan Zhang, Tina Chaves, Greg Boverman, Radford Neal, and Jari Kaipio, among many others. We acknowledge funding by National Institutes of Health grants 2P01-NS35611-06, R29-NS38842, P41-RR14075, and from the Center for Innovative Minimally Invasive Therapies. This research was funded in part by the U.S. Army under Cooperative Agreement DAMD17-99-2-9001.

References and Notes

1. D. A. Boas, D. H. Brooks, E. L. Miller, C. A. DiMarzio, M. Kilmer, R. J. Gaudette, and Q. Zhang, "Imaging the body with diffuse optical tomography," IEEE Signal Process. Mag. **18**(6), 57–75 (2001).
2. B. Chance, D. T. Delpy, C. E. Cooper, and E. O. R. Reynolds, eds., "Near-infrared spectroscopy and imaging of living systems," Philos. Trans. R. Soc. London Ser. B **352,** 649–763 (1997).
3. D. A. Boas, M. A. Franceschini, A. K. Dunn, and G. Strangman, "Noninvasive imaging of cerebral activation with diffuse optical tomography," in *In Vivo Optical Imaging of Brain Function*,

R. D. Frostig, ed. (CRC Press, Boca Raton, Fla., 2002), pp. 193–221.
4. A. Villringer and B. Chance, "Non-invasive optical spectroscopy and imaging of human brain function," Trends Neurosci. **20,** 435–442 (1997).
5. S. Fantini, S. A. Walker, M. A. Franceschini, K. T. Moesta, P. M. Schlag, M. Kaschke, and E. Gratton, "Assessment of the size, position, and optical properties of breast tumors *in vivo* by noninvasive optical methods," Appl. Opt. **37,** 1982–1989 (1998).
6. B. W. Pogue, T. O. McBride, S. Osterman, S. Poplack, U. Osterberg, and K. D. Paulsen, "Quantitative hemoglobin tomography with diffuse near-infrared spectroscopy: pilot results in the breast," Radiology **218**(1), 261–266 (2001).
7. J. C. Hebden, H. Veenstra, H. Dehghani, E. M. C. Hillman, M. Schweiger, S. R. Arridge, and D. T. Delpy, "Three-dimensional time-resolved optical tomography of a conical breast phantom," Appl. Opt. **40,** 3278–3287 (2001).
8. N. Shah, A. Cerussi, C. Eker, J. Espinoza, J. Butler, J. Fishkin, R. Hornung, and B. Tromberg, "Noninvasive functional optical spectroscopy of human breast tissue," Proc. Natl. Acad. Sci. USA **98,** 4420–4425 (2001).
9. S. Fantini, M. A. Franceschini-Fantini, J. S. Maier, S. A. Walker, B. Barbieri, and E. Gratton, "Frequency-domain multichannel optical detector for noninvasive tissue spectroscopy and oximetry," Opt. Eng. **34,** 32–42 (1995).
10. A. Klose, A. H. Hielscher, K. M. Hanson, and J. Beuthan, "Three-dimensional optical tomography of a finger joint model for diagnostic of rheumatoid arthritis," in *Photon Propagation in Tissue IV*, D. A. Benaron, B. Chance, M. Ferrari, and M. Kohl, eds., Proc. SPIE **3566,** 151–160 (1998).
11. G. Gratton, M. Fabiani, P. M. Corballis, D. C. Hood, M. R. Goodman-Wood, J. Hirsch, K. Kim, D. Friedman, and E. Gratton, "Fast and localized event-related optical signals (EROS) in the human occipital cortex: comparisons with the visual evoked potential and fMRI," NeuroImage **6,** 168–180 (1997).
12. J. Steinbrink, M. Kohl, H. Obrig, G. Curio, F. Syre, F. Thomas, H. Wabnitz, H. Rinneberg, and A. Villringer, "Somatosensory evoked fast optical intensity changes detected non-invasively in the adult human head," Neurosci. Lett. **291,** 105–108 (2000).
13. V. Ntziachristos, A. G. Yodh, M. Schnall, and B. Chance, "Concurrent MRI and diffuse optical tomography of breast after indocyanine green enhancement," Proc. Natl. Acad. Sci. USA **97,** 2767–2772 (2000).
14. G. Strangman, J. P. Culver, J. H. Thompson, and D. A. Boas, "A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation," NeuroImage **17,** 719–731 (2002).
15. S. R. Arridge, "Optical tomography in medical imaging," Inverse Probl. **15,** R41–R93 (1999).
16. D. A. Boas, M. A. O'Leary, B. Chance, and A. G. Yodh, "Detection and characterization of optical inhomogeneities with diffuse photon density waves: a signal-to-noise analysis," Appl. Opt. **36,** 75–92 (1997).
17. S. Wray, M. Cope, and D. T. Delpy, "Characteristics of the near infrared absorption spectra of cytochrome aa3 and hemoglobin for the noninvasive monitoring of cerebral oxygenation," Biochim. Biophys. Acta **933,** 184–192 (1988).
18. W. D. Heiss, "Ischemic penumbra: evidence from functional imaging in man," J. Cereb. Blood Flow Metab. **20,** 1276–1793 (2000).
19. S. R. Hintz, W.-F. Cheong, J. P. van Houten, D. K. Stevenson, and D. A. Benaron, "Bedside imaging of intracranial hemorrhage in the neonate using light: comparison with ultrasound, computed tomography, and magnetic resonance imaging," Pediatr. Res. **45,** 54–59 (1999).
20. S. R. Hintz, D. A. Benaron, A. M. Siegel, A. Zourabian, D. K. Stevenson, and D. A. Boas, "Bedside functional imaging of the premature infant brain during passive motor activation," J. Perinat. Med. **29,** 335–343 (2001).
21. R. Barbour, H. Graber, Y. Wang, J. Chang, and R. Aronson, "A perturbation approach for optical diffusion tomography using continuous-wave and time-resolved data," in *Medical Optical Tomography: Functional Imaging and Monitoring*, Vol. IS 11 of the SPIE Institute Series (SPIE, Bellingham, Wash., 1993), pp. 87–120.
22. A. J. Devaney, "Reconstruction tomography with diffractive wave-fields," Inverse Probl. **2,** 161–183 (1986).
23. S. R. Arridge and J. C. Hebden, "Optical imaging in medicine: II. Modelling and reconstruction," Phys. Med. Biol. **42,** 841–854 (1997).
24. X. Cheng and D. A. Boas, "Systematic diffuse optical image errors resulting from uncertainty in the background optical properties," Opt. Exp. **4,** 299–307 (1999); http://www.opticsexpress.org.
25. V. Ntziachristos, A. G. Yodh, M. D. Schnall, and B. Chance, "MRI-guided diffuse optical spectroscopy of malignant and benign breast lesions," Neoplasia **4,** 347–354 (2002).
26. S. R. Arridge and M. Schweiger, "Reconstruction in optical tomography using MRI based prior knowledge," in *Information Processing in Medical Imaging*, Y. Bizais, C. Barillot, and R. di Paola, eds. (Kluwer, Dordrecht, The Netherlands, 1995), pp. 77–88.
27. M. Schweiger and S. R. Arridge, "Optical tomographic reconstruction in a complex head model using a priori region boundary information," Phys. Med. Biol. **44,** 2703–2722 (1999).
28. M. Hämäläinen, R. Hari, R. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain," Rev. Mod. Phys. **65,** 413–497 (1993).
29. J. D. Oakley, "Magnetic resonance imaging based correction and reconstruction of positron emission tomography images," Ph.D. dissertation (Service Hospitalier Frederic Joliot, CEA, Orsay, France, 2000).
30. V. Kolehmainen, M. Vauhkonen, J. P. Kaipio, and S. R. Arridge, "Recovery of piecewise constant coefficients in optical diffusion tomography," Opt. Exp. **7,** 468–481 (2000); http://www.opticsexpress.org.
31. M. Kilmer, E. Miller, D. A. Boas, and D. Brooks, "A shape-based reconstruction technique for DPDW data," Opt. Exp. **7,** 481–491 (2000); http://www.opticsexpress.com.
32. M. S. Patterson, B. Chance, and B. C. Wilson, "Time-resolved reflectance and transmittance for the noninvasive measurement of tissue optical properties," Appl. Opt. **28,** 2331–2336 (1989).
33. R. Cubeddu, A. Pifferi, P. Taroni, A. Torricelli, and G. Valentini, "Time-resolved imaging on a realistic tissue phantom: μ_s' and μ_a images versus time-integrated images," Appl. Opt. **35,** 4533–4540 (1996).
34. A. Torricelli, A. Pifferi, P. Taroni, E. Giambattistelli, and R. Cubeddu, "In vivo optical characterization of human tissues from 610 to 1010 nm by time-resolved reflectance spectroscopy," Phys. Med. Biol. **46,** 2227–2237 (2001).
35. A. Kienle, M. S. Patterson, N. Dögnitz, R. Bays, G. Wagnières, and H. van den Bergh, "Noninvasive determination of the optical properties of two-layered media," Appl. Opt. **37,** 779–791 (1998).
36. A. Kienle, T. Glanzmann, G. Wagnières, and H. van den Bergh, "Investigation of two-layered turbid media with time-resolved reflectance," Appl. Opt. **37,** 6852–6862 (1998).
37. A. Pifferi, A. Torricelli, P. Taroni, and R. Cubeddu, "Reconstruction of absorber concentrations in a two-layer structure by use of multidistance time-resolved reflectance spectroscopy," Opt. Lett. **26,** 1963–1965 (2001).
38. C. K. Hayakawa, J. Spanier, F. Bevilacqua, A. K. Dunn, J. S. You, B. J. Tromberg, and V. Venugopalan, "Perturbation

Monte Carlo methods to solve inverse photon migration problems in heterogeneous tissues," Opt. Lett. **26,** 1335–1337 (2001).
39. A. Kienle and T. Glanzmann, "In vivo determination of the optical properties of muscle with time-resolved reflectance using a layered model," Phys. Med. Biol. **44,** 2689–2702 (1999).
40. K. M. Hanson, G. S. Cunningham, and S. S. Saquib, "Inversion based on computational simulations," in *Maximum Entropy and Bayesian Methods*, G. J. Erickson, J. T. Rychert, and C. R. Smith, eds. (Kluwer Academic, Dordrecht, The Netherlands, 1998), pp. 121–135.
41. S. S. Saquib, K. M. Hanson, and G. S. Cunningham, "Model-based image reconstruction from time-resolved diffusion data," in *Medical Imaging 1997: Image Processing*, K. M. Hanson, ed., Proc. SPIE **3034,** 369–380 (1997).
42. G. Nicholls and C. Fox, "Prior modelling and posterior sampling in impedance imaging," in *Bayesian Inference for Inverse Problems*, A. Mohammad-Djafari, ed., Proc. SPIE **3459,** 116–127 (1998).
43. J. P. Kaipio, V. Kolehmainen, E. Somersalo, and M. Vauhkonen, "Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography," Inverse Probl. **16,** 1487–1522 (2000).
44. D. M. Schmidt, J. S. George, and C. C. Wood, "Bayesian inference applied to the electromagnetic inverse problem," Hum. Brain Mapp. **7,** 195–212 (1999).
45. S. Oh, A. B. Milstein, R. P. Millane, C. A. Bouman, and K. J. Webb, "Source–detector calibration in three-dimensional Bayesian optical diffusion tomography," J. Opt. Soc. Am. A **19,** 1983–1993 (2002).
46. I. Kwee, "Towards a Bayesian framework for optical tomography," Ph.D. dissertation (Department of Medical Physics and Bioengineering, University College London, London, 1999).
47. M. J. Eppstein, D. E. Dougherty, T. L. Troy, and E. M. Sevick-Muraca, "Biomedical optical tomography using dynamic parameterization and Bayesian conditioning on photon migration measurements," Appl. Opt. **38,** 2138–2150 (1999).
48. K. J. Friston, "Bayesian estimation of dynamical systems: an application to fMRI," NeuroImage **16,** 513–530 (2002).
49. D. A. Boas, T. J. Gaudette, and S. R. Arridge, "Simultaneous imaging and optode calibration with diffuse optical tomography," Opt. Exp. **8,** 263–270 (2001); http://www.opticsexpress.com.
50. D. J. C. MacKay, "Information theory, inference, and learning algorithms," Chap. 3, available at http://www.inference.phy.cam.ac.uk/mackay/book.html.
51. J. Berger, *Statistical Decision Theory and Bayesian Analysis* (Springer, New York, 1985).
52. S. J. Press, *Bayesian Statistics: Principles, Models, and Applications*, Wiley Series in Probability and Statistics (Wiley, New York, 1989).
53. D. S. Sivia, *Data Analysis: A Bayesian Tutorial* (Oxford U. Press, Oxford, U.K., 1996).
54. S. F. Gull, "Bayesian inductive inference and maximum entropy," in *Foundations*, Vol. 1 of Maximum Entropy and Bayesian Methods in Science and Engineering, G. R. Erickson and C. R. Smith, eds. (Kluwer, Dordrecht, The Netherlands, 1988).
55. R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," Tech. Rep. CRG-TR-93-1 (Department of Computer Science, University of Toronto, Toronto, 1993); available at http://www.cs.toronto.edu/~radford/review.abstract.html.
56. S. R. Arridge and M. Schweiger, "A gradient-based optimisation scheme for optical tomography," Opt. Exp. **2,** 213–226 (1998); http://www.opticsexpress.org.
57. A. H. Hielscher, A. D. Klose, and K. M. Hanson, "Gradient-based iterative image reconstruction scheme for time-resolved optical tomography," IEEE Trans. Med. Imaging **18,** 262–271 (1999).
58. F. Gao, H. Zhao, and Y. Yamada, "Improvement of image quality in diffuse optical tomography by use of full time-resolved data," Appl. Opt. **41,** 778–791 (2002).
59. A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical surface-based analysis. I. Segmentation and surface reconstruction," NeuroImage **9,** 179–194 (1999).
60. E. Okada, M. Firbank, M. Schweiger, S. R. Arridge, M. Cope, and D. T. Delpy, "Theoretical and experimental investigation of near-infrared light propagation in a model of the adult head," Appl. Opt. **36,** 21–31 (1997).
61. F. Bevilacqua, D. Piguet, P. Marquet, J. D. Gross, B. J. Tromberg, and C. Depeursinge, "*In vivo* local determination of tissue optical properties: applications to human brain," Appl. Opt. **38,** 4939–4950 (1999).
62. J. C. Tamraz and Y. G. Comair, *Atlas of Regional Anatomy of the Brain Using MRI: With Functional Correlations* (Springer, New York, 2000).
63. J. Ripoll, S. R. Arridge, and M. Nieto-Vesperinas, "Effect of roughness in nondiffusive regions within diffusive media," J. Opt. Soc. Am. A **18,** 940–947 (2001).
64. F. E. Schmidt, M. E. Fry, E. M. C. Hillman, J. C. Hebden, and D. T. Delpy, "A 32-channel time-resolved instrument for medical optical tomography," Rev. Sci. Instrum. **71,** 256–265 (2000).
65. K. M. Yoo, F. Liu, and R. R. Alfano, "When does the diffusion approximation fail to describe photon transport in random media?," Phys. Rev. Lett. **64,** 2647–2650 (1990).
66. A. Ishimaru, *Wave Propagation and Scattering in Random Media* (Academic, New York, 1978), Vol. 1.
67. A. H. Hielscher, R. E. Alcouffe, and R. L. Barbour, "Comparison of finite-difference transport and diffusion calculations for photon migration in homogeneous and heterogeneous tissues," Phys. Med. Biol. **43,** 1285–1302 (1998).
68. T. B. Durduran, B. Chance, A. G. Yodh, and D. A. Boas, "Does the photon diffusion coefficient depend on absorption?," J. Opt. Soc. Am. A **14,** 3358–3365 (1997).
69. Note that for us this acronym does not imply association with hyperbolic equations. We are evolving a parabolic equation.
70. H. Dehghani, S. R. Arridge, M. Schweiger, and D. T. Delpy, "Optical tomography in the presence of void regions," J. Opt. Soc. Am. A **17,** 1659–1670 (2000).
71. S. R. Arridge, M. Hiraoka, and M. Schweiger, "Statistical basis for the determination of optical pathlength in tissue," Phys. Med. Biol. **40,** 1539–1558 (1995).
72. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge U. Press, 1992); available at http://lib-www.lanl.gov/numerical/bookcpdf.html.
73. S. J. Press, *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Interference*, 2nd ed. (Krieger, Malabar, Fla., 1982).
74. R. D. Richtmeyer and K. W. Morton, *Difference Methods for Initial-Value Problems* (Wiley, New York, 1967).
75. J. W. Thomas, *Numerical Partial Differential Equations: Finite Difference Methods* (Springer-Verlag, New York, 1995).

Parametric evolution for a deformed cavity

Doron Cohen,[1] Alex Barnett,[1] and Eric J. Heller[1,2]
[1]*Department of Physics, Harvard University, Cambridge, Massachusetts 02138*
[2]*Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138*
(Received 31 August 2000; published 27 March 2001)

We consider a classically chaotic system that is described by a Hamiltonian $\mathcal{H}(Q,P;x)$, where (Q,P) describes a particle moving inside a cavity, and x controls a deformation of the boundary. The quantum eigenstates of the system are $|n(x)\rangle$. We describe how the parametric kernel $P(n|m)=|\langle n(x)|m(x_0)\rangle|^2$, also known as the local density of states, evolves as a function of $\delta x = x - x_0$. We illuminate the nonunitary nature of this parametric evolution, the emergence of nonperturbative features, the final nonuniversal saturation, and the limitations of random-wave considerations. The parametric evolution is demonstrated numerically for two distinct representative deformation processes.

DOI: 10.1103/PhysRevE.63.046207 PACS number(s): 05.45.Mt, 03.65.−w, 73.23.−b

I. INTRODUCTION

A. The local density of states

Consider a system that is described by an Hamiltonian $\mathcal{H}(Q,P;x)$, where (Q,P) are canonical variables and x is a constant parameter. Our interest in this paper is in the case where (Q,P) describe the motion of a particle inside a cavity, and x controls the deformation of the confining boundary. The one-dimensional (1D) version of a cavity, also known as a ''potential well,'' is illustrated in Fig. 1. However, we are mainly interested in the case of chaotic cavities in $d>1$ dimensions. Cavities in $d=2$ dimensions, also known as billiard systems, are prototype examples in the studies of classical and quantum chaos, and we shall use them for the purpose of numerical illustrations.

The eigenstates of the quantized Hamiltonian are $|n(x)\rangle$ and the corresponding eigenenergies are $E_n(x)$. The eigenenergies are assumed to be ordered, and the mean level spacing will be denoted by Δ. We are interested in the parametric kernel

$$P(n|m)=|\langle n(x)|m(x_0)\rangle|^2=\mathrm{tr}(\rho_n\rho_m). \tag{1}$$

In the equation above $\rho_m(Q,P)$ and $\rho_n(Q,P)$ are the Wigner functions that correspond to the eigenstates $|m(x_0)\rangle$ and $|n(x)\rangle$, respectively. The trace stands for $dQdP/(2\pi\hbar)^d$ integration. The difference $x-x_0$ will be denoted by δx. We assume a dense spectrum, so that our interest is in ''classically small'' but ''quantum mechanically large'' energy scales. It is important to realize that the kernel $P(n|m)$ has a well defined classical limit. The classical approximation (see remark [1]) is obtained by using microcanonical distributions instead of Wigner functions.

Fixing n, the vector $P(n|m)$ describes the shape of the nth eigenstate in the $\mathcal{H}_0=\mathcal{H}(Q,P;x_0)$ representation. By averaging over several eigenstates one obtains the average shape of the eigenstate (ASOE). We can also identify $P(n|m)$ as the local density of states (LDOS), by regarding it as a function of n, where m is considered to be a fixed reference state. In the latter case an average over few m states is assumed. We shall denote the LDOS by $P(r)$, where $r=(n-m)$. The ASOE is just $P(-r)$. Note that the ASOE and the LDOS are given by the same function. One would have to be more careful with these definitions if $\mathcal{H}_0$ were integrable while $\mathcal{H}$ nonintegrable.

A few words are in order regarding the definition of the LDOS, and its importance in physical applications. The LDOS, also known as strength function [4–6], describes an energy distribution. Conventionally it is defined as follows:

$$\rho(E)=-\frac{1}{\pi}\langle m|\mathrm{Im}\,\mathbf{G}(E)|m\rangle=\sum_n P(n|m)\,\delta(E-E_n), \tag{2}$$

where $\mathbf{G}(E)=1/(E-\mathcal{H}+i0)$ is the retarded Green function. We are interested in chaotic systems, so it should be clear that our $P(r)$ is related by trivial change of variable $(E\mapsto r)$ to the above defined $\rho(E)$. Our $P(r)$ also incorporates an average over the reference state. The LDOS is important in studies of either chaotic or complex conservative quantum systems that are encountered in nuclear physics as well as in atomic and molecular physics. Related applications may be found in mesoscopic physics. Going from $\mathcal{H}_0$ to $\mathcal{H}$ may signify a physical change of an external field, or switching on of a perturbation, or a sudden change of an effective interaction (as in molecular dynamics [7]). The so-called ''line shape'' of the LDOS is important for the understanding of the asso-

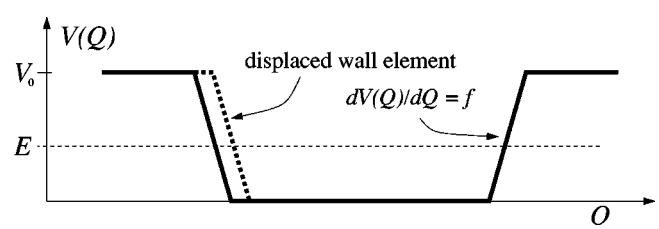


FIG. 1. The shape of a cavity in d dimensions is defined by its $d-1$ boundary. The confining potential is $V(Q)$. The figure illustrates $V(Q)$ for one dimension well. It also can be regarded as a cross section of the $d>1$ cavity. The kinetic energy of the particle is $E=\frac{1}{2}mv^2$. The walls of the cavity exert a field of force f on the bouncing particle. The hard wall limit corresponds to $f\to\infty$ and $V_0\to\infty$. For theoretical considerations it is convenient to assume that f and V_0 are large but finite. Mathematically it is also convenient to think of the embedding space as having some huge but finite volume (not illustrated).

 ©2001 The American Physical Society

ciated dynamics. It is also important to realize that the LDOS is the Fourier transform of the so-called ''survival probability amplitude'' [7] (see Ref. [8] for concise presentation of this point).

B. Parametric evolution

Textbook [9] formulations of perturbation theory can be applied in order to find the LDOS. Partial summations of diagrams to infinite order can be used in order to get an improved Lorentzian-type approximation. However, most textbooks do not illuminate the limitations and the subtleties which are involved in using the conventional perturbative schemes. It is therefore interesting to take a somewhat different approach to the study of LDOS. The roots of this alternate approach can be traced back to the work of Wigner [10] regarding a simple banded random matrix (BRM) model $\mathcal{H}=\mathbf{E}+\delta x\mathbf{B}$. Here $\mathbf{E}$ is a diagonal matrix whose elements are the ordered energies $\{E_n\}$, and $\mathbf{B}$ is a banded matrix. The study of this model can be motivated by the realization that in generic circumstances it is possible to write $\mathcal{H}(Q,P;x) \approx \mathcal{H}_0(Q,P)+\delta x\mathcal{F}(Q,P)$. Using a simple semiclassical argument [11] it turns out that the matrix representation of any generic $\mathcal{F}$, in the eigenbasis that is determined by the chaotic Hamiltonian $\mathcal{H}_0$, is a banded matrix.

The important ingredient (from our point of view) in the original work by Wigner, is the emphasis on the parametric-evolution (PE) of the LDOS. The LDOS describes an energy distribution: For $\delta x=0$ the kernel $P(r)$ is simply a Kroneker delta function. As δx becomes larger, the width as well as the whole profile of this distribution ''evolves.'' Wigner has realized that for his BRM model there are three parametric regimes. For very small δx we have the standard perturbative structure where most of the probability is concentrated in $r=0$. For larger δx we have a Lorentzian lineshape. But this Lorentzian line shape does not persist if we further increase δx. Instead we get a semicircle line shape. Many works about the LDOS have followed [4–6], but the issue of PE has not been further discussed there. The emphasis in those works is mainly on the case where $\mathcal{H}_0$ is an integrable or noninteracting system, while $\mathcal{H}$ is possibly (but not necessarily) chaotic due to some added perturbation term.

The line of study which is pursued in the present work has been originated and motivated by studies of quantum dissipation [12–14]. Understanding PE can be regarded as a preliminary stage in the analysis of the energy spreading process in driven mesoscopic systems. Note that the LDOS gives the energy redistribution due to a ''sudden'' (very fast) change of the Hamiltonian. Unlike the common approach for studies of LDOS, we assume both $\mathcal{H}$ and $\mathcal{H}_0$ to be chaotic. Both correspond to the same parametrically dependent Hamiltonian $\mathcal{H}(Q,P;x)$, and there is nothing special in choosing a particular value $x=x_0$ as a starting point for the PE analysis.

C. Main results

The theory of PE, as discussed in general in Refs. [12–14] and in particular in Refs. [8,15] takes us beyond the random-matrix-theory considerations of Wigner. There appear five (rather than two) different parametric scales (see remark [2]). These are summarized by Table I.

TABLE I. The parametric scales in the general theory of PE are listed (left column) along with the questions that motivate their introduction. The distribution $P(r)$ may contain perturbative tail regions (for $\delta x \ll \delta x_{\rm prt}$), and nonperturbative core regions (for $\delta x > \delta x_c^{\rm qm}$). Nonuniversal (system specific) features may manifest themselves in the core structure for $\delta x \gg \delta x_{\rm NU}$. In generic examples $\delta x \gg \delta x_{\rm SC}$ allows a classical approximation for $P(r)$. We are going to explain that only two independent parametric scales survive in the hard wall limit.

$\delta x_c^{\rm cl}$	Is it possible to linearize $\mathcal{H}(x+\delta x)$?
$\delta x_c^{\rm qm}$	Is it possible to use standard perturbation theory?
$\delta x_{\rm prt}$	Do perturbative tail regions survive?
$\delta x_{\rm NU}$	Do nonuniversal core features show up?
$\delta x_{\rm SC}$	Is it possible to use semiclassical approximation?

In the present paper we consider cavities with hard walls. We are going to explain that because of the ''hard wall limit'' there are only *two* independent parametric scales: One is $\delta x_c^{\rm qm}$ and the others (see remark [3]) coincide with $\delta x_{\rm NU}$. Assuming that $\delta x_c^{\rm qm}$ and $\delta x_{\rm NU}$ are well separated, it follows that there are *three* distinct parametric regimes in the PE of our system. These are the standard perturbative regime ($\delta x \ll \delta x_c^{\rm qm}$), the core-tail regime ($\delta x_c^{\rm NU} \ll \delta x \ll \delta x^{\rm qm}$), and the nonuniversal regime ($\delta x \gg \delta x_{\rm NU}$).

The exploration of the three parametric regimes in the PE of a deformed cavity with hard walls is the main issue of the present paper. To the best of our knowledge such detailed exploration has not been practical in the past. We owe our ability to carry out this task to a powerful technique for finding clusters of billiard eigenstates [19,20]. There are also some secondary issues that we are going to address.

(a) In the strict limit of hard walls the PE becomes nonunitarity. We shall use the 1D well example in order to shed light on this confusing issue. In particular we demonstrate that any truncation of the PE equation leads to false unitarity due to a finite-size edge effect.

(b) For special deformations, namely, those that constitute linear combination of translations, rotations, and dilations, the parametric scales $\delta x_c^{\rm qm}$ and $\delta x_{\rm NU}$ coincide. Consequently there is no longer distinct core-tail regime, and the PE features a quite sharp transition from the standard perturbative regime to the nonuniversal regime.

(c) In the nonuniversal regime we demonstrate that our numerical results are in accordance with our theoretical expectation [8]. Namely, the width of the LDOS profile is determined by time-domain semiclassical considerations, rather then by phase-space or random-wave considerations.

(d) The last section puts our specific study in a larger context. We explain why Wigner's scenario of PE is not followed once hard walls are considered.

II. CAVITY SYSTEM

We consider a particle moving inside a d-dimensional cavity whose volume is V. The kinetic energy of the particle is $E=\frac{1}{2}mv^2$, where m is its mass, and v is its velocity. It is

assumed that this motion is classically chaotic. The ballistic mean free path is $l_{\rm bl}$. One can use the estimate $l_{\rm bl}\sim V/A$, where A is the total area of the walls. The associated time scale is $\tau_{\rm bl}=l_{\rm bl}/v$.

The penetration distance upon a collision is $l=E/f$, where f is the force that is exerted by the wall. Upon quantization we have an additional length scale, which is the De Broglie wavelength $\lambda_B=2\pi\hbar/(mv)$. We shall distinguish between the *hard walls* case where we assume $l<\lambda_B\ll l_{\rm bl}$, and *soft walls* for which $\lambda_B\ll l$. Note that taking $\hbar\to 0$ implies soft walls.

There is a class of special deformations that are shape preserving. These are generated by translations, rotations, and dilations of the cavity. A general deformation need not preserve the billiard shape nor its volume. We can specify any deformation by a function $D(\mathbf{s})$, where $\mathbf{s}$ specifies the location of a wall element on the boundary (surface) of the cavity, and $D(\mathbf{s})\delta x$ is the normal displacement of this wall element. In many practical cases it is possible to use the convention $|D(\mathbf{s})|\sim 1$. With this convention δx has units of length, and its value has the meaning of typical wall displacement.

The eigenenergies of a particle inside the cavity are in general x dependent, and can be written as $E_n=(\hbar k_n)^2/(2m)$. The mean level spacing is

$$\Delta=\hbar v\times\frac{2\pi}{\Omega_d}\frac{1}{\mathcal{V}}\lambda_B^{d-1} \tag{3}$$

where $\Omega_d=2\pi,4\pi,\ldots,$ for $d=2,3,\ldots$. In our numerical study we shall consider a quarter stadium with curved edge of radius 1 and straight edge of length 1. The ''volume'' of the quarter stadium is $\mathcal{V}=1+\pi/4$. The ''area'' of its boundary $A=4+\pi/2$ is just the perimeter. We shall look on the parametric evolution of eigenstates around $k\sim 400$ where the mean level spacing in k units is $\tilde{\Delta}=\Delta/(\hbar v)\approx 0.0088$.

III. PARAMETRIC EVOLUTION

Consider the quantum-mechanical state $\psi=|m(x_0)\rangle$. We can write $\psi=\Sigma_n a_n(x)|n(x)\rangle$. The parametric kernel can be written as $P(n|m)=|a_n(x)|^2$. It is a standard exercise to obtain (from $d\psi/dx=0$ and differentiating by parts) the following equation for the amplitudes:

$$\frac{da_n}{dx}=-\frac{i}{\hbar}\sum_m \mathbf{W}_{nm}(x)a_m. \tag{4}$$

In order to get $P(n|m)$ one should solve this equation with the initial conditions $a_n(x_0)=\delta_{nm}$. The transitions between levels are induced by the matrix elements

$$\mathbf{W}_{nm}=\frac{i\hbar}{E_n-E_m}\left(\frac{\partial\mathcal{H}}{\partial x}\right)_{nm} \tag{5}$$

and we use the ''gauge'' convention $\mathbf{W}_{nm}=0$ for $n=m$. (Only one parameter is being changed and therefore Berry's phase is not an issue.)

Equation (4) is a possible starting point for constructing a perturbation theory for the PE of $P(r)$ (see Ref. [14] for more details). As an input for this equation we need the matrix elements of $\partial\mathcal{H}/\partial x$. These can be calculated using a simple boundary integral formula [16] whose simplest derivation [14] is as follows: The position of the particle in the vicinity of a wall element is $Q=(z,\mathbf{s})$, where $\mathbf{s}$ is a surface coordinate and z is a perpendicular ''radial'' coordinate. We take $f=\infty$ so that

$$\frac{\partial\mathcal{H}}{\partial x}=-D(\mathbf{s})V_0\delta(z). \tag{6}$$

The logarithmic derivative of the wave function on the boundary is $\varphi(\mathbf{s})/\psi(\mathbf{s})$, where $\varphi(\mathbf{s})=\mathbf{n}\cdot\nabla\psi$, and $\mathbf{n}$ is a unit vector in the z direction. For $z>0$ the wave function $\psi(Q)$ is a decaying exponential. If V_0 is large enough, then the exponential decay is fast, and we can treat the boundary as if it were locally flat. It follows that the logarithmic derivative of the wave function on the boundary should be equal to $-\sqrt{2mV_0}/\hbar$. Consequently one obtains the following expression for the matrix elements:

$$\left(\frac{\partial\mathcal{H}}{\partial x}\right)_{nm}=-\frac{\hbar^2}{2m}\oint\varphi_n(\mathbf{s})\varphi_m(\mathbf{s})D(\mathbf{s})d\mathbf{s}. \tag{7}$$

In the one-dimensional case the boundary integral is replace by the sum $\Sigma_s\varphi_n(s)\varphi_m(s)D(s)$ where $s=1,2$ are the two turning points of the potential well.

IV. HARD WALLS AND NONUNITARITY

For the purpose of the following argumentation it is convenient to take $f=\infty$, but to keep V_0 large but finite. Mathematically it is also convenient to think of the embedding space as having some huge but finite volume. We would like to illuminate a subtlety which is associated with the hard wall limit $V_0\to\infty$. For any finite V_0 the parametric kernel satisfies

$$\sum_n P(n|m)=p_{\rm total} \tag{8}$$

with $p_{\rm total}=1$. This follows from the fact that $|n(x)\rangle$ is a complete orthonormal basis for any x. However, for hard walls ($V_0=\infty$) this statement is not true. This implies that for hard walls the PE is nonunitary. We are going to explain this point below.

Let us denote the volume of the original cavity by $\mathcal{V}_0$ and of the deformed cavity by $\mathcal{V}$. The volume shared by the deformed and the undeformed cavities will be denoted by $\mathcal{V}_0\cap\mathcal{V}$ and we shall use the notation $\eta=(\mathcal{V}_0\cap\mathcal{V})/\mathcal{V}_0$. For the purpose of the following argument let us consider a reference state m whose energy E_m is well below V_0. Let us also assume that the wall displacement is large compared to De Broglie wavelength. Consequently the expression for $P(n|m)$ has the following semiclassical structure:

$$P(n|m)\approx\eta\times f(E_n-E_m)+(1-\eta)\times g(E_n-E_c), \tag{9}$$

where $E_c=(\mathcal{V}_0+E_m)\sim\mathcal{V}_0$. The above result can be deduced by assuming that the wave functions look ergodic in space, but still that they are characterized by a well-defined *local* wavelength. An equivalent derivation is obtained by using the phase-space picture of Ref. [8]. Both f and g in the above expression have unit normalization, and therefore $p_{\text{total}}=1$ for any finite $\mathcal{V}_0$. However, for hard walls ($\mathcal{V}_0=\infty$) we have $E_c=\infty$ and therefore $p_{\text{total}}=\eta$. We may say the the operation of taking the hard wall limit does not commute with the summation in Eq. (8). An analogous statement can be derived regarding the summation $\Sigma_m P(n|m)$, with the respective definition $\eta=(V_0\cap V)/V$.

The correctness of the above observation becomes less trivial if we consider Eq. (4) with expressions (5) and (7) substituted for the matrix elements. Looking at Eq. (5) with Eq. (7) it seems as if the matrix $\mathbf{W}_{nm}$ is Hermitian, and therefore should generate unitary PE. But this statement is mathematically correct only for (any) finite truncation N of the PE equation. For $N=\infty$ the matrix $\mathbf{W}_{nm}$ becomes non-Hermitian. It turns out that for any finite N, there is a pile-up of probability in the edges of the spreading profile, due to finite-size effect. We shall demonstrate this effect in the next section using a simple 1D example. In other words, if we solve Eq. (4) for hard-walled cavity, we get as a result Eq. (9) with $E_c=E_N$. For $N=\infty$ we get $E_c=\infty$ and therefore $p_{\text{total}}=\eta$ in accordance with the conclusion of the previous paragraph.

Thus if either $V_0<\infty$ or $N<\infty$ then we have unitary PE. But for hard walls, meaning $V_0=\infty$ with $N=\infty$, we have nonunitary PE. The lost probability is associated with the second term in Eq. (9). This term is peaked around a high energy E_c. For hard walls $E_c=\infty$ and consequently some probability is lost. The above picture is supported by the simple pedagogical example of the next section.

V. PARAMETRIC EVOLUTION FOR A 1D BOX

Consider a 1D box with hard walls, where the free motion of the particle is within $0<Q<\mathbf{a}$. The eigenstates of the Hamiltonian are

$$|n(\mathbf{a})\rangle\to(-1)^n\sqrt{\frac{2}{\mathbf{a}}}\sin(k_nQ), \tag{10}$$

where $k_n=n\times(\pi/\mathbf{a})$ is the wave number, and $n=1,2,\ldots,$ is the level index. The phase factor $(-1)^n$ has been introduced for convenience. We consider now the parametric evolution as a function of $\mathbf{a}$. One easily obtains

$$\langle n(\mathbf{a})|m(\mathbf{a}_0)\rangle=(-1)^n\sqrt{\eta}\,\frac{\sin(\pi\eta n)}{\pi}\,\frac{2m}{\eta^2n^2-m^2}, \tag{11}$$

where $\eta\equiv\mathbf{a}_0/\mathbf{a}$ is assumed to be smaller than 1, corresponding to expansion of the box. The probability kernel is $P(n|m)=|\langle n|m\rangle|^2$. One can verify that the parametric evolution in the $\mathbf{a}_0\mapsto\mathbf{a}$ direction is unitary, meaning that $\Sigma_n P(n|m)=1$. On the other hand, in the $\mathbf{a}\mapsto\mathbf{a}_0$ direction the parametric evolution is nonunitary, because $\Sigma_m P(n|m)=\eta$. The profile of $P(n|m)$ for fixed n is illustrated by a dashed line in Fig. 2.

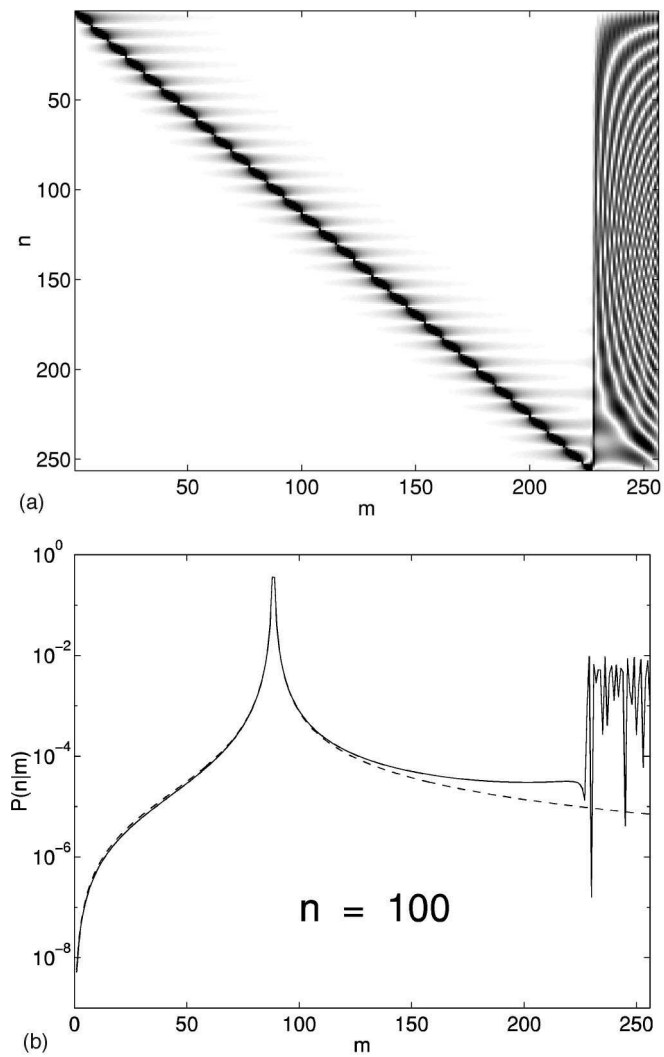


FIG. 2. (a) An image of the kernel $P(n|m)$ for 13% expansion of the box (i.e., $\mathbf{a}/\mathbf{a}_0=1.13$). The kernel has been calculated numerically using Eq. (13) with finite truncation $N=256$. (b) The profile of a representative row of $P(n|m)$. The dashed line is the $N=\infty$ analytical result using Eq. (11).

We can restore unitarity by making V_0 large but finite. In such case, a variation of the above calculation leads to the following picture: Consider the overlap of a reference level $n(\mathbf{a})$ with the levels $m(\mathbf{a}_0)$. As in the case $V_0=\infty$ there is a probability η which is located in the levels whose energies are $E_m\approx E_n$. But now the ''lost'' probability $(1-\eta)$ is located in the levels whose energies are $E_m\approx E_n+V_0$. Thus we have $\Sigma_m P(n|m)=1$ rather than $\Sigma_m P(n|m)=\eta$.

We again consider the case $V_0=\infty$. The normal derivative on the boundary is $\varphi_n(\mathbf{a})=\sqrt{(2/\mathbf{a})}k_n$. Hence we can easily get the following result:

$$\frac{1}{\hbar}\mathbf{W}_{nm}=\frac{-i}{k_n^2-k_m^2}\varphi_n(\mathbf{a})\varphi_m(\mathbf{a})=-i\frac{1}{\mathbf{a}}\frac{2nm}{n^2-m^2}. \tag{12}$$

It is more convenient to use $\alpha=\ln(\mathbf{a})$ for parametrization. Hence the equation that describes the parametric evolution is

$$\frac{da_n}{d\alpha}=-\sum_m \frac{2nm}{n^2-m^2}a_m. \tag{13}$$

For any finite truncation $N<\infty$ this equation manifestly generates unitary parametric evolution. It is only for $N=\infty$ that it becomes equivalent to the nonunitary evolution of the 1D box. Again, one can wonder where the ''lost'' probability is located if $N<\infty$. The answer is illustrated in Fig. 2. We see that the ''lost'' probability piles up at the edge of the (truncated) tail.

VI. MATRIX ELEMENTS FOR CHAOTIC CAVITY

It is possible to use semiclassical considerations [11] in order to determine the band profile of the matrix Eq. (7). The application to the cavity example has been introduced in Ref. [14], and numerically demonstrated in Ref. [17]. The accuracy of this semiclassical estimate is remarkable. Here we summarize the recipe. First one should generate a very long (ergodic) trajectory, and define for it the fluctuating quantity

$$\mathcal{F}(t)=-\frac{\partial\mathcal{H}}{\partial x}=\sum_{\rm col} 2mv\cos(\theta_{\rm col})D_{\rm col}\delta(t-t_{\rm col}), \tag{14}$$

where $t_{\rm col}$ is the time of a collision, $D_{\rm col}$ stands for $D(\mathbf{s})$ at the point of the collision, and $v\cos(\theta_{\rm col})$ is the normal component of the particle's velocity. Each delta spike (for soft walls it is actually a narrow rectangular spike) corresponds to one collision. Now one can calculate the correlation function $C(\tau)$ of the fluctuating quantity $\mathcal{F}(t)$, and its Fourier transform $\tilde{C}(\omega)$. The semiclassical estimate for the band profile is

$$\left\langle\left|\left(\frac{\partial\mathcal{H}}{\partial x}\right)_{nm}\right|^2\right\rangle\approx\frac{\Delta}{2\pi\hbar}\tilde{C}\left(\frac{E_n-E_m}{\hbar}\right) \tag{15}$$

Reference [18] contains a systematic study of the function $\tilde{C}(\omega)$. For large ω, meaning $\omega\gg 1/\tau_{\rm bl}$, one can use

$$\tilde{C}(\omega)\approx 2m^2v^3\langle|\cos\theta|^3\rangle\frac{1}{\mathcal{V}}\oint[D(\mathbf{s})]^2d\mathbf{s}, \tag{16}$$

where the geometric factor is $\langle|\cos(\theta)|^3\rangle=1,4/3\pi,\ldots,$ for $d=1,2,\ldots$. A lengthy calculation [14] reveals that Eq. (15) with (16) substituted, is an exact global result if we could assume that the cavity eigenfunctions look similar to ''random waves,'' and that different wave functions are uncorrelated. However, it turns out that to take this random wave result as a global approximation is an oversimplification. For $\omega\ll 1/\tau_{\rm bl}$, using the semiclassical recipe and assuming strongly chaotic cavity, one obtains

$$\tilde{C}(\omega)\approx\tilde{C}(\infty)\times(\tau_{\rm bl}\omega)^{\gamma} \tag{17}$$

with $\gamma=4$ for dilations and translations, $\gamma=2$ for rotations, and $\gamma=0$ for normal deformations. We use the term ''special deformations'' [18] in order to distinguish those deformations that have the property $\tilde{C}(\omega)\to 0$ in the limit $\omega\to 0$. Any combination of dilations, translations, and rotations is a special deformation. Around the bouncing frequency ($\omega\sim 1/\tau_{\rm bl}$) the function $\tilde{C}(\omega)$ typically displays some nonuniversal (system and deformation specific) structure. This is true for any typical deformation, but for some deformations the nonuniversal features are more pronounced. If the cavity has bouncing ball modes, we may get also a modified (nonuniversal) behavior in the small frequency limit. For the purpose of general discussion it is convenient to assume that the interpolation between Eqs. (17) and (16) is smooth, but in actual numerical calculation the actual $\tilde{C}(\omega)$ is computed (see below).

As a numerical example we have picked the stadium billiard. We have found the eigenstates of a desymmetrized (quarter) stadium as described in Ref. [17]. We have selected those eigenstates whose eigenenergies k_n are in the vicinity of $k=400$. Our two representative deformations are (a) rotation around the stadium center and (b) generic (nonspecial) deformation involving the curved edge. In the latter case the curved edge of the quarter stadium ($0<s<\pi/2$) is pushed outwards with $D(s)=[\cos(s)]^2$, while for the straight edges $D(s)=0$. (The corner $s=0$ is the 90° intersection of the curved edge with the long straight edge). The respective band profiles are displayed in Fig. 3. The band profile has been defined as

$$B(\kappa)=\frac{1}{4k^2}\left\langle\left|\oint\varphi_n(\mathbf{s})\varphi_m(\mathbf{s})\,D(\mathbf{s})d\mathbf{s}\right|^2\right\rangle, \tag{18}$$

where $\kappa=(k_n-k_m)$ is the distance from the diagonal. Note that $B(\kappa)$ is just a scaled version of the semiclassical $\tilde{C}(\omega)$ as implied by Eq. (15) with (7). The remarkable agreement of $B(\kappa)$ with the semiclassical calculation has been demonstrated in Refs. [17,18].

It is important to realize that in the hard wall limit (which is assumed here) the matrix $(\partial\mathcal{H}/\partial x)_{nm}$ is not a banded matrix. It would become banded if we were assuming *soft* walls. For soft walls $\tilde{C}(\omega)$ becomes vanishingly small for $\omega\gg v/l$. The bandwidth in energy units is $\Delta_b=\hbar v/l$, and in dimensionless units it is

$$b=\frac{\Delta_b}{\Delta}=\frac{\mathcal{V}}{l\lambda_B^{d-1}}. \tag{19}$$

Unless stated otherwise we have $b=\infty$.

VII. PARAMETRIC EVOLUTION—NUMERICAL RESULTS

The parametric evolution of $P(r)$ for rotation and for generic deformation of the stadium is illustrated by the images of Fig. 4 and by the plots of Fig. 5 and Fig. 6. The calculation of each $P(r)$ profile is carried out as follows: Given δx we use the method which is described in Ref. [20] in order to calculate the matrix $P(n|m)$. Then we plot the elements of $P(n|m)$ versus $\kappa=[k_n(x)-k_m(0)]$. In order to obtain the

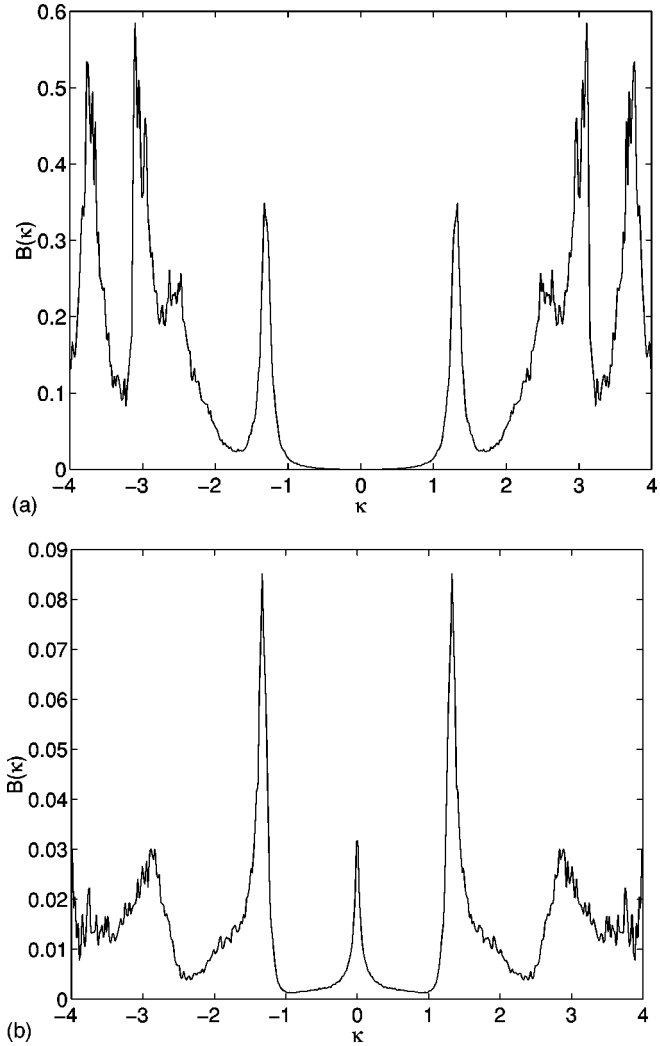


FIG. 3. Band profiles for deformations of the quarter stadium as defined by Eq. (18). (a) Rotation around the stadium center. (b) Generic (nonspecial) deformation involving displacement of the curved edge. It is important to notice that for the special deformation we have $\tilde{C}(\omega)\to 0$ in the limit $\omega\to 0$.

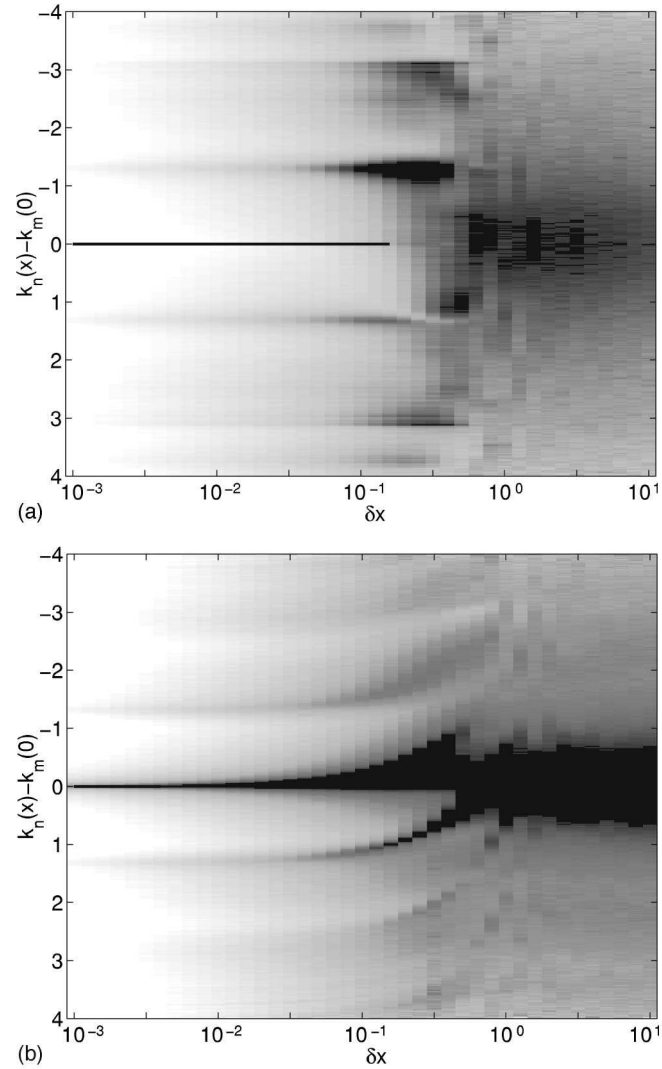


FIG. 4. Each column is an image of $P(r)$ versus κ for a different value of δx. There are 41 columns. The value $\delta x\sim 1$ corresponds roughly to λ_B displacement of the boundary. (a) is for rotation and (b) is for generic deformation. Note that the $r=0$ component is excluded from the image. Instead we have plotted over the image an $r=0$ line wherever this component contains more than 50% of the probability. In the lower figure this line cannot be resolved from the developing core region.

average profile the plot is smeared using standard procedure (see remark [21]). The transformation from κ to $r=(n-m)$ is done using the relation (see remark [21]):

$$\kappa=\tilde{\Delta}\cdot r-\frac{1}{d}k\times\frac{\delta\mathcal{V}}{\mathcal{V}}. \tag{20}$$

Above $\tilde{\Delta}$ is the mean level spacing of the $\{k_n\}$ spectrum, and $\delta\mathcal{V}$ is the volume change that is associated with the deformation (it is approximately proportional to δx). If the deformation is volume preserving (as in the case of rotation) then the second term equals zero. But for the generic deformation that we have picked in our second numerical example, the volume is not preserved, and the systematic ''downward'' shift of the levels should be taken into account.

Looking first at the case of rotation, we clearly see two parametric regimes: The standard perturbative regime ($\delta x<0.2$) and the nonuniversal regime ($\delta x>0.2$). Let us clarify this observation. We see that for $\delta x<0.2$ most of the probability is well concentrated in $r=0$. This implies that we can use standard perturbation theory in order to estimate the $r=0$ probabilities. On the other hand, for $\delta x>0.2$ the perturbative nature of $P(r)$ is destroyed. Now $P(r)$ becomes smoother, and eventually (for $\delta x>0.5$) there is a very good fitting with Lorentzian (see lower plot in Fig. 5).

The qualitative explanation for the Lorentzian profile is as follows. For $\delta x>0.5$ the typical displacement of the walls is of the order of λ_B. Therefore the $|n(x)\rangle$ eigenstates become uncorrelated with the $|m(0)\rangle$ eigenstates. Consequently $P(r)$ becomes δx independent. The Lorentzian profile agrees with the assumption of uncorrelated random waves as explained in the Appendix.

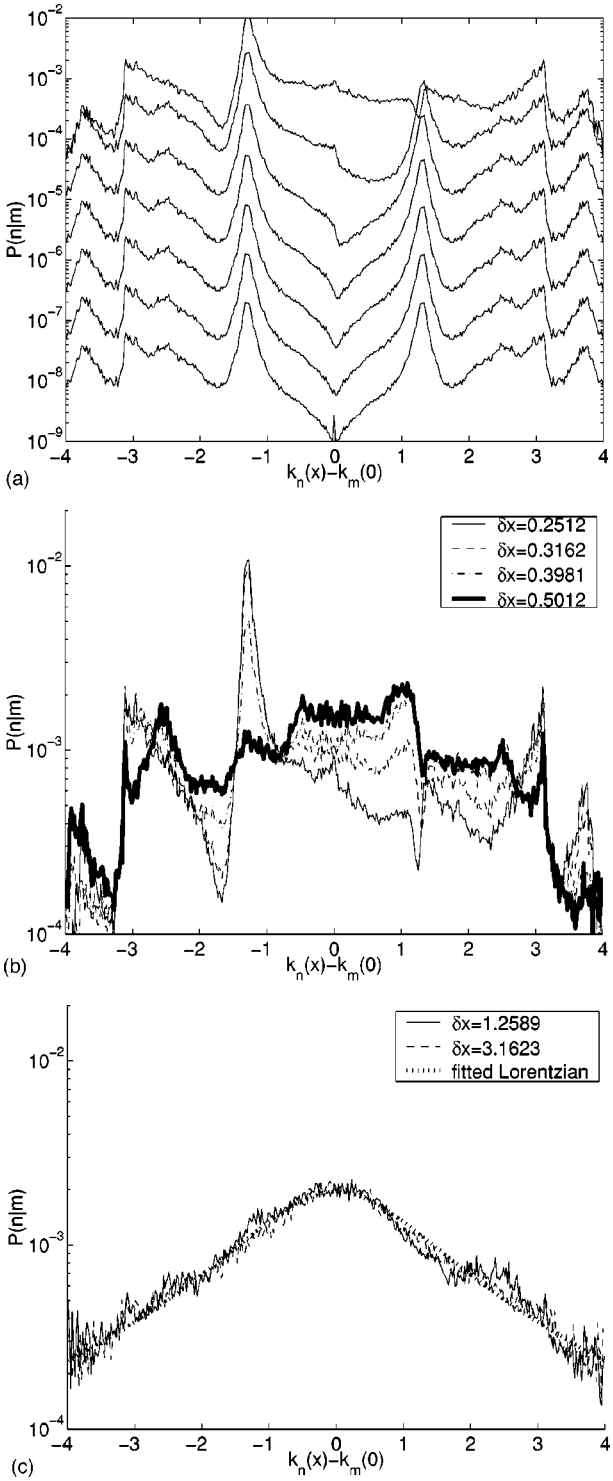


FIG. 5. Representative plots of $P(r)$ for the case of rotation. (a) is for $0.0010 \leqslant \delta x \leqslant 0.2512$. The $r=0$ component is excluded. In (b) the fitted Lorentzian is indistinguishable from the actual profile.

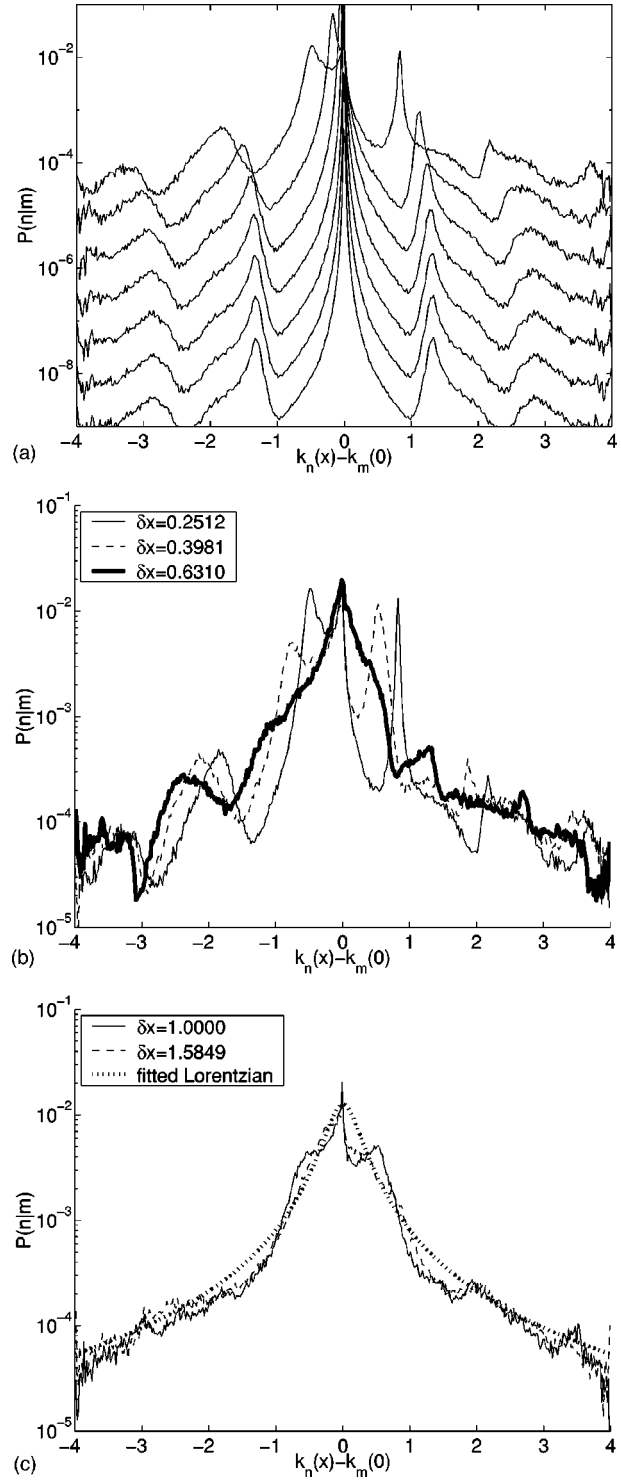


FIG. 6. Representative plots of $P(r)$ for the case of generic deformation. (a) is for $0.0010 \leqslant \delta x \leqslant 0.2512$. In (b) a fitted Lorentzian is overlayed for the purpose of comparison.

Let us now look at the case of generic deformation. Here we clearly see three parametric regimes: The standard perturbative regime ($\delta x < 0.004$), the core-tail regime ($0.004 < \delta x < 0.2$), and the nonuniversal regime ($\delta x > 0.2$). Let us clarify this observation. As in the case of rotation there is a standard perturbative regime ($\delta x < 0.004$) where most of the probability is well concentrated in $r=0$. For larger deformation, namely for $\delta x > 0.004$, standard perturbation theory is no longer applicable because the $n=m$ level is mixed nonperturbatively with other (neighboring) levels. As a result $P(r)$ contains a nonperturbative ''core'' component. However, for $0.004 < \delta x < 0.2$ we definitely do not get a Lorent-

zian. Rather the tails of $P(r)$ keep growing in the same way as in the standard perturbative regime.

In the case of the generic deformation, as in the case of rotation, we enter the nonuniversal regime, and eventually (for $\delta x > 0.6$) we get a smooth Lorentzian-like distribution. However, the Lorentzian-like distribution is not identical to that of the rotation case. Also the similarity to proper Lorentzian is far from being satisfactory (see lower plot in Fig. 6). This means that the random-wave picture of the Appendix is an oversimplification.

In the following sections we are going to summarize the theoretical considerations [8] that explain the observed parametric scenario. In particular we are going to illuminate the way in which nonperturbative features emerge; to clarify the crossover to the nonuniversal regime; and to explain the specific nature of the nonuniversal distribution.

VIII. THE STANDARD PERTURBATIVE REGIME

Standard perturbation theory gives the following first order expression for the LDOS

$$P(n|m) \approx \delta_{nm} + \delta x^2 \frac{|(\partial \mathcal{H}/\partial x)_{nm}|^2}{(E_n - E_m)^2}. \tag{21}$$

This expression is most straightforwardly obtained by inspecting Eqs. (4) and (5). We can define the (total) transition probability as

$$p(\delta x) = \sum_{r \neq 0} P(r). \tag{22}$$

Using Eq. (21) combined with (15) we get the following estimate:

$$p(\delta x) \approx \delta x^2 \times \frac{1}{\hbar^2} \int_{|\omega| > \Delta/\hbar} \frac{d\omega}{2\pi} \frac{\tilde{C}(\omega)}{\omega^2}. \tag{23}$$

Standard perturbation theory is applicable as long as $p(\delta x) \ll 1$. This can be converted into an equivalent inequality $\delta x \ll \delta x_c^{\rm qm}$. By this definition $\delta x_c^{\rm qm}$ is the parametric deformation which is needed in order to mix the initial level m with other levels $n \neq m$.

If we use Eq. (23) for a special deformation, then we have $\gamma > 1$, and consequently the integral is not sensitive to the exclusion of the $|r| < \Delta$ region. As a result we have $\delta x_c^{\rm qm} \propto \hbar$. Using Eq. (17) with (16) we get

$$\delta x_c^{\rm qm}|_{\rm special} = \left(l_{\rm bl} \frac{1}{\mathcal{V}} \oint [D(\mathbf{s})]^2 d\mathbf{s} \right)^{-1/2} \times \lambda_B. \tag{24}$$

In the case of generic deformation, Eq. (24) is not valid because the value of the integral in Eq. (23) is predominantly determined by the $\omega \sim \Delta/\hbar$ lower cutoff rather than by an effective $\omega \sim 1/\tau_{\rm bl}$ lower cutoff. As a result one obtains

$$\delta x_c^{\rm qm}|_{\rm generic} = \left(l_H \frac{1}{\mathcal{V}} \oint [D(\mathbf{s})]^2 d\mathbf{s} \right)^{-1/2} \times \lambda_B, \tag{25}$$

where $\ell_H = v t_H$ is the length which is associated with the Heisenberg time $t_H = 2\pi\hbar/\Delta$.

It is illuminating to use the convention $D(\mathbf{s}) \sim 1$, such that δx measures the typical displacement of a wall element. With this convention we get from Eqs. (24),(25) the following:

$$\delta x_c^{\rm qm}|_{\rm special} \approx \lambda_B, \tag{26}$$

$$\delta x_c^{\rm qm}|_{\rm generic} = \left(\frac{\lambda_B^{d-1}}{A_{\rm eff}} \right)^{1/2} \times \lambda_B, \tag{27}$$

$$\delta x_c^{\rm qm}|_{\rm diffractive} \sim \lambda_B. \tag{28}$$

In the generic case the effective area of the deformed boundary $A_{\rm eff}$ may be smaller than the total area A of the boundary. The effective area $A_{\rm eff}$ can be formally defined by comparing Eq. (27) with Eq. (25). Equation (28) has been added for sake of completeness of our presentation. It correspond to the diffractive limit $A_{\rm eff} \to \lambda_B^{d-1}$. Note that this limit is beyond the scope of the present study. Thus in the generic case, the wall displacement needed to mix levels is much smaller than λ_B. In the generic case $\delta x_c^{\rm qm} \propto \hbar^{(d+1)/2}$ rather than $\delta x_c^{\rm qm} \propto \hbar$. What happens with perturbation theory beyond $\delta x_c^{\rm qm}$? This is the subject of the next section.

IX. THE CORE-TAIL REGIME

For $\delta x > \delta x_c^{\rm qm}$ standard perturbation theory diverges due to the nonperturbative mixing of neighboring levels on small scale: Once δx becomes of the order of $\delta x_c^{\rm qm}$ several levels are mixed, and as δx becomes larger, more levels are being mixed nonperturbatively. Consequently it is natural to distinguish between *core* and *tail* regions [13–15]. Most of the spreading probability is contained within the core region, which implies a natural extension of first-order perturbation theory (FOPT): The first step is to transform Eq. (4) into a new basis where transitions within the core are eliminated. The second step is to use FOPT (in the new basis) in order to analyze the core-to-tail transitions. Details of this procedure, which is in the spirit of degenerate perturbation theory, were discussed in Ref. [14]. The most important (and nontrivial) consequence of this procedure is the observation that mixing on small scales does not affect the transitions on large scales. Therefore we have in the tail region $P(n|m) \propto \delta x^2$ rather than $P(n|m) \propto \delta x$. The validity of this observation has been numerically illustrated in Ref. [15].

Following the above reasoning we define the *tail* region as consists of those levels whose ''occupation'' can be calculated using perturbation theory, while the *core* is the nonperturbative component in the vicinity of $r = 0$. Assuming that only one scale characterize the core width, one arrives at the following practical approximation

$$P_{\rm prt}(r) = \frac{\Delta}{2\pi\hbar} \tilde{C}\left(\frac{E_n - E_m}{\hbar} \right) \frac{\delta x^2}{[\Gamma(\delta x)]^2 + (E_n - E_m)^2}. \tag{29}$$

It is implicit in this definition that $(E_n - E_m)$ should be regarded as a function of $r=(n-m)$. The parameter $\Gamma(\delta x)$ is determined (for a given δx) such that $P_{\rm prt}(r)$ has a unit normalization. One may say that $\Gamma(\delta x)$ regularizes the behavior around $r=0$. For generic deformation we get

$$\Gamma(\delta x) \equiv b_0 \Delta \approx \left(\frac{\delta x}{\delta x_c^{\rm qm}} \right)^2 \times \Delta. \tag{30}$$

The core is defined as the region $|r|<b_0$, and the outer $(|r|>b_0)$ regions are the tails. For $\delta x \ll \delta x_c^{\rm qm}$ we get $b_0 \ll 1$ and the core-tail structure Eq. (29) becomes equivalent to the standard perturbative result Eq. (21). It should be clear that the core-tail structure is a generalization of Wigner's Lorentzian. It is indeed a Lorentzian in the special case of a ''flat'' band profile.

We turn now to analyze our numerical results. We have verified (see, e.g., Fig. 7) that for $\delta x < 0.2$ we have good agreement with perturbation theory irrespective of whether we have a core component (which is the case for the generic deformation) or not (which is the case for the rotation). As we come closer to $\delta x \sim 0.2$ the agreement becomes worse, and for $\delta x > 0.3$ we have a total collapse of perturbation theory. In Fig. 8 we display the total transition probability $p(\delta x)$ as a function of δx. This plot should be used in order to numerically determine the value of $\delta x_c^{\rm qm}$, say as the value where $p(\delta x)=1/2$.

Figure 8 also displays comparison with the corresponding perturbative calculation [using Eq. (22) with (29)]. The agreement (for $\delta x < \delta x_c^{\rm qm}$) in case of the rotation is remarkable. For larger values of δx one observes a linear drop of the total probability due to the nonunitary nature of the evolution. (This drop is clearly linear in a linear-linear scale which is not displayed.) The eventual rise of the total transition probability (beyond 1) in case of the generic deformation, reflects numerical errors in the determination of the small κ overlaps. See Ref. [20] for further details.

The lower subfigure in Fig. 8 displays the calculated core width b_0 for the generic deformation. Recall that b_0 is determined, given δx, such that $P_{\rm prt}(r)$ of Eq. (29) is normalized. Also calculated is the width b_c of the region that contains 50% of the probability. The width b_c is calculated for both $P_{\rm prt}(r)$ and $P(r)$. Note that $b_c=1$ in the standard perturbative regime. The determination of $\delta x_c^{\rm qm}$ for the generic deformation becomes more convenient by using this plot. Also the crossover (at $\delta x_{\rm NU} \sim 0.2$) to the nonuniversal regime is most pronounced.

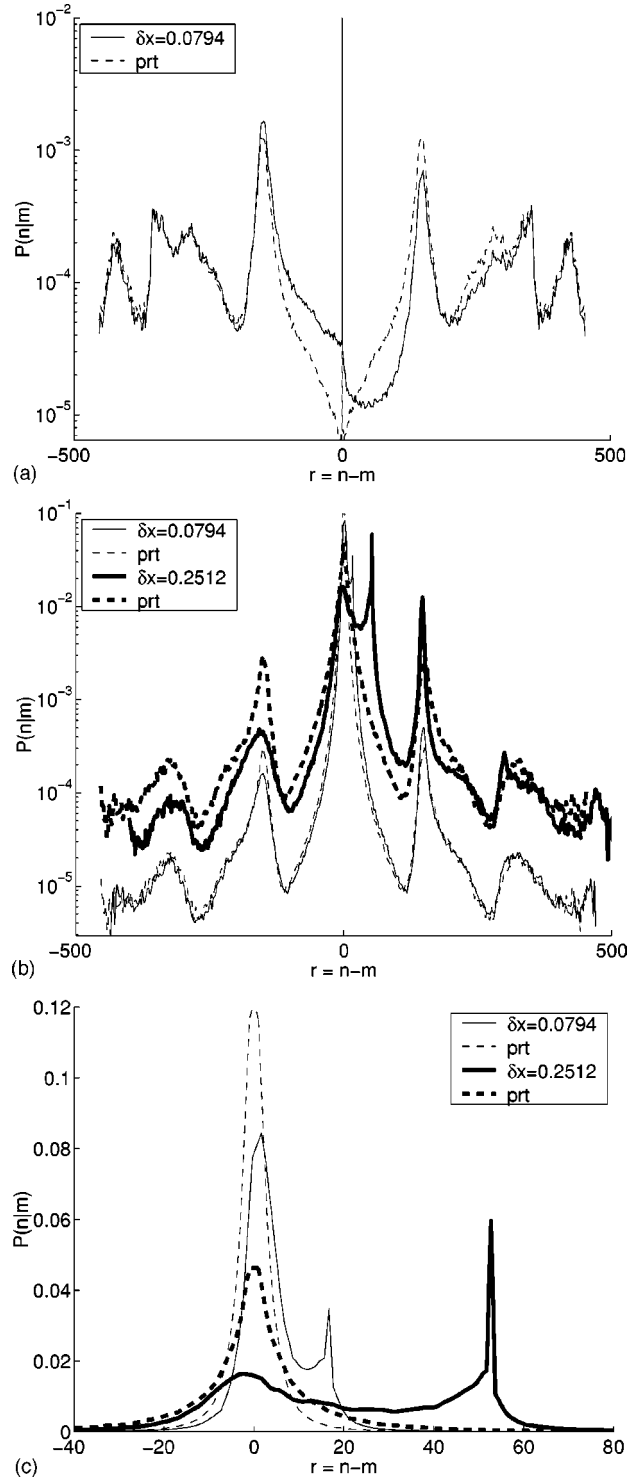


FIG. 7. Comparison between the perturbative calculation $P_{\rm prt}(r)$ and the actual $P(r)$. (a) is for the rotation and (b) is for the generic deformation, (c) is the same as (b), giving a zoom over the ''birth'' of the nonperturbative semiclassical saturation profile.

X. NONUNIVERSAL REGIME

The validity of Eq. (29) as a global approximation relies on the assumption that the core is characterized only by one scale (which is b_0). But this assumption ceases to be true if δx is large enough. We have explained in Ref. [8] that the width of the *core* defines a ''window'' through which we can view the ''landscape'' of the semiclassical analysis. As δx becomes larger, this ''window'' becomes wider, and eventually some of semiclassical structure is exposed. This is marked by the nonuniversal parametric scale $\delta x_{\rm NU}$. For δx larger than $\delta x_{\rm NU}$, the nonuniversal structure of the core is exposed.

What is the semiclassical structure of the LDOS? Time-domain semiclassical considerations [See Eq. (10) of Ref. [8]

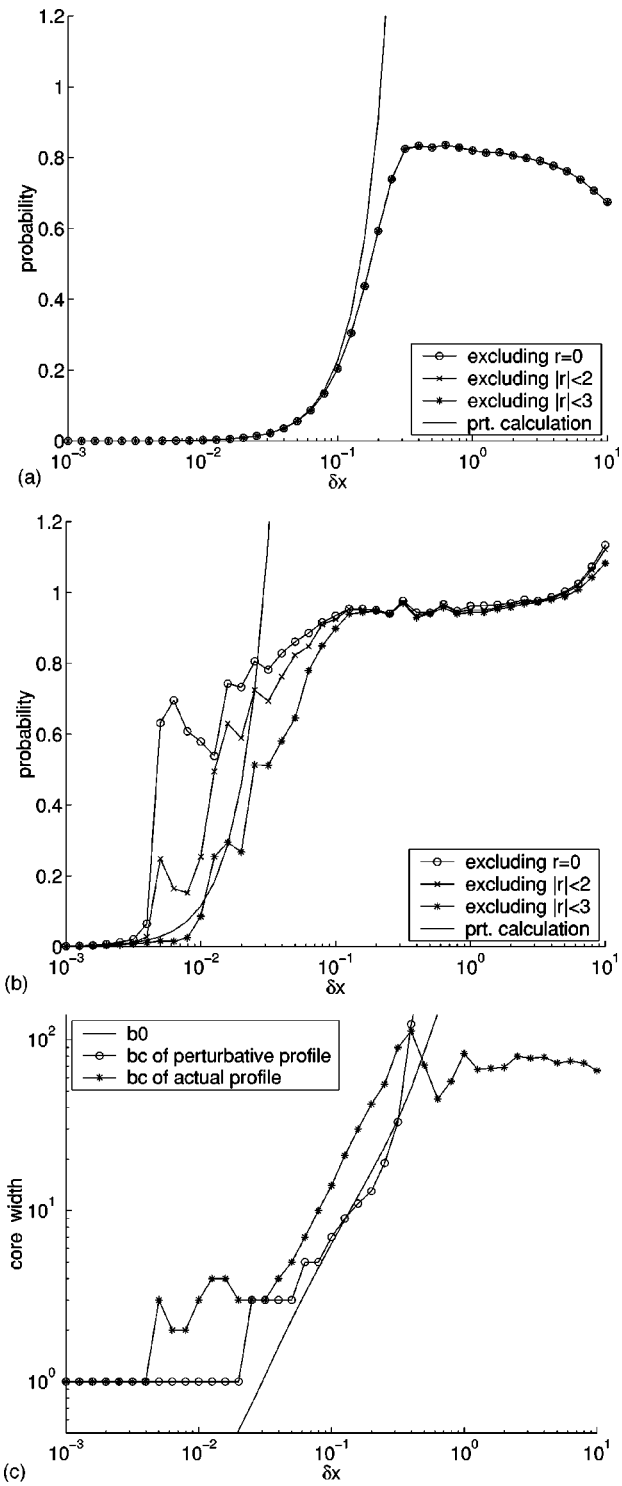


FIG. 8. (a) and (b) are $p(\delta x)$ for the rotation and for the generic deformation respectively. (c) is the perturbative calculation of the core width b_0 for the generic deformation. Also displayed is the calculated b_c for $P_{\rm prt}(r)$, and the actual b_c for $P(r)$.

and related discussion there] imply that the nonuniversal structure of the core is

$$P(n-m)\approx\frac{1}{\pi}\frac{\delta E_{\rm SC}}{\delta E_{\rm SC}^2+(E_n-E_m)^2} \tag{31}$$

with $\delta E_{\rm SC}=\hbar/\tau_{\rm col}$. The definition of the collision rate $1/\tau_{\rm col}$ is similar to that of $1/\tau_{\rm bl}$. The former is the collision rate with the deformed area $A_{\rm dfr}$, and therefore it may be smaller than $1/\tau_{\rm bl}$, because $\tau_{\rm bl}$ is related to the total area A.

If, by mistake, we identified $\tau_{\rm col}$ with $\tau_{\rm bl}$, then we would get the random wave result which is derived in the Appendix. This would be an over simplification. If it were true, it would imply that for $\delta x>\delta x_{\rm NU}$ we should have got the same Lorentzian distribution in both cases (the rotation and the generic deformation). What we see, as a matter of fact, is that for the rotation (Fig. 5) we have a reasonably good agreement with Lorentzian whose width is $\delta k_{\rm SC}=1.5$, whereas for the generic deformation (Fig. 6) there is rough agreement with Lorentzian whose width is $\delta k_{\rm SC}=0.26$. The smaller width in the latter case clearly reflects having larger $\tau_{\rm col}$. For the generic deformation we also have pronounced non-Lorentzian features. Actually the global fitting to Lorentzian is quite bad. Our understanding is that these features are due to the bouncing-ball trapping: It leads to non-exponential decay of the time-dependent survival probability (see Ref. [8] for definition of the latter term), and hence to the observed non-Lorentzian features of the spreading profile.

We turn now to explain how $\delta x_{\rm NU}$ is determined. By definition it is the deformation which is required in order to expose features of the semiclassical landscape. These features start to be exposed once $\Gamma(\delta x)\sim\delta E_{\rm SC}$ which should be converted into an equivalent expression $\delta x\sim\delta x_{\rm NU}$. Thus we get

$$\delta x_{\rm NU}=\left(\frac{A_{\rm dfr}}{A_{\rm eff}}\right)^{1/2}\times\lambda_B\sim\lambda_B\,. \tag{32}$$

Here $A_{\rm dfr}$ is the geometric area of the deformation (in the sense of scattering cross section), while $A_{\rm eff}$ is the effective area of the deformation. The definition of the latter is implied by comparing Eq. (27) with Eq. (25). By rescaling $D(\mathbf{s})\mapsto\alpha D(\mathbf{s})$ and $\delta x\mapsto\delta x/\alpha$, we can make $A_{\rm eff}=A_{\rm dfr}$ by convention. For special deformations $\delta x_{\rm NU}$ coincides with $\delta x_c^{\rm qm}$ implying that we get into the non-universal regime as soon as we have a breakdown of standard perturbation theory.

Our theoretical consideration so far do not imply a total collapse of perturbation theory. We may have in principle a coexistence of nonuniversal core component and perturbative tails. Actually we see such coexistence in the lower subfigures of Fig. 7, mainly for $\delta x=0.2512$. The right peak around $\kappa=0$ is clearly nonperturbative, while the rest of the profile is in reasonable (though not very good) agreement with the perturbative calculation. We are going to explain in the next section that the total collapse of perturbation theory for $\delta x>0.3$ is actually not related at all to the appearance of nonuniversal features in the core structure. It is only circumstantial that in the hard wall limit this collapse happens as soon as we enter the nonuniversal regime.

XI. THE COLLAPSE OF PERTURBATION THEORY

A good starting for the following discussion is to consider the classical approximation for $P(r)$. Namely, in Eq. (1) one

approximates $\rho_n(Q,P)$ and $\rho_m(Q,P)$ by microcanonical distributions. A phase-space illustration of the energy surfaces which support $\rho_n(Q,P)$ and $\rho_m(Q,P)$ can be found in Fig. 1 of Ref. [8]. The classical $P(n|m)$ equals to the overlap of these surfaces.

If we were dealing with a generic system we could introduce a linearized version of the Hamiltonian $\mathcal{H}(Q,P;x) = \mathcal{H}_0(Q,P) + \delta x \mathcal{F}(Q,P;\delta x=0)$, where $\mathcal{F}(Q,P;x) = \partial\mathcal{H}/\partial x$. By definition this linearization is a good approximation provided $\delta x \ll \delta x_c^{\rm cl}$, where $\delta x_c^{\rm cl}$ is the classical correlation scale of $\mathcal{F}(Q,P;x)$ with respect to x. In the classical linear regime the classical $P(r)$ has the scaling property $P(r) = 1/\delta x \hat{P}(r/\delta x)$. See Ref. [15] for details.

An equivalent definition of $\delta x_c^{\rm cl}$ in the quantum-mechanical case is obtained by looking on the x dependence of the matrix elements of $\mathcal{F}(Q,P;x)$ in some *fixed basis*. Again we define $\delta x_c^{\rm cl}$ as the respective correlation scale. It is quite clear that for cavity with soft walls we have

$$\delta x_c^{\rm cl} = l \quad [\text{soft walls}], \tag{33}$$

where $l = E/f$ has been defined as the penetration distance upon collision. From a purely classical point of view the hard wall limit $l \to 0$ is a *nonlinear* limit. But this is not true quantum mechanically. Here we have

$$\delta x_c^{\rm cl} \sim \lambda_B \quad [\text{hard walls}] \tag{34}$$

for $\ell < \lambda_B$. The terminology ''classical correlation scale'' while referring to $\delta x_c^{\rm cl}$ becomes misleading here, but we shall keep using it anyway.

The theory of the core-tail structure [14] is valid only in the linear regime $\delta x \ll \delta x_c^{\rm cl}$. Let us assume for a moment soft walls. We can ask is $\delta x \ll \delta x_c^{\rm cl}$ a sufficient condition for having a core-tail structure? The answer is definitely not. Perturbation theory has a final collapse once the core width $b_0(\delta x)$ becomes of the order of the band width b. This defines a parametric scale $\delta x_{\rm prt}$. For $\delta x > \delta x_{\rm prt}$ the LDOS becomes purely nonperturbative. In Wigner's theory of random banded matrices this corresponds to the crossover from Lorentzian to semicircle line shape.

However, in the limit of hard walls the above mechanism of collapse becomes irrelevant because the band width is infinite ($b=\infty$). On the other hand we still have to satisfy the inequality $\delta x \ll \delta x_c^{\rm cl}$. Thus, for $\delta x > \lambda_B$ we expect a total collapse of perturbation theory, as indeed observed in the numerical study.

ACKNOWLEDGMENTS

This work was funded by ITAMP and the National Science Foundation.

APPENDIX: OVERLAP OF UNCORRELATED RANDOM WAVES

It is possible to estimate the overlap $|\langle n|m\rangle|^2$ if we assume that $|n\rangle$ and $|m\rangle$ are *uncorrelated* random superpositions of plane waves: A random superposition of plane waves is characterized by the correlation function

$$\langle \psi_R(\mathbf{x}_1)\psi_R(\mathbf{x}_2)\rangle = \frac{1}{\mathcal{V}}\cos(k|\mathbf{x}_2-\mathbf{x}_1|),$$

where $\cos(kr) \equiv \langle \exp(i\mathbf{k}\cdot\mathbf{r})\rangle_\Omega$ is a generalized Bessel function (for further details see Appendix D of Ref. [14]). Assuming that the wave function of $|n\rangle$ is uncorrelated with the wave function of $|m\rangle$ one obtains

$$\langle |\langle n|m\rangle|^2\rangle = \left(\frac{1}{\mathcal{V}}\right)^2 \int\int \cos(k_n|\mathbf{x}_2-\mathbf{x}_1|) \times \cos(k_m|\mathbf{x}_2-\mathbf{x}_1|)d\mathbf{x}_1 d\mathbf{x}_2.$$

The integration is over the whole volume of the cavity. Using the definition of the cos function we can cast this expression into the form $\langle |\langle n|m\rangle|^2\rangle = \langle f(\mathbf{q})\rangle_q$ where the average is over the difference $\mathbf{q} = (\mathbf{k}_2 - \mathbf{k}_1)$, with all possible orientations for $|\mathbf{k}_1| = k_m$ and for $|\mathbf{k}_2| = k_n$. The function $f(\mathbf{q})$ is defined as

$$f(\mathbf{q}) = \left|\int e^{i\mathbf{q}\cdot\mathbf{x}} d\mathbf{x}\right|^2.$$

The function $f(\mathbf{q})$ depends mainly on $q = |\mathbf{q}|$. We can obtain an estimate for $f(q)$ by considering a spherical cavity $|\mathbf{x}| < R$ with the same volume. Using spherical coordinates it is straightforward to obtain

$$f(q) \approx \left(\mathrm{FT}\left[\frac{1}{d-1}(R^2-x^2)^{(d-1)/2}\right]\right)^2,$$

where FT is a Fourier transform from x to q. For $q \ll 1/R$ we have simply $f(q) = (\mathcal{V})^2$. For $q \gg 1/R$ we have $f(q) = (\mathcal{V})^2/(Rq)^{d+1}$. This is because of the singularity at $x = \pm R$. The average over q can be done using again spherical coordinate. We have $q \approx [(\Delta k)^2 + 2k^2(1-\cos\theta)]^{1/2}$, where $\Delta k = |k_n - k_m|$, and we can transform the $d\theta$ integration into a dq integration

$$\langle f(\mathbf{q})\rangle_q = \frac{\Omega_{d-2}}{\Omega_{d-1}} \times \int \frac{q\,dq}{k^2}\left(\frac{\sqrt{[q^2-(\Delta k)^2][(2k)^2-q^2]}}{2k^2}\right)^{d-3} f(q),$$

where Ω_d is the solid angle in d dimensions. There is no point in trying to carry an exact integration. Rather, it is important to observe that a practical approximation for the overlap is

$$\langle |\langle n|m\rangle|^2\rangle \approx \left(\frac{1}{kR}\right)^{d-1} \frac{1}{1+[R\cdot(k_n-k_m)]^2}.$$

In the latter expression we have neglected the d-dependent normalization prefactor. It is a ''practical'' approximation since it gives the correct behavior for both small and large values of Δk. The interpolation around $\Delta k \sim 1/R$ cannot be trusted.

[1] The classical approximation for $P(n|m)$ works well for ergodic eigenstates. As explained later in the text we actually consider the averaged profile. Namely, $P(n|m)$ is regarded as a function of $r=n-m$, and it is averaged over the reference state m. With this procedure the effect of the minority of nonergodic eigenstates can be neglected in the classical limit.

[2] The two parametric scales of Wigner's RMT model correspond to δx_c^{qm} and δx_{prt} of Table I. Consequently there are three parametric regimes in Wigner's theory. These are the standard perturbative regime ($\delta x \ll \delta x_c^{\mathrm{qm}}$), the Lorentzian regime ($\delta x_c^{\mathrm{qm}} \ll \delta x \ll \delta x_{\mathrm{prt}}$), and the semi-circle regime ($\delta x \gg \delta x_{\mathrm{prt}}$). As an artifact of this RMT model, there actually exists a forth regime (Anderson strong localization regime) that in the present context does not have a semiclassical analog.

[3] As explained in Ref. [8] the generic hierarchy is $\delta x_{\mathrm{NU}} \ll \delta x_{\mathrm{prt}} \ll \delta x_{\mathrm{SC}} \ll \delta x_c^{\mathrm{cl}}$. This hierarchy is realized in the classical limit (small $\hbar$) where we have soft walls (see Sec. II). The hard wall limit is nongeneric, and these four parametric scales coincide. In the present paper we assume hard walls unless stated otherwise.

[4] G. Casati, B. V. Chirikov, I. Guarneri, and F. M. Izrailev, Phys. Rev. E **48**, R1613 (1993); Phys. Lett. A **223**, 430 (1996).

[5] V. V. Flambaum, A. A. Gribakina, G. F. Gribakin, and M. G. Kozlov, Phys. Rev. A **50**, 267 (1994).

[6] G. Casati, B. V. Chirikov, I. Guarneri, and F. M. Izrailev, Phys. Lett. A **223**, 430 (1996); F. Borgonovi, I. Guarneri, and F. M. Izrailev, Phys. Rev. E **57**, 5291 (1998).

[7] E. Heller in *Chaos and Quantum Physics*, Proceedings of Session LII of the Les-Houches Summer School, edited by A. Voros and M.-J. Giannoni (North-Holland, Amsterdam, 1990).

[8] D. Cohen and E. J. Heller, Phys. Rev. Lett. **84**, 2841 (2000).

[9] C. Cohen-Tannoudji, J. Dupont-Roc, and G. Grynberg, *Atom-photon Interactions: Basic Processes and Applications* (Wiley, New York, 1992).

[10] E. Wigner, Ann. Math. **62**, 548 (1955); **65**, 203 (1957).

[11] M. Feingold and A. Peres, Phys. Rev. A **34**, 591 (1986); M. Feingold, D. Leitner, and M. Wilkinson, Phys. Rev. Lett. **66**, 986 (1991); M. Wilkinson, M. Feingold, and D. Leitner, J. Phys. A **24**, 175 (1991); M. Feingold, A. Gioletta, F. M. Izrailev, and L. Molinari, Phys. Rev. Lett. **70**, 2936 (1993).

[12] D. Cohen, Phys. Rev. Lett. **82**, 4951 (1999).

[13] D. Cohen, in *Proceedings of the International School of Physics ''Enrico Fermi'' Course CXLIII*, edited by G. Casati, I. Guarneri, and U. Smilansky (IOS Press, Amsterdam, 2000).

[14] D. Cohen, Ann. Phys. (N.Y.) **283**, 175 (2000).

[15] D. Cohen and T. Kottos, Phys. Rev. E (to be published).

[16] M. V. Berry and M. Wilkinson, Proc. R. Soc. London **392**, 15 (1984).

[17] A. Barnett, D. Cohen, and E. J. Heller, Phys. Rev. Lett. **85**, 1412 (2000).

[18] A. Barnett, D. Cohen, and E. J. Heller, J. Phys. A (to be published).

[19] E. Vergini and M. Saraceno, Phys. Rev. E **52**, 2204 (1995); E. Vergini, Ph. D. thesis, Universidad de Buenos Aires, 1995.

[20] A. Barnett, Ph.D. thesis, Harvard, 2000.

[21] In order to find the average profile we divide the κ axis into small bins. We find the average $P(n|m)$ value in each bin. In order to get $P(r)$ we resample the average profile. The distance between the sampling points is equal to the mean level spacing. We are not interested in features on scale of the mean level spacing. Therefore the transformation from the continuous variable κ to the integer variable r should not be considered as problematic.

Mesoscopic scattering in the half plane: Squeezing conductance through a small hole

A. H. Barnett, M. Blaauboer, A. Mody, and E. J. Heller
Department of Physics, Harvard University, Cambridge, Massachusetts 02138
(Received 26 September 2000; revised manuscript received 9 February 2001; published 4 June 2001)

We model the two-probe conductance of a quantum-point contact (QPC), in linear response. If the QPC is highly nonadiabatic or near to scatterers in the open-reservoir regions, then the usual distinction between leads and reservoirs breaks down and a technique based on scattering theory in the full two-dimensional half plane is more appropriate. Therefore we relate conductance to the transmission *cross section* for incident plane waves. This is equivalent to Landauer's formula using a radial partial-wave basis. We derive the result that an arbitrarily small (tunneling) QPC can reach a *p*-wave channel conductance of $2e^2/h$ when coupled to a suitable reflector. If two or more resonances coincide the total conductance can even exceed this. This relates to recent mesoscopic experiments in open geometries. We also discuss reciprocity of conductance, and the possibility of its breakdown in a proposed QPC for atom waves.

DOI: 10.1103/PhysRevB.63.245312 PACS number(s): 72.10.−d, 73.23.−b, 42.25.Fx, 32.80.Qk

I. INTRODUCTION

The quantum-point contact[1,2] (QPC) has played a central role in the understanding of mesoscopic conductance. It is the simplest example of a two-dimensional electron gas 2DEG system where the quantum coherent nature of the electron controls the bulk transport properties. The Landauer-Büttiker (LB) formalism[3,4,2] reduces the calculation of quantum conductance in the linear response regime to the evaluation of single-particle wave-function transmission amplitudes. Traditionally, these amplitudes are measured between traveling wave basis states in the ''leads.'' Far from the scattering system the leads have constant profiles of finite width, and support a finite number of transverse modes (channels). Eventually it is assumed that the leads are impedence matched (that is, without reflection) into ''reservoirs'' that act as thermalized sources of electrons at their respective potentials; these potentials are taken to reflect the measured bias voltage. Such theoretical constructs have been remarkably successful at describing transport phenomena, for instance conductance quantization,[5,1,2] because the scattering systems involved have generally had good lead-to-reservoir matching.

We consider ''open'' two-terminal mesoscopic systems, namely, those where a QPC is *nonadiabatic* (possessing rapid longitudinal variation in transverse profile[6,1,7]) and has short or nonexistent leads (for instance, if it suddenly abutts onto the ''reservoir'' regions), or those where there can be scattering off nearby objects in the ''reservoir'' region. We call such systems ''open'' because the fully two-dimensional (2D) nature of the ''reservoirs'' (i.e., the surrounding semi-infinite regions of free space) is important, and therefore they cannot be modeled using the quasi-1D approach described above. This includes a variety of recent mesoscopic experiments, for example, the combination of QPC's with nearby resonator structures[8] or with a nearby depletion region caused by a moveable atomic force microscope (AFM) tip.[9] It also includes any QPC system where elastic backscattering from disorder in the reservoirs is significant,[10] or generally where the lead-reservoir matching is bad. In such systems, the conventional quasi-1D picture does not apply: the scattering system is not coupled to leads in the usual sense, indeed the distinction between leads and reservoirs is no longer clear.[11] The main aim of the present work is to introduce a 2D scattering theory approach that can handle such systems, and to apply it to the calculation of the maximum conductance of an open resonator structure of experimental relevance.

We imagine a geometry where a 2DEG exists in two semi-infinite half-plane regions, separated by an impenetrable potential barrier that we align with the *y* axis [see Fig. 1(a)]. Our general ''QPC scattering system'' is any gap in this barrier that allows coupling of the wave function on the left and right sides. This gap can be defined by an arbitrary form of the elastic potential, and may include other nearby scattering objects or disorder [which would all be placed

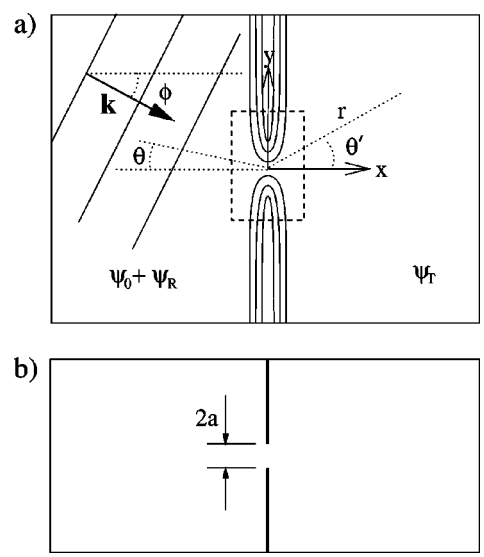


FIG. 1. Schematic QPC geometry in 2D: (a) general point-contact scatterer coupling two semi-infinite regions of free space. The solid curves are contours of an elastic scattering potential $V(\mathbf{r})$. The ''system'' size L (dashed box) we take to be the region where $V(\mathbf{r})$ has not yet reached its asympotic form (which is zero apart from a *y*-invariant profile around the *y* axis). Also shown are an incoming plane wave, and the coordinate system. (b) The idealized ''slit'' aperture in a thin, hard wall considered in Sec. IV.

 ©2001 The American Physical Society

within the box shown in Fig. 1(a)]. The only important limitation is that this coupling region (the ''system'') be of finite y extent, so that electrons that leave the system do so via a well-defined terminal: either the left ($x<0$) or the right ($x>0$). We also assume that the system size L is much smaller than both the dephasing length l_ϕ and the momentum relaxation (elastic scattering) length l_e. The former requirement allows treatment using a coherent wave function across the system; the latter allows free-space elastic scattering concepts to be applied. We will stay within the noninteracting quasiparticle picture, consider zero-applied magnetic field, and assume spin degeneracy of 2 throughout.

The conventional distinction between ''reservoir'' and ''lead'' is no longer applicable, however at short distances outside the system ($r>L$ but $r\ll l_\phi$ and $r\ll l_e$) the two semi-infinite free-space regions behave like leads, since they support scattering-free ''channels'' (see Sec. III). At large distances the *same* regions behave as reservoirs: for $r\gg l_e$ ergodicity ensures that the momentum distribution is uniform in angle, and for $r\gg l_\phi$ the energy is redistributed to ensure equilibrium at the relevant (experimentally measured) chemical potential of each terminal. In the intermediate region, there is a broad crossover from lead to reservoir.

In this work we first derive a general relation between transmission cross section (a concept we define using scattering in the half plane) and conductance for this open geometry, in Sec. II. In Sec. III we show that partial-wave-type states, defined in the half-plane regions, can take the place of transverse lead modes in the Landauer formula. In Sec. IV we discuss the maximum conductance through an idealized, highly nonadiabatic QPC (a hole in a thin hard wall) that is reached when a resonator is placed on one side of the QPC. We find a universal result, namely, a single conductance quantum, *regardless* how small the hole is. This illuminates the findings of a recent experiment[8] in such an open geometry. In Sec. V we discuss attempts to exceed this universal quantum of conductance through a single channel. A reciprocity relation for cross section is derived in Sec. VI, and the possibility of breaking this reciprocity, due to a nonthermal reservoir occupation, is described. We discuss an application to matter-wave ''conductance'' through a 3D QPC. We conclude in Sec. VII.

II. CONDUCTANCE IN TERMS OF CROSS SECTION

We consider scattering of a single-quasiparticle wave function from the general two-terminal system described in the Introduction [see Fig. 1(a)]. The Hamiltonian is $\mathcal{H}=-(\hbar^2/2m)\nabla^2+V(\mathbf{r})$, for a quasiparticle mass m. The elastic scattering potential $V(\mathbf{r})$ completely defines the system. We imagine a monochromatic unit plane wave $\psi_{\rm I}=e^{i\mathbf{k}\cdot\mathbf{r}}$ incident from the free-space left-hand region.[12] The wave vector is $\mathbf{k}\equiv(k,\phi)$ in polar coordinates, ϕ being the angle of incidence. The free-space wave-vector magnitude is taken as $k=k_{\rm F}$ (corresponding to a total energy $E=\hbar^2k^2/2m$ equal to the Fermi energy), unless stated otherwise.

We are at liberty to choose our definition of the ''unscattered'' wave ψ_0. We take it to be the wave function which would result from reflection of the incident wave off a wall uniform in the y direction. We can imagine creating such a wall by replacing the system box shown in Fig. 1(a) by the surrounding y-invariant wall profile. Note that ψ_0 exists only on the left side. In the left free-space region it is

$$\psi_0=e^{i(k_xx+k_yy)}-e^{\{i(-k_xx+k_yy+\gamma_{\mathbf{k}})\}}, \tag{1}$$

where the first term is $\psi_{\rm I}$, and the angle-dependent reflection phase $\gamma_{\mathbf{k}}$ of the second term depends on both (k,ϕ) and the wall profile.[13] Upon introduction of our true system potential, the full wave function becomes

$$\psi\equiv\psi_0+\psi_{\rm R}+\psi_{\rm T}, \tag{2}$$

where the change in reflected wave $\psi_{\rm R}$ exists only on the left side, and the new transmitted wave $\psi_{\rm T}$ exists only on the right. These scattered waves have the asymptotic ($r>L$ and $kr\gg1$) forms of 2D scattering theory,[14]

$$\psi_{\rm R}=f_{\rm R}(\theta)\frac{e^{ikr}}{\sqrt{r}},\quad \psi_{\rm T}=f_{\rm T}(\theta')\frac{e^{ikr}}{\sqrt{r}}. \tag{3}$$

See Fig. 1(a) for definitions of θ and θ'.

The transmission cross section $\sigma_{\rm T}(k,\phi)$ is the ratio of $\Gamma_{\rm T}$, the transmitted particle flux (number per unit time), to $j_{\rm I}$, the incident particle flux per unit length normal to the incident beam:

$$\sigma_{\rm T}(k,\phi)\equiv\frac{\Gamma_{\rm T}}{j_{\rm I}}. \tag{4}$$

Physically, $\sigma_{\rm T}(k,\phi)$ is the length required of an aperture-oriented normal to the incident beam in order to transmit an equivalent flux of classical particles. [Note that $\sigma_{\rm T}(k,\phi)$ is proportional to the *injection distribution*[1] that can be measured in mesoscopic systems.[15]] It depends on the incident angle because $V(\mathbf{r})$ has no radial symmetry. $j_{\rm I}$ is the magnitude of the incoming probability flux density vector $\mathbf{j}\equiv(\hbar/m)\mathrm{Im}[\psi_{\rm I}^*\boldsymbol{\nabla}\psi_{\rm I}]$, which for a unit wave gives $j_{\rm I}=v$, the particle speed. The transmitted flux is defined as

$$\Gamma_{\rm T}\equiv\int dl\ \hat{\mathbf{n}}\cdot\mathbf{j}=\frac{\hbar}{m}\int dl\ \hat{\mathbf{n}}\cdot\mathrm{Im}[\psi_{\rm T}^*\boldsymbol{\nabla}\psi_{\rm T}], \tag{5}$$

where the line integral encloses the entire transmitted wave, and the (rightwards pointing) surface normal is $\hat{\mathbf{n}}$. Applying this and Eq. (4) to the asymptotic form gives

$$\sigma_{\rm T}(k,\phi)=\int_{-\pi/2}^{\pi/2}d\theta'|f_{\rm T}(\theta')|^2, \tag{6}$$

familiar from scattering theory apart from the restriction to the right half plane. There is a corresponding form

$$\sigma_{\rm R}(k,\phi)=\int_{-\pi/2}^{\pi/2}d\theta|f_{\rm R}(\theta)|^2, \tag{7}$$

for the reflective cross section (removal from the unscattered wave without being transmitted).

We will calculate the conductance by assuming the chemical potential is slightly higher on the left side than the right, and as is usual[1,4] consider only the left-to-right trans-

port of the states in this narrow energy range. We take the left region to be a large ($\gg l_\phi$) closed region of area A containing single-particle states, and find their decay rate through the QPC into the right side. Semiclassically each single-particle state occupies a phase-space volume h^d, where we have $d=2$. Therefore the phase-space density in the 2DEG Fermi sea is $2/h^2$ where the factor of 2 comes from the spin degeneracy. We can project this density onto momentum space in order to find the effective number of plane-wave states impinging on the wall:[16] this corresponds to a uniform density of states in **k** space given by

$$\rho(k,\phi)k\ dk\,d\phi=\frac{A}{2\pi^2}k\ dk\,d\phi. \tag{8}$$

Each state has an amplitude $A^{-1/2}$ due to the requirement of unity area normalization in the left region, so has incoming flux density $j_{\mathrm{I}}=v/A$. Substituting this into Eq. (4) gives the decay rate of a state i as

$$\Gamma_{\mathrm{T}}^{(i)}=\frac{v}{A}\sigma_{\mathrm{T}}(k_i,\phi_i). \tag{9}$$

We can now sum the decay rates of all the left-hand states in a given wave-vector range k_{F} to $k_{\mathrm{F}}+\delta k$, to get the current

$$\delta I=e\sum_i\ \Gamma_{\mathrm{T}}^{(i)}=\frac{ev}{A}\int_{-\pi/2}^{\pi/2}d\phi\int_{k_{\mathrm{F}}}^{k_{\mathrm{F}}+\delta k}k\ dk\,\rho(k,\phi)\sigma_{\mathrm{T}}(k,\phi)$$

$$=\frac{ev\ k_{\mathrm{F}}\delta k}{2\pi^2}\int_{-\pi/2}^{\pi/2}d\phi\ \sigma_{\mathrm{T}}(k_{\mathrm{F}},\phi), \tag{10}$$

where the last step incorporated the linear-response assumption that σ_{T} is constant over the range δk.

When a potential difference δV is applied across the QPC, the energy range carrying current is $\delta E=e\,\delta V$, which we can equate with $\hbar v\,\delta k$ using the dispersion relation. This can be used with Eq. (10) to write the conductance

$$G\equiv\frac{\delta I}{\delta V}=\frac{2e^2}{h}\frac{1}{\lambda_{\mathrm{F}}}\int_{-\pi/2}^{\pi/2}d\phi\ \sigma_{\mathrm{T}}(k_{\mathrm{F}},\phi) \tag{11a}$$

$$=\frac{2e^2}{h}\frac{k_{\mathrm{F}}}{2}\langle\sigma_{\mathrm{T}}\rangle_\phi, \tag{11b}$$

where the particle wavelength is $\lambda_{\mathrm{F}}\equiv 2\pi/k_{\mathrm{F}}$. The latter form is written in terms of the angle-averaged cross section at the Fermi energy. The weighting of this average is uniform because of the ergodic assumption that incoming states are uniformly distributed in angle.

Equation (11) is a key result of this paper (an independent derivation is given by Barnett[17]). Like the Landauer formula, it directly connects conductance and scattering. In a scattering measurement from the left side, σ_{T} appears to be the QPC's inelastic cross section (since the transmitted waves never return to this side). In a current measurement the corresponding conductance is given by Eq. (11). Our derivation was for temperature $T=0$, but it applies at a finite T as long as σ_{T} does not change significantly over the energy range $k_{\mathrm{B}}T$. This can be seen by generalizing the above to include integration over the Fermi distribution.

In the limit where a QPC is adiabatic, its conductance is known to be quantized:[5,1,2] $G=(2e^2/h)N$, where N is the integer number of open channels at the Fermi energy. Looking at Eq. (11a), this corresponds to quantization of the angular integral of the cross section in units of λ_{F}.

III. PARTIAL-WAVE CHANNEL MODES FOR A TWO-TERMINAL SYSTEM

In free-space scattering theory, partial waves form a basis in which to decompose the asymptotic ($r\to\infty$) form of the full wave function ψ into incoming and outgoing states of definite angular momentum l. In 2D the basis functions are the cylindrical solutions to the free-space wave equation; the S matrix that takes incoming to outgoing waves can then be written in this basis.[14] Because there is only a single set of incoming channels and a single set of outgoing channels, this is equivalent to a scattering system (a stub) connected to a single lead, with an infinite number of open-channel modes. This contrasts the open two-terminal geometry we study, where we need to account for two new related facts: (1) in the $r\to\infty$ limit the potential V no longer preserves angular momentum, and (2) there are now distinct ways the particle can enter and exit the system, via different leads.

We define a ''half-plane partial-wave basis'' as the subset of the cylindrical free-space solutions that go to zero on the entire y axis. This gives independent basis functions existing on either the left or right side of the y axis. The basis is expressed in terms of Hankel functions[18] on either side

$$\phi_l^{-L}(kr)\equiv H_l^{(2)}(kr)\sin\left[l\left(\frac{\pi}{2}-\theta\right)\right]$$

$$\phi_l^{+L}(kr)\equiv H_l^{(1)}(kr)\sin\left[l\left(\frac{\pi}{2}-\theta\right)\right]$$

$$\phi_l^{-R}(kr)\equiv H_l^{(2)}(kr)\sin\left[l\left(\frac{\pi}{2}-\theta'\right)\right]$$

$$\phi_l^{+R}(kr)\equiv H_l^{(1)}(kr)\sin\left[l\left(\frac{\pi}{2}-\theta'\right)\right], \tag{12}$$

where on the left (L) side θ is the angle from the negative x axis and on the right (R) side θ' is the angle from the positive x axis [see Fig. 1(a)]. The channel index is $l=1,2,\cdots\infty$, and $+(-)$ refers to outgoing (incoming) traveling waves. We note that the s wave $l=0$ is excluded because of the y-axis barrier, leaving the first channel as the p wave $H_1(kr)\cos(\theta)$. Assuming the width of the barrier is finite and constant as $|y|\to\infty$ [see Fig. 1(a)], then any wave function in the $r\to\infty$ limit can be written as a sum of the above basis functions. The separability of this basis in (r,θ) is directly analogous to the separability of conventional (constant-width) lead basis states[3] into a product of transverse modes and longitudinal traveling waves.

Our basis (12) is chosen such that unit-amplitude coefficients carry equal fluxes in all incoming and outgoing chan-

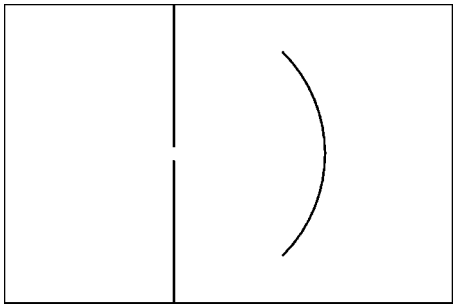


FIG. 2. A tunneling-regime QPC combined with a nearby circular reflector, forming a stable resonant cavity open at the sides.

nels, so flux conservation implies the unitarity of the S matrix when written in this basis. As with a conventional transverse lead mode basis, the familiar Landauer formula

$$G=\frac{2e^2}{h}\mathrm{Tr}(t^\dagger t), \tag{13}$$

holds.[19,20,4,17] The transmission matrix t is defined by $q_l^+ = \Sigma_m t_{lm} p_m^-$, where the outgoing (incoming) amplitude coefficients are p_l^+ (p_l^-) on the left and q_l^+ (q_l^-) on the right. Note that it is possible to ''mix and match'' different basis set types (for instance define a transmission matrix between transverse lead modes on the left side and partial-wave modes on the right), as long as equal-flux normalization, and transverse orthogonality, are preserved.

IV. POINT CONTACT COUPLED TO A RESONATOR

Figure 2 illustrates a QPC-plus-reflector system whose conductance has been experimentally measured.[8] The circular arc reflector and the vertical wall together form a cavity that can support long-lived resonances; the energy of these resonances can be swept by sweeping the reflector gate voltage. The classical condition[8] for stability of the cavity modes is that the arc center must lie at, or to the left of the wall ($x=0$). The cavity modes are coupled to the left terminal via the QPC, and to the right terminal via leakage of the modes out through the cavity top and bottom. The system is interesting because it is ''open'' in the sense that it has no Coulomb blockade,[1] but ''closed'' in the sense that the dwell time is much greater than the ballistic time (the resonances are long lived). It has also been studied recently in our laboratory using microwave measurements.[21]

The actual potential in a mesoscopic experiment differs from the illustration: it has soft walls (on the scale $1/k_F$), it may have deviations from the circle due to lithographic error, and it has modulations of the background potential due to elastic disorder.[8] However, we will not be interested in details of the resonator on the right-hand side. Rather, we will adopt the view of a 2D scattering-theorist ''looking'' from the left-hand side. In this section we discuss the maximum conductance of this system, when the ''bare'' QPC (i.e., without the reflector) is in the tunneling regime (conductance $\ll 2e^2/h$).

We use an idealized slit QPC model [see Fig. 1(b)] in which the potential V is zero everywhere except along a hard, thin wall where it is taken as infinite. The QPC is a gap in the wall of size $2a$. This model is highly nonadiabatic (see Ref. 17 for a review of its transmission properties). The hard wall simplifies the treatment of the left-hand side scattering problem, and we do not believe it alters our basic conclusion. We consider the ''unscattered'' wave to be the incident plus reflected wave Eq. (1) when the QPC is closed ($a=0$). This we expand in Bessel functions,

$$\begin{aligned}\psi_0(\mathbf{r}) &= e^{[i(k_x x+k_y y)]} - e^{[i(-k_x x+k_y y)]}\\ &= -4iJ_1(kr)\cos(\theta)\cos(\phi) + \text{higher order terms.}\end{aligned} \tag{14}$$

The first term in the expansion is the incoming plus outgoing p wave, which in the tunneling limit will dominate in our consideration of the absorption.[17]

Now we open the slit, and replace $2J_1(kr)$ in the above by $H_1^{(2)}(kr)+e^{2i\delta}H_1^{(1)}(kr)$, where δ follows the usual definition of partial-wave phase shift.[14] The closed slit corresponds to $\delta=0$. An open slit leading into a closed resonator (imagine extending the arc in Fig. 2 to seal off the entire right side), in the case of infinite dephasing length, corresponds to $\delta=$ real, and would appear from the left side as an elastic dipole scatterer. An open slit with an open resonator corresponds to complex δ with positive imaginary part, and would appear as a general inelastic dipole scatterer. Therefore transmission though the QPC appears, to an observer on the left side, to be *absorption* of incident waves. σ_T is interpreted as an ''inelastic'' cross section (since exiting the right-hand terminal is equivalent to leaving in a new channel), and σ_R as an ''elastic'' one. $\sigma_T(k,\phi)$ can be found from integrating the net incoming flux [as in Eq. (5)] of the total wave function on the *left* side. Substitution into Eq. (4) then gives $\sigma_T(k,\phi)=4/k(1-|e^{2i\delta}|^2)\cos^2(\phi)$. For $\delta\to i\infty$ the maximal cross section is reached,

$$\sigma_{T,\max}(k,\phi)=\frac{4}{k}\cos^2(\phi). \tag{15}$$

This corresponds to an effective classical ''area'' (size) $a_{\mathrm{eff}} = \lambda_F/2$. This is analogous to the fact[22] that in 3D the effective area of an arbitrarily small electromagnetic dipole aerial can be of order λ^2. To an observer on the left side who was able to ''see'' the electron waves living in the energy range $e\delta V$ responsible for conductance, the QPC would stand out as a ''black dot'' of size $\sim\lambda_F$ against the surrounding uniform ''gray'' thermal luminosity reflected in the vertical wall mirror.

The associated maximum conductance is found easily using Eqs. (15) and (11) to be

$$G_{\max}=\frac{2e^2}{\hbar}, \tag{16}$$

the universal quantum of conductance (for 2 spin channels), independent of the size of the QPC hole, even for an *arbitrarily small* hole ($ka\to 0$). This universal resonant-tunneling maximum conductance was first found numerically;[23,24,1] however our system differs from those of

Xue and Lee[23] and Kalmeyer and Laughlin[24] because the resonance does not involve transmission though an *isolated* (zero-dimensional) quantum dot. The dramatic increase over the conductance of the bare QPC [which vanishes as $(ka)^4$, see Ref. 17] runs counter to the naive classical expectation, namely, that the reflector would *decrease* the left-to-right flow of electrons because it sends back into the QPC particles that would otherwise exit to the right.

How do we know that it is possible to build a resonant geometry that corresponds to $\delta \to i\infty$? The reflector can be described by r, the amplitude with which it returns an outgoing p wave back to the QPC as an incoming p wave. If $|r|^2 = 1 - |t_{11}|^2$, where the p-wave transmission of the QPC is t_{11} as defined in Sec. III, then the p-wave channel becomes a 1D Fabry-Perot resonator with mirrors of matched reflectivity. Sweeping the round-trip phase then produces peaks of complete transmission (corresponding to complete p-wave absorption on the left side). The ratio of peak separation to peak width is the quality factor $Q \sim 1/|t_{11}|^2$. Such peaks, with heights much greater than the bare tunneling QPC conductance, were observed in the experiments of Katine *et al.*[8] However, Eq. (16) has not yet been tested quantitatively because of the difficulty of matching the Fabry-Perot reflectivities in a real 2DEG experiment. Note that the maximum conductance (16) also follows immediately from the Landauer formula when we realize that there can be complete transmission of the incoming $l=1$ channel state (from Sec. III).

An interesting possibility arises when we realize[17] that higher l channels are still *slightly* transmitted by the bare QPC, when $ka \ll 1$, even though they are increasingly evanescent. If the resonator has a high enough reflectivity for these modes, then additional Fabry-Perot conductance peaks will be produced.[23,25] The peaks may be extremely narrow, but can carry a full quantum of conductance because they can transmit another incoming l channel. By careful arrangement of the cavity, one or more of these peaks could be brought into conjunction with an already-existing $l=1$ peak at the Fermi energy. (For instance, the $l=1$ and $l=2$ resonances are in different symmetry classes in Fig. 2 so there can be an exact level crossing). Therefore, we have the surprising result that, in theory, a conductance of $(2e^2/h)n$ can pass through an arbitrarily small QPC hole if n resonances (from n different channels) coincide at the Fermi energy. However, due to their extremely small width, such large conductance peaks are unlikely to be observable in a real mesoscopic tunneling QPC due to finite dephasing length and finite-temperature smearing.[1]

Finally, we should not overlook the fact that our expressions for partial cross sections are a factor of 4 greater than those conventionally arising in 2D scattering theory from a radial potential,[14] because we are measuring cross section on the reflective boundary of a semi-infinite half plane. For instance, the maximum inelastic partial cross section for a single channel in free space[14] is $\sigma_r = 1/k$, compared to our maximum ''inelastic'' cross section per channel Eq. (15). Similarly, the maximum elastic result in free space is $\sigma_e = 4/k$, compared to our maximum (normal-incidence) ''elastic'' cross section per channel $\sigma_{\mathrm{R,max}} = 16/k$. This latter case occurs when $\delta = (\text{integer} + \frac{1}{2})\pi$.

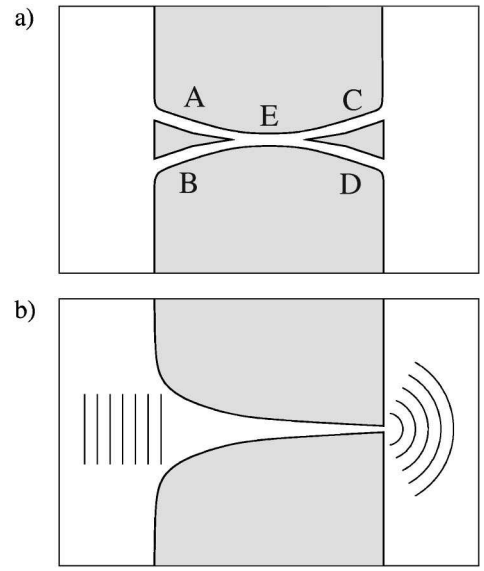


FIG. 3. (a) An attempt to increase conductance through a single channel by multiple connections feeding from the reservoirs. All channels are single mode and sufficiently long that the evanescent tunneling of higher modes is negligible. (b) An illustrative hard-walled exponential horn system that has differing acceptance angles on each side: very narrow on the left, and very wide on the right. Such a mesoscopic 2DEG system would exhibit symmetric conductance, however, in an atom beam context the conductance can become unsymmetric.

V. WHAT IS THE MAXIMUM CONDUCTANCE OF A SINGLE QUANTUM CHANNEL?

The surprising theoretical results of the previous section might lead one to question the conductance limit $2e^2/h$ for a single quantum channel (by which we mean a single transverse mode for which the longitudinal degree of freedom is a 1D Fermi gas; this includes both conventional and partial-wave basis sets). For this gedanken-experiment we will consider conventional electron waveguides that are single mode and long enough that evanescent waves are negligible, but which are also $\ll l_\phi$. We try to encourage more current to pass down a single-mode channel (E) by connecting it to a reservoir via multiple routes (A,B,C,D), as shown in Fig. 3(a), where two routes are used on each side. It is possible to match the junctions so that a wave entering down A,B,C, or D has no reflection back along the same lead. In this case we might guess that the hypothetical left-side observer (from the previous section) would see the single-mode entrances to guides A and B as two ''black dots,'' giving twice the effective absorption cross section, and therefore infer a conductance of twice $2e^2/h$. We might also justify this by saying that waves traveling down A and B will meet and continue down E, and since they have no particular phase relation, their currents will add to give a doubled current through E, as would be necessary.

However there is a fundamental flaw in the above reason-

ing. The ABE junction can be designed so that if waves come down A and B in phase, they will be adiabatically transformed into the lowest transverse mode of E, so will propagate through to the right side without reflection, carrying a current of twice that of a usual single-mode guide. However, if A and B are π out of phase, the same adiabatic transformation must result in the second transverse mode, which is evanescent. So this latter wave will reflect perfectly back out of the left side, and carry no current. Plane waves are impinging from the left reservoir uniformly over all angles, and because of the $>\lambda$ separation of the entrances, an average over angles gives an average over relative phase in A and B. Thus we are left with no increase above the single-channel conductance. This property of the ABE junction is not merely practical; rather, it is easy to show that its 3×3 S matrix cannot be unitary if a junction is to couple both $A\rightarrow E$ and $B\rightarrow E$ with unity transmissions. Such an appealing junction is therefore ruled out on the grounds of flux conservation. A consequence is that the entrances to A and B can at most appear ''half black'' to the observer, due to waves that enter A then exit B and vice versa.

This suggests another way to try and defeat the conductance limit: direct the incoming plane waves in a narrow enough angular distribution so that waves *always* come down A and B in phase, and this will double the conductance. (This is similar to experiments[26] where the series resistance of two QPC's was found to be less than the sum of the individual QPC resistances, because collimation at the exit of the first QPC illuminated the second with a narrow beam, increasing its conductance). However, this beam is no longer a *thermal* occupation of incoming states. This illustrates the inextricable link between thermal Fermi occupation of reservoir states and the universal quantum of conductance. At $T=0$, thermal occupation at a given chemical potential difference implies that *all* quantum states lying in the appropriate energy range are filled in the left reservoir and empty in the right. Semiclassically, this corresponds to a uniform distribution in phase space, or when projected into momentum states, uniform in angle, as exemplified by Eq. (11). The semiclassical viewpoint allows one to see that since transformations in phase space cannot change the phase-space density (Liouville's theorem), the universal conductance per quantum channel cannot be changed. This reminds us that unitarity in quantum mechanics is analogous to Liouville's theorem in classical mechanics.

VI. RECIPROCITY AND ''CONDUCTANCE'' OF ATOM WAVES

We can ask if the conductance (11) computed using transmission of left-side reservoir plane-wave states through the QPC is equal to that using right-side reservoir states. Since the two directions correspond to opposite signs of δV, then in order to have linear response (well-defined constant G around $\delta V=0$) we would hope that they are equal. That the angular average of transmission cross section is equal from the left and right sides is not immediately apparent in a general asymmetric system. For instance, consider Fig. 3(b) that has a small acceptance angle from the left but a large from the right, therefore very different forms of the transmission cross sections $\sigma_{\mathrm{T}}^{\mathrm{L}\rightarrow\mathrm{R}}(k,\phi)$ and $\sigma_{\mathrm{T}}^{\mathrm{R}\rightarrow\mathrm{L}}(k,\phi)$.

If we assume *classical* motion then we can imagine a map from a Poincaré section (PS) (y,p_y) at a vertical slice at $x=-x_0$ to another PS (y',p_y') at $x=+x_0$. At each PS we consider only rightwards-moving $(p_x>0)$ particles, and take $x_0>L$. A certain area of phase space (y,p_y) is transmitted and is mapped to an *equal area*[27] in phase space (y',p_y'). Time-reversal invariance holds since we consider magnetic field $B=0$, so we can negate the momenta (now considering $p_x<0$) and find that the *same* phase-space area is transmitted right to left. When it is realized that the angle-averaged cross section is proportional to the transmitted phase-space area on a PS, then the symmetry of the angle-averaged *classical* cross sections follows.

The same symmetry is not obvious for quantum cross sections, but it also holds true. Comparing Eq. (11a) with Eq. (13) gives

$$\int_{-\pi/2}^{\pi/2} d\phi\ \sigma_{\mathrm{T}}^{\mathrm{L}\rightarrow\mathrm{R}}(k_{\mathrm{F}},\phi)=\lambda_{\mathrm{F}}\ \mathrm{Tr}(t^{\dagger}t), \tag{17}$$

where t is measured from left-to-right states. It is instructive to derive this directly.[17] This relation ties together the cross section and Landauer views of conductance. Time-reversal invariance and flux conservation together imply[28] that $\mathrm{Tr}(t^{\dagger}t)$ is unchanged by swapping the labeling of the leads,[1,20] thus we immediately have from Eq. (17) the reciprocity of angle-integrated quantum cross section

$$\int_{-\pi/2}^{\pi/2} d\phi\ \sigma_{\mathrm{T}}^{\mathrm{L}\rightarrow\mathrm{R}}(k,\phi)=\int_{-\pi/2}^{\pi/2} d\phi\ \sigma_{\mathrm{T}}^{\mathrm{R}\rightarrow\mathrm{L}}(k,\phi). \tag{18}$$

So in Fig. 3(b) it is now clear that the ratio of acceptance angles must be balanced by the ratio of effective areas.

We now discuss a case in which nonthermal occupation of incoming states is possible: the rapidly developing field of coherent matter-wave optics, in which potentials are defined by microfabricated structures.[29–31] There is a recent proposal[30] for observation of quantization of atomic flux through a micron-sized 3D QPC defined by the Zeeman effect potential of a magnetic field. The device is illuminated by a beam of atoms passing through a vacuum, whose angular distribution is an experimental parameter (for instance, a collimated oven source or a dropped cloud of cold atoms[32]). The atomic flux transmitted (per unit k, at wave vector k) will be $F(k)=G_{\mathrm{atom}}(k)J_0(k)$ where $J_0(k)$ is the flux incident per unit wall area, and we define the atomic ''conductance'' by

$$G_{\mathrm{atom}}(k)\equiv\int d\Omega\ w(k,\Omega)\sigma_{\mathrm{T}}(k,\Omega). \tag{19}$$

As before, the quantum-transmission cross section is $\sigma_{\mathrm{T}}(k,\Omega)$, but now there is a *weighting function* $w(k,\Omega)$ which defines the angular distribution of the incident beam.[33] The weight has the normalization $\int d\Omega\ w(k,\Omega)\cos(\theta)=1$. [All integrals over solid angle $\Omega\equiv(\theta,\phi)$ are over a range of 2π appropriate for the half-sphere.] Following the analogy of Thywissen,[30] $F(k)$ plays the role of current, $J_0(k)$ that of

bias voltage. However, the name ''conductance'' does not imply any definite chemical potential difference as in the 2DEG case. For classical particles, the ''conductance'' of an aperture of area $A_{\rm eff}$ in a thin wall is simply $G_{\rm atom}(k) = A_{\rm eff}$, regardless of the incident angular distribution. Thus $G_{\rm atom}(k)$ gives the effective area $A_{\rm eff}$ of a QPC, in an analogous fashion to $a_{\rm eff}$ in 2D.

For an integer number of quantum channels, the 2D quantization of $a_{\rm eff}$ in units of $\lambda/2$ becomes in 3D the quantization[30,1,34] of $A_{\rm eff}$ in units of λ^2/π, a result well known from work on 3D metallic point contacts.[35] As stated by Thywissen,[30] this accurate flux quantization requires the incident beam width to be much larger than the QPC acceptance angle.

Equation (19) is the matter-wave equivalent of Eq. (11a), with the important difference that it has a general weight function. Possible nonuniformity of this weight function leads to a key result: that *asymmetry* of the conductance is possible given identical illumination on either side, even though the (center of mass) motion is time-reversal invariant. For example, if the incident flux used to illuminate the horn QPC of Fig. 3(b) is narrow in angular spread, then the left-to-right conductance will be much larger than the right-to-left conductance. This contrasts with the 2DEG case where the conductance is always symmetric.

Finally, it is interesting to note that for the nonthermal incident (reservoir) distributions discussed above, the Landauer formula takes the modified form

$$G \propto \mathrm{Tr}(t^{\dagger} t \rho), \tag{20}$$

where ρ is the density matrix of the incident beam.

VII. CONCLUSIONS

Quantum-scattering theory in the 2D half plane can provide an alternative description of the mesoscopic conductance of noninteracting particles. It is especially useful in ''open'' systems (e.g., those with nearby scatterers in the reservoir regions) where the usual transverse-channel approach is inappropriate. We have considered elastic potentials in zero magnetic field, in linear response in the low-temperature limit. Conductance is proportional to the transmission cross section integrated over all incident angles, Eq. (11). We also define a half-plane partial-wave basis applicable with the usual Landauer formula, and relate this to our transmission cross section result. A difference between this and previous work is the ability to treat a direct ''leadless'' connection to the reservoir.

Using the example of a slit QPC combined with an open-cavity structure, we show that an arbitrarily small QPC can carry up to a single quantum of conductance via resonant tunneling (equal to the limit in the closed-dot resonant tunneling case). This requires a resonance at the Fermi energy. If n coincident resonances occur for different incoming channels, then n conductance quanta can in theory be achieved through this same tunneling QPC, a result that we believe has not been noted until now.

We emphasize that conductance is proportional to phase-space density of the reservoir states. Therefore the universal quantum of conductance e^2/h per spin in Fermi gas systems is a direct result of the uniform phase-space density (angular distribution) in a thermal occupation of the Fermi sea. This insight is supported by discussion of attempts to exceed this universal value. When the reservoir occupation differs from thermal, the conductance formula requires generalization: an angle-dependent weight is included in the cross section integral (19); equivalently for 2DEG systems the Landauer formula requires inclusion of the incoming ensemble (20). This result, and our approach in general, is relevant to the emerging field of matter-wave conductance by microfabricated structures (for instance, a quantum-point contact in 3D), under general illumination by atom waves. We hope this work provides new tools for the study of coherent electron and matter-wave systems.

ACKNOWLEDGMENTS

We thank Adam Lupu-Sax for his early contributions to this work, also J. Thywissen, C. Marcus, and D. Fisher for stimulating discussions. This work was supported by the National Science Foundation (U.S.A.) via Grant No. CHE-9610501, ITAMP (at the Harvard-Smithsonian Center for Astrophysics and the Harvard Physics Department), and the Netherlands Organization for Scientific Research (NWO).

[1] For a review, see C.W.J. Beenakker and H. van Houten, Solid State Phys. **44**, 1 (1991).

[2] T. Dittrich, P. Hänggi, G.-L. Ingold, B. Kramer, G. Schön, and W. Zwerger, *Quantum Transport and Dissipation* (Wiley-VCH, Weinheim, 1998).

[3] R. Landauer, IBM J. Res. Dev. **1**, 233 (1957); Z. Phys. B: Condens. Matter **68**, 217 (1987); M. Büttiker, Phys. Rev. Lett. **57**, 1761 (1986).

[4] S. Datta, *Electronic Transport in Mesoscopic Systems* (Cambridge University Press, New York, 1995).

[5] B.J. van Wees, H. van Houten, C.W.J. Beenakker, J.G. Williamson, L.P. Kouwenhoven, D. van der Marel, and C.T. Foxon, Phys. Rev. Lett. **60**, 848 (1988).

[6] By nonadiabatic, we mean that even at a QPC's narrowest region the transverse profile is changing rapidly. Clearly every QPC becomes ''nonadiabatic'' at the coupling to infinite-width reservoirs: this type of nonadiabaticity we do not include because it does not cause significant impedance mismatch, as explained by Yacoby and Imry (Ref. 7).

[7] A. Yacoby and Y. Imry, Phys. Rev. B **41**, 5341 (1990).

[8] J.A. Katine, M.A. Eriksson, A.S. Adourian, R.M. Westervelt, J.D. Edwards, A. Lupu-Sax, E.J. Heller, K.L. Campman, and A.C. Gossard, Phys. Rev. Lett. **79**, 4806 (1997).

[9] M.A. Topinka, B.J. LeRoy S.E.J. Shaw, E.J. Heller R.M. Westervelt, K.D. Maranowski, and A.C. Gossard, Science **289**, 2323 (2000).

[10] A.K. Geim, P.C. Main, R. Taboryski, E. Veje, H.A. Carmona, C.V. Brown, T.J. Foster, and L. Eaves, Phys. Rev. B **49**, 2265 (1994).

[11] Although the quasi-1D approach can be retained by modelling very wide leads attached to such systems, following A. Szafer and A.D. Stone, Phys. Rev. Lett. **62**, 300 (1989), this has both numerical and conceptual limitations.

[12] Of course, throughout this paper we could imagine the incident wave on the right-hand side and the same conductance would result (since we are in linear response); see Sec. VI.

[13] We could equally well imagine that the QPC can be ''closed off'' (no transmission) by varying a parameter (this is often true experimentally), and define ψ_0 as the full wave function in this closed-off state. Thus ψ_0 would be the sum of an incident plane wave and a more complicated outgoing wave. This alternative definition may be better in systems where the wall has disorder, or where there is more complicated structure on the left-hand side than shown in Fig. 1(a). The two definitions are equivalent as far as Sec. IV is concerned.

[14] L.D. Landau and E.M. Lifschitz, *Quantum Mechanics* (Elsevier, New York, 1998); J.J. Sakurai, *Modern Quantum Mechanics* (Addison-Wesley, Reading, MA, 1994), Chap. 7. Presentations of two-dimensional scattering theory are rare; see I.R. Lapidus, Am. J. Phys. **50**, 45 (1982); S.K. Adhikari, *ibid.* **54**, 362 (1986); Y.S. Chan, Ph.D. thesis, Harvard University, 1997.

[15] K.L. Shepard, M.L. Roukes, and B.P. Van der Gaag, Phys. Rev. Lett. **68**, 2660 (1992).

[16] This argument can also be verified in the more specific case of the left region being a rectangular Dirichlet box, in which case the exact eigenfunctions are known and can be written explicitly in terms of a sum of ψ_0 for incidences ϕ and $-\phi$. However the phase-space presentation is more general and applies to the real situation where the left region is chaotic (diffusive elastic scattering).

[17] A.H. Barnett, Ph.D. thesis, Harvard University, 2000.

[18] G. Arfken, *Mathematical Methods for Physicists*, 2nd ed. (Academic Press, New York, 1985).

[19] D.S. Fisher and P.A. Lee, Phys. Rev. B **23**, 6851 (1981).

[20] A.D. Stone and A. Szafer, IBM J. Res. Dev. **32**, 317 (1988).

[21] J.S. Hersch, M.R. Haggerty, and E.J. Heller, Phys. Rev. Lett. **83**, 5342 (1999); Phys. Rev. E **62**, 4873 (2000).

[22] J.D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975).

[23] W. Xue and P.A. Lee, Phys. Rev. B **38**, 3913 (1988).

[24] V. Kalmeyer and R.B. Laughlin, Phys. Rev. B **35**, 9805 (1987).

[25] G.W. Bryant, Phys. Rev. B **39**, 3145 (1989).

[26] D.A Wharam, M. Pepper, H. Ahmed, J.E.F. Frost, D.G. Hasko, D.C. Peacock, D.A. Richie, and G.A.C. Jones, J. Phys.: Condens. Matter **21**, L887 (1998); also see Ref. 1 and references within.

[27] M.C. Gutzwiller, *Chaos in Classical and Quantum Mechanics* (Springer-Verlag, New York, 1990), Chap. 7.

[28] With $B \neq 0$, the conductance is *still* symmetric under swapping of the leads. This results from the two-terminal special case of unitarity sum rules (Ref. 4), namely, the rows and columns of the matrix of absolute-value-squared S-matrix elements must all sum to 1. Thus the reciprocity derived here is preserved for $B \neq 0$. How the classical argument from the previous paragraph generalizes for $B \neq 0$ is not known by the authors.

[29] R. Folman, P. Krüger, D. Cassettari, B. Hessmo, T. Maier, and J. Schmiedmayer, Phys. Rev. Lett. **84**, 4749 (2000).

[30] J.H. Thywissen, R.M. Westervelt, and M. Prentiss, Phys. Rev. Lett. **83**, 3762 (1999).

[31] A.H. Barnett, S.P. Smith, M. Olshanii, K.S. Johnson, A.W. Adams, and M. Prentiss, Phys. Rev. A **61**, 023608 (2000).

[32] M. Key, I.G. Huges, W. Rooijakkers, B.E. Sauer, E.A. Hinds, D.J. Richardson, and P.G. Kazansky, Phys. Rev. Lett. **84**, 1371 (2000).

[33] Because we wish to consider general illumination and general $st(k,\Omega)$, our definition of ''conductance'' coincides with that of Thywissen (Ref. 30) only in the case of isotropic illumination $w(k,\Omega)=1/\pi$. The beam *brightness* per unit k range, that is, its phase-space density, is assumed uniform in position space and is proportional to $J_0(k)w(k,\Omega)$. This is also proportional to $a(\mathbf{k})$ defined by Thywissen.

[34] Yu.V. Sharvin, Zh. Éksp. Teor. Fiz. **48**, 984 (1965) [Sov. Phys. JETP **21**, 655 (1965)].

[35] J.M. Krans, J.M. van Ruitenbeek, V.V. Flsun, I.K. Yanson, and L.J. de Jongh, Nature (London) **375**, 767 (1995); P. García-Mochales, P.A. Serena, N. García, and J.L. Costa-Krämer, Phys. Rev. B **53**, 10 268 (1996).

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **34** (2001) 413–437 www.iop.org/Journals/ja PII: S0305-4470(01)15538-1

Rate of energy absorption for a driven chaotic cavity

Alex Barnett[1], **Doron Cohen**[1] **and Eric J Heller**[1,2]

[1] Department of Physics, Harvard University, Cambridge, MA 02138, USA
[2] Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA

Received 13 July 2000, in final form 8 November 2000

Abstract
We consider the response of a chaotic cavity in d dimensions to periodic driving. We are motivated by older studies of one-body dissipation in nuclei, and also by anticipated mesoscopic applications. For calculating the rate of energy absorption due to time-dependent deformation of the confining potential, we introduce an improved version of the wall formula. Our formulation takes into account that a special class of deformations causes no heating in the zero-frequency limit. We also derive a mesoscopic version of the Drude formula, and explain that it can be regarded as a special example of our calculations. Specifically we consider a quantum dot driven by an electro-motive force which is induced by a time-dependent homogeneous magnetic field.

PACS numbers: 0545, 7323, 8530V

1. Introduction

The dynamics of a particle inside a cavity (billiard) in $d = 2$ or 3 dimensions is major theme in studies of classical and quantum chaos. Whereas the physics of time-independent chaotic systems is extensively explored, less is known about the physics of time-dependent chaotic systems. The main exceptions are the studies of the kicked rotator and related systems [1]. However, the rotator (with no kicks) is a one-dimensional integrable system, whereas we are interested in chaotic (two- or three-dimensional) cavities.

Driven cavities were of special interest in the studies of one-body dissipation in nuclei [2–5]. A renewed interest in this problem is anticipated in the field of mesoscopic physics. Quantum dots can be regarded as small two-dimensional cavities whose shape is controlled by electrical gates. Another variation is driving a quantum dot by time-dependent magnetic field. In section 6 we will explain that the calculation of the system response in the latter case can be regarded as a special example of the study in this paper. A similar observation applies to the case of a quantum dot driven by a homogeneous time-dependent electric field. However, in the latter case it is essential to take screening into account [6], and therefore our calculations no longer apply.

We consider a system of non-interacting particles inside a cavity whose walls can be deformed. We define a single parameter x that controls this deformation. We would like to

consider the case where $x(t) = A\sin(\Omega t)$ is being changed periodically in time, where A is the amplitude and Ω is the driving frequency. In particular we are interested in the small-frequency limit, meaning $\Omega \ll 1/\tau_{\rm col}$. Here $\tau_{\rm col}$ is the typical time between collisions with the moving walls of the cavity.

We will be interested in general deformations which need not preserve the billiard shape nor its volume. We can specify any deformation by a function $D(\boldsymbol{s})$, where $\boldsymbol{s}$ specifies the location of a wall element on the boundary (surface) of the cavity, and $D(\boldsymbol{s})\delta x$ is the normal displacement of this wall element. There is a restricted class of deformations that are shape preserving: they involve translations, rotations and dilations of the cavity. We will see that this class has special properties. Note that translations and rotations are also volume preserving, in which case the associated time-dependent deformations can be described as 'shaking' the cavity.

What is the rate at which the 'gas' inside the cavity is heated up? The answer depends on the shape of the cavity and the deformation $D(\boldsymbol{s})$ involved, as well as on the amplitude A and the driving frequency Ω. Also the number of particles N and their energy distribution $\rho(E)$ should be specified.

For non-interacting particles the solution of this problem is reduced to the analysis of one-particle physics. This observation is self-evident for non-interacting classical particles, but it is also true for non-interacting fermions (see appendix A). We would like to work within the framework of linear response theory (LRT). In such a case one can write

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\mathcal{H}\rangle = \mu(\Omega) \times \frac{1}{2}(A\Omega)^2 \tag{1}$$

where the dissipation coefficient $\mu(\Omega)$ is amplitude independent. The small-Ω version of this formula can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\mathcal{H}\rangle = \mu V^2 \tag{2}$$

where $\mu = \mu(\Omega \to 0)$ is known as the friction coefficient, and $V = A\Omega/\sqrt{2}$ is the average root-mean-square (RMS) deformation velocity. For convenience let us define x as having units of length, such that V characterizes the velocity of the moving walls.

A necessary classical condition for the validity of LRT is $V \ll v_{\rm E}$ where $v_{\rm E} \equiv (2E/m)^{1/2}$ is the velocity of the particle [7, 8]. We also assume that the motion of the particle inside the cavity is globally chaotic, meaning no mixed phase space [9]. The criteria for having such a cavity are discussed in [10, 11]. The justification of LRT in the quantum mechanical case is more subtle [8, 12, 13], and does not constitute a theme in this paper, although we do connect with the quantum case in section 3. The theory to be presented assumes that LRT is a valid formulation of the problem.

As explained in appendix A, LRT implies that the dissipation coefficient $\mu(\Omega)$ is related via a fluctuation–dissipation (FD) relation to a spectral function $\tilde{C}_{\rm E}(\Omega)$. Namely,

$$\mu(\Omega) = \frac{1}{2}\int_0^\infty \rho(E)\,\mathrm{d}E\,\frac{1}{g(E)}\frac{\partial}{\partial E}[g(E)\tilde{C}_{\rm E}(\Omega)]. \tag{3}$$

Here $\rho(E)$ is the energy distribution of the particles, and $g(E)$ is the density of states. The function $\tilde{C}_{\rm E}(\omega)$ is the noise power spectrum of the generalized 'force' associated with the parameter x. This function is the main object of the present study, and its precise definition is in section 2. We shall chiefly explore how $\tilde{C}_{\rm E}(\omega)$ depends on the type of deformation involved, but also discuss effects due to the cavity shape.

In particular we are interested in the small-frequency limit where μ is related to the

fluctuation intensity

$$\nu_{\mathrm{E}} = \tilde{C}_{\mathrm{E}}(0) = \int_{-\infty}^{\infty} C_{\mathrm{E}}(\tau)\,\mathrm{d}\tau. \tag{4}$$

The simplest estimate for ν_{E}, which we are going to call the 'white-noise approximation' (WNA), leads (in the case of a 3D cavity) to the well known 'wall formula' [2]

$$\mu_{\mathrm{E}} = \frac{N}{\mathsf{V}} m v_{\mathrm{E}} \oint D(\boldsymbol{s})^2\,\mathrm{d}\boldsymbol{s} \tag{5}$$

where the subscript E implies that we are considering a microcanonical ensemble $\rho(E)$, the number of particles is N and the volume of the cavity is V. The above version of the wall formula has been derived for the purpose of calculating the so-called one-body dissipation rate in nuclei. The original derivation of this formula is based on a simplified kinetic picture [2]. For alternate derivations using the LRT approach see [3]. For the generalization to any dimension d using the LRT–FD strategy see [8] and further references therein.

Our main purpose is to introduce an improved version of the wall formula, and to analyse the frequency dependence of $\mu(\Omega)$. This will involve a demonstration [14] that for special types of deformation (namely dilations, translations and rotations) the small-Ω dissipation rate is remarkably different from the naive expectation. As an application, the mesoscopic version of the Drude formula for the *conductance* of a quantum dot in a uniform time-dependent magnetic field reduces to the the calculation of $\tilde{C}_{\mathrm{E}}(\omega)$ for one of these special deformations (namely rotation), and leads to (see section 6)

$$\mu(\Omega) \sim \frac{N}{\mathsf{A}} \left(\frac{e^2}{m}\tau_{\mathrm{col}}\right) \frac{1}{1 + (\tau_{\mathrm{col}}\Omega)^2} \tag{6}$$

where A is the area of the dot.

For our improved wall formula, we show that it is essential to project out the special components of a general deformation, and only then to estimate ν_{E} using the WNA. If the assumption of strong chaos cannot be justified, further corrections are required due to correlations between successive bounces.

The effect of interaction between the particles is not discussed in this paper. If the mean free path for inter-particle collisions is large compared with the size of the cavity, then we expect that our analysis still applies. If the mean free path is much smaller, then we get into the hydrodynamic regime. In the latter case we have a drag effect, and the dissipation rate is determined by the viscosity of the gas via Stokes' law.

2. The model system

Consider a particle whose canonical coordinates are $(\boldsymbol{r}, \boldsymbol{p})$ moving inside a cavity. The Hamiltonian is

$$\mathcal{H}(\boldsymbol{r}, \boldsymbol{p}; x) = \boldsymbol{p}^2/2m + U(\boldsymbol{r} - x\boldsymbol{D}(\boldsymbol{r})) \tag{7}$$

where $U(\boldsymbol{r})$ is the confining potential. We have introduced a (unitless) deformation 'field' $\boldsymbol{D}(\boldsymbol{r})$, and x is the controlling parameter. In this paper we assume that $U(\boldsymbol{r}) = 0$ inside the cavity. The volume of the cavity is V. Outside the cavity the potential $U(\boldsymbol{r})$ becomes very large. To be specific, one may assume that the walls exert a normal force f, and we take the hard-wall limit $f \to \infty$. With the above assumptions about $U(\boldsymbol{r})$ it is clear that the deformation is completely specified by the boundary function $D(\boldsymbol{s}) \equiv \hat{\boldsymbol{n}}(\boldsymbol{s}) \cdot \boldsymbol{D}(\boldsymbol{s})$, where $\hat{\boldsymbol{n}}(\boldsymbol{s})$ is an outwards unit normal vector at the boundary point $\boldsymbol{s}$.

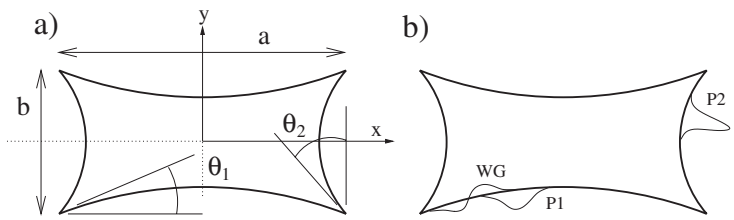


Figure 1. (*a*) The generalized two-dimensional Sinai billiard which has been used for our numerical studies. (*b*) Three example deformations are illustrated. Note that they are shown exaggerated in strength.

Table 1. Key to deformation types used for numerical two-dimensional billiard experiments in this paper. L is the billiard perimeter. The deformation is described by a function $D(s)$, where s is measured anticlockwise along the perimeter with $s = 0$ at the upper left corner. In the 'fracture' and 'shift-x' cases we use the horizontal Cartesian coordinate $x(s)$.

Key	Description	Surface deformation function $D(s)$
CO	Constant	1
Wn	n periods	$\cos(2\pi ns/L)$
DF	Diffuse	random$[-1, 1]$ (equivalent to W∞)
FR	Fracture	$\mathrm{sgn}(x(s))$ if on top or bottom, else 0
SX	Shift-x	$\mathrm{sgn}(x(s))$ if on left or right, else 0
P1	Piston 1	$10\exp(-\frac{1}{2}\alpha^2)$, $\alpha = (s/L - 0.3)/0.01$
P2	Piston 2	$10\exp(-\frac{1}{2}\alpha^2)$, $\alpha = (s/L - 0.6)/0.005$
WG	Wiggle	$5\alpha\exp(-\frac{1}{2}\alpha^2)$, $\alpha = (s/L - 0.25)/0.02$

Table 2. Key to the four 'special' deformations in two dimensions. The unit vectors $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$ are in the plane (see figure 1), and $\boldsymbol{e}_z$ is in the perpendicular direction. In the case of dilation and rotation $\boldsymbol{D}$ could be made unitless by dividing by a constant length.

Key	Description	Deformation field
DI	Dilation about origin	$\boldsymbol{D}(\boldsymbol{r}) = \boldsymbol{r}$
TX	x-translation	$\boldsymbol{D}(\boldsymbol{r}) = \boldsymbol{e}_x$
TY	y-translation	$\boldsymbol{D}(\boldsymbol{r}) = \boldsymbol{e}_y$
RO	Rotation about origin	$\boldsymbol{D}(\boldsymbol{r}) = \boldsymbol{e}_z \times \boldsymbol{r}$

Most of our numerical tests will refer to the two-dimensional cavity illustrated in figure 1(*a*). It is a generalized two-dimensional Sinai billiard formed from concave arcs of circles with two different radii. Typical parameters used are $a = 2$, $b = 1$, $\theta_1 = 0.2$ and $\theta_2 = 0.5$, for which the average collision rate with the wall is $(1/\tau_{\mathrm{bl}}) \approx 0.63$. This billiard has been chosen because it has 'hard chaos': there is no mixed phase space, and there are no marginally stable orbits (see section 7). In figure 1(*b*) we show three example deformations. For illustration purposes we have selected three 'localized' deformations. See tables 1 and 2 for a full list of deformations that have been tested in our numerical work.

Associated with the parameter x is the fluctuating quantity

$$\mathcal{F}(t) = -\left.\frac{\partial \mathcal{H}(\boldsymbol{r}, \boldsymbol{p}; x)}{\partial x}\right|_{x=0} \tag{8}$$

where the time dependence arises from that of the trajectory $(\boldsymbol{r}(t), \boldsymbol{p}(t))$. This quantity can be thought of as the generalized time-dependent 'force' associated with the parameter x. For the

Hamiltonian (7) we can write

$$\mathcal{F}(t) = \boldsymbol{D}(\boldsymbol{r}) \cdot \boldsymbol{\nabla} U(\boldsymbol{r}) = -\boldsymbol{D}(\boldsymbol{r}) \cdot \dot{\boldsymbol{p}}. \tag{9}$$

Recognizing $\dot{\boldsymbol{p}}$ as the force on the gas particle, we see that $\mathcal{F}(t)$ is a train of spikes (see figure 2(*a*)). Namely, the fluctuating force $\mathcal{F}(t)$ consists of impulses whose maximum duration is $\tau_0 = 2mv_{\mathrm{E}}/f$. In the hard-wall limit $\tau_0 \to 0$, and we can write

$$\mathcal{F}(t) = \sum_i 2mv_{\mathrm{E}} \cos(\theta_i) D_i \delta(t - t_i) \tag{10}$$

where i labels collisions: t_i is the time of a collision, D_i stands for $D(\boldsymbol{s}_i)$ at the location $\boldsymbol{s}_i$ of a collision and $v_{\mathrm{E}} \cos(\theta_i)$ is the normal component of the particle's collision velocity. The above sequence of impulses is characterized by an average rate of collisions $1/\tau_{\mathrm{col}}$. The quantitative definition of τ_{col} is postponed to section 4. Note however that τ_{col} may be much larger that the ballistic time τ_{bl}. The ballistic time is the average time between collisions with the boundary. We have $\tau_{\mathrm{col}} \gg \tau_{\mathrm{bl}}$ whenever a deformation involves only a small piece of the boundary. Finally we note that if the deformation is volume preserving then $\langle \mathcal{F}(t) \rangle = 0$. Otherwise it is convenient to subtract the (constant) average value $F(x)$ from the above definition of $\mathcal{F}(t)$. This convention is reflected in our illustration (figure 2(*a*)).

We define the auto-correlation function of $\mathcal{F}(t)$ as follows:

$$C_{\mathrm{E}}(\tau) \equiv \langle \mathcal{F}(t)\mathcal{F}(t+\tau) \rangle_{\mathrm{E}}. \tag{11}$$

The subscript E, whenever used, suggests that the average over initial conditions is of microcanonical type, with energy E. Note that $C_{\mathrm{E}}(\tau)$ is defined using the time-independent ('frozen') Hamiltonian, and therefore is independent of t. The auto-correlation function $C_{\mathrm{E}}(\tau)$ can be handled as a time average rather than an ensemble average (by ergodicity). The resulting construction is illustrated in figure 2(*b*), where we illustrate the projection of $\mathcal{F}(t_1)\mathcal{F}(t_2)$ onto the $\tau \equiv t_2 - t_1$ axis. The contribution for the self-correlation is shaded. The forms of the resultant $C_{\mathrm{E}}(\tau)$ and its Fourier transform $\tilde{C}_{\mathrm{E}}(\omega) \equiv \int C_{\mathrm{E}}(\tau) \exp(\mathrm{i}\omega\tau)\, \mathrm{d}\tau$ are illustrated schematically in figures 2(*c*) and (*d*). Note that the $\omega \to 0$ limit of $\tilde{C}(\omega)$ is equal to the area under $C(\tau)$.

The auto-correlation function $C_{\mathrm{E}}(\tau)$ consists of a $\tau = 0$ ('self') peak due to the self-correlation of the spikes, and of an additional smooth ('non-self') component due to correlations between successive bounces. This implies[3] that pronounced correlations are usually characterized by the time scale τ_{bl}, rather than τ_{col}. Consequently the associated frequency scale for non-universal structures is $\omega \sim 1/\tau_{\mathrm{bl}}$. Another relevant timescale is the ergodic time t_{erg}, which is the inverse of the average Lyapunov (instability) exponent. Beyond t_{erg} the correlations become vanishingly small. Non-negligible tails may arise only if the motion has marginally stable orbits.

As explained in the introduction, we shall be most interested in the noise intensity ν_{E} defined by (4). Observing that $\mathcal{F}(t)$ is linear in $D(\boldsymbol{s})$, it follows that the noise intensity is a quadratic functional

$$\nu_{\mathrm{E}} = \oint\!\!\oint \mathrm{d}\boldsymbol{s}_1\, \mathrm{d}\boldsymbol{s}_2\, D(\boldsymbol{s}_1)\gamma_{\mathrm{E}}(\boldsymbol{s}_1, \boldsymbol{s}_2) D(\boldsymbol{s}_2) \tag{12}$$

where the kernel γ_{E} depends on both the cavity shape and the particle energy E [3]. Furthermore, billiards are *scaling systems* in the sense that a change in E leaves the trajectories unchanged. From this and (10) we have the scaling relation $\gamma_{\mathrm{E}}(\boldsymbol{s}_1, \boldsymbol{s}_2) = m^2 v_{\mathrm{E}}^3 \cdot \hat{\gamma}(\boldsymbol{s}_1, \boldsymbol{s}_2)$,

[3] Consider the case of a deformation which involves only a small piece of the boundary. Typically, the time between collisions with the deforming piece is τ_{col}. However, correlations are dominated by the rare events when the time between collisions is $\sim\tau_{\mathrm{bl}}$.

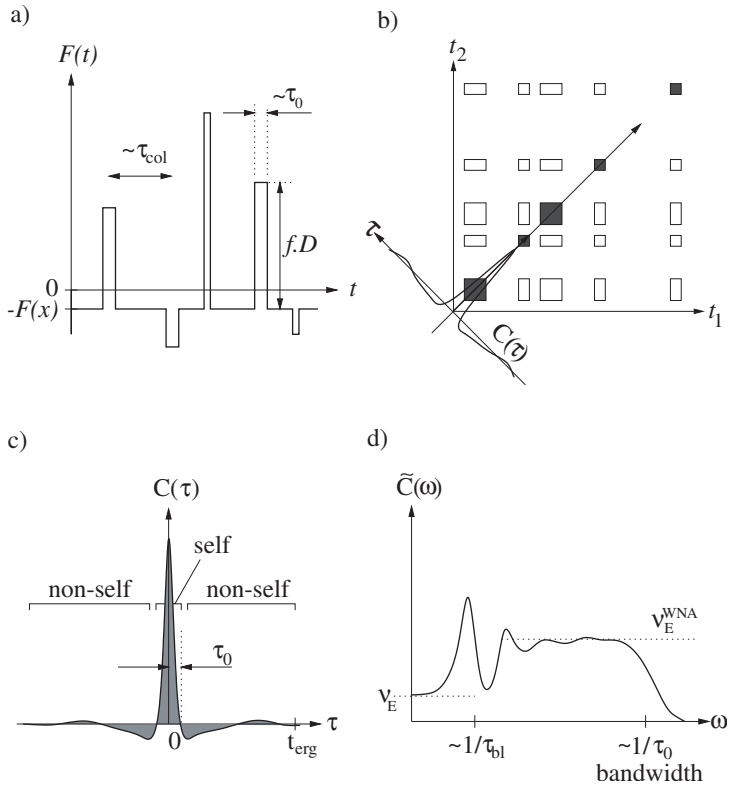


Figure 2. The fluctuating force $\mathcal{F}(t)$ looks like a train of impulses (*a*). Due to ergodicity the autocorrelation function $C(\tau)$ can be regarded as a time average (*b*). The resultant autocorrelation function (*c*) and the associated power spectrum (*d*) may be characterized by non-universal features. See the text for further explanations.

where the scaled kernel depends entirely on the geometrical shape of the cavity. However, the reason for being interested in *approximations* for ν_{E} is that the exact result for the kernel $\hat{\gamma}$ is very complicated to evaluate, and involves a sum over all classical paths from $\boldsymbol{s}_1$ to $\boldsymbol{s}_2$ (see [3]).

3. Quantum–classical correspondence

This paper applies classical physics in order to analyse the response of a wide class of systems, including mesoscopic systems where quantum mechanics may play a role. How much of a compromise is a classical analysis of the dissipation? This question has been addressed in [8, 12]. At the level of one-particle physics the answer is as follows: within the framework of LRT the only difference between the classical formulation and the quantal one is involved in replacing the classical definition of $C_{\mathrm{E}}(\tau)$ by the corresponding quantum mechanical definition. In the level of many (non-interacting) particles the only further modification is associated with the application of the FD relation, as discussed in appendix A (see equation (A.6)). We would like to re-emphasize that we assume in this paper that we are in an (A, Ω) regime where

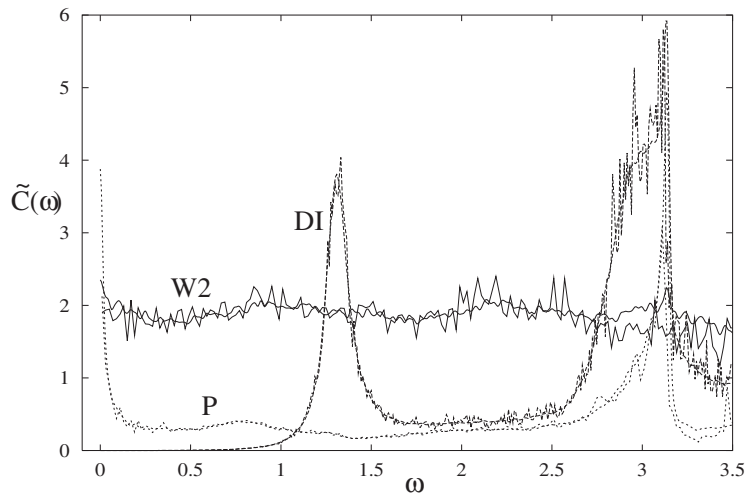


Figure 3. Agreement between quantum and classical $\tilde{C}_{\mathrm{E}}(\omega)$ in the two-dimensional quarter-stadium billiard for three example deformations (see text). In each case classical is shown as a thick line, and quantum a thin line. The y-axis has been displaced to clearly show the $\omega \to 0$ behaviour. The singular peak at $\omega = \pi$ is due to the 'bouncing ball' orbit.

LRT is a valid formulation. The quantum adiabatic regime (extremely small Ω), and the non-perturbative regime (see discussion in [13]) are excluded from our considerations.

Thus the only remaining question is whether a classical calculation of $\tilde{C}_{\mathrm{E}}(\omega)$ is a good approximation quantum mechanically. The answer is that the quantum–classical correspondence here is remarkable. It has been tested for a few example systems [14–16]. In figure 3 we demonstrate correspondence for the stadium billiard for three types of deformation: DI (dilation), W2 (periodic oscillation around the perimeter) and P (wide 'piston' existing only on the top edge). The RMS estimation error is 3% for the classical calculation and 10% for the quantum calculation. The quantum estimate of $\tilde{C}_{\mathrm{E}}(\omega)$ amounts to computing boundary overlap integrals of the eigenfunctions (see [14]). We have used all 451 states lying in the range of wavenumbers $398 < k < 402$, where the mean level spacing is $\Delta \approx 8.8 \times 10^{-3}$ in ω units. Note that there are $\sim 10^2$ de Broglie wavelengths across the system. The stadium was chosen because it enables efficient quantization using the method of Vergini and Saraceno [17, 18]. An especially good basis set is known for this shape [19].

4. The white-noise approximation

The most naive estimate of the fluctuation intensity is based on the WNA. Namely, one assumes that the correlation between bounces can be neglected. This corresponds [3] to the *local* part of the kernel (12). In such case only the self-correlation of the spikes is taken into consideration and one obtains [8]

$$\nu_{\mathrm{E}} \approx (2mv_{\mathrm{E}})^2 \left\langle \sum_i \cos^2(\theta_i) D_i^2 \delta(t - t_i) \right\rangle_{\mathrm{E}} \tag{13}$$

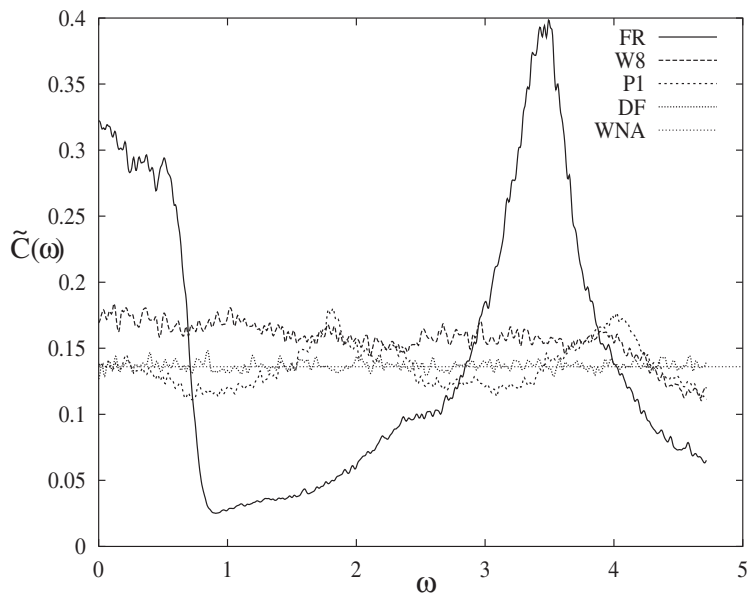


Figure 4. The white-noise approximation estimate (WNA is the horizontal dotted line) compared to actual $\tilde{C}_{\mathrm{E}}(\omega)$ power spectra for some example deformations of the two-dimensional generalized Sinai billiard, with $m = v = 1$. (The RMS estimation error of 3% can be seen as multiplicative noise with short correlation length in ω.) Deformation functions are defined in table 1, and discussed further in the text.

and from here (see [8]) using ergodicity,

$$\nu_{\mathrm{E}} \approx 2m^2 v_{\mathrm{E}}^3 \langle |\cos\theta|^3 \rangle \frac{1}{\mathrm{V}} \oint [D(\boldsymbol{s})]^2 \, \mathrm{d}\boldsymbol{s} \tag{14}$$

where the geometric factor for $d = 2, 3, \ldots$ is $\langle |\cos(\theta)|^3 \rangle = 4/(3\pi), 1/4, \ldots$. If we can use the convention $|D(\boldsymbol{s})| \sim 1$ over the deformed region (and zero otherwise), then we can write the WNA as $\nu_{\mathrm{E}} = (2mv_{\mathrm{E}})^2 \times (1/\tau_{\mathrm{col}})$ where $(1/\tau_{\mathrm{col}})$ defines the effective collision rate. For a more careful discussion see appendix F of [8]. Note again that τ_{col} can be much larger than the ballistic time τ_{bl} in the case where only a small piece of the boundary is being deformed.

The use of the WNA can be justified whenever successive collisions are effectively uncorrelated. The applicability of such an assumption depends on the shape of the cavity (which will determine the decay of correlations via the typical Lyapunov exponent) as well as on the type of deformation involved. If we have the cavity of figure 1(a), and the deformation involves only a small piece of the boundary (e.g. see figure 1(b)), then successive collisions with the *deformed part* of the boundary are effectively uncorrelated. This is so because there are many collisions with static pieces of the boundary before the next effective collision (with non-zero D_i) takes place. If the deformation involves a large piece (or all) of the boundary, we can still argue that successive collisions are effectively uncorrelated provided $D(\boldsymbol{s})$ is 'oscillatory' enough (i.e. changes sign many times along the boundary). These expectations are qualitatively confirmed by the numerical results of figure 4. Here we show a sequence of deformation types for which the WNA performs increasingly well: FR (for which sensitivity to the vertical least unstable periodic orbit causes large correlation effects and large deviations from WNA), W8 (oscillatory deformation changes sign many times around the perimeter,

giving better agreement with WNA), P1 (localized 'piston' type deformation, for which WNA is good) and DF (random function of zero correlation-length along the perimeter, showing complete WNA agreement).

The numerical evaluation of $\tilde{C}_{\mathrm{E}}(\omega)$ throughout this paper is performed by squaring the Fourier transform of a single long sample of $\mathcal{F}(t)$ ($\sim 10^6$ consecutive collisions). Ergodicity ensures that the properties of a single trajectory reproduce the desired ensemble average $\langle \cdots \rangle_{\mathrm{E}}$. In practice the power spectrum of a single sample is a stochastic quantity with no correlations in ω-space. To estimate the underlying noise spectrum $\tilde{C}_{\mathrm{E}}(\omega)$ a smoothing convolution in ω-space is performed. In the figures a smoothing width of 10^{-2} is typical, giving 3% RMS estimation error. The δ-function nature of $\mathcal{F}(t)$ is handled by convolving in the time-domain with a suitably narrow Gaussian. This enables the signal to be sampled uniformly in time, and hence we can benefit from use of the fast Fourier transform procedure.

It might be asked whether the exponential growth in sensitivity to numerical round-off error invalidates the computation of the properties of a long classical trajectory. The answer is no: it has been shown that in simple two-dimensional chaotic maps such as ours, a numerically generated 'pseudo-trajectory' *shadows* (is very close to) a true trajectory with slightly different initial conditions [20]. However, as we shall see, the differences in $\omega \to 0$ behaviour (in the hard-chaos case) do not in fact rely on correlation properties over times any longer than t_{erg}.

5. 'Special' deformations

The WNA dramatically fails (see figure 5) for dilation, translations and rotations (see table 2 for their definitions in two dimensions). It is not surprising that the WNA is 'bad' for these deformations because their $\boldsymbol{D}(\boldsymbol{s})$ are slowly changing delocalized functions of $\boldsymbol{s}$. However, what is remarkable is that $\tilde{C}_{\mathrm{E}}(\omega)$ for this type of deformation *vanishes* in the limit $\omega \to 0$. Such deformations we would like to call 'special' [14]. More generally, we would like to say that a deformation is 'special' if the associated fluctuation intensity is $\nu_{\mathrm{E}} = 0$.

A result that follows from the considerations of appendix B is that a linear combination of special deformation is also special. Therefore the special deformations constitute a linear space of functions. We believe that this linear space is spanned by the following basis functions: one dilation, d translations and $d(d-1)/2$ rotations. However we are not able to give a rigorous mathematical argument that excludes the possibility of having a larger linear space. In other words, we believe that any special deformation can be written as a linear combination of dilation, translations and rotations.

We will explain the observed $\nu_{\mathrm{E}} = 0$, starting with the case of translations and dilations. For translations we have $\boldsymbol{D} = \boldsymbol{e}$, where $\boldsymbol{e}$ is a constant vector that defines a direction in space. We can write $\mathcal{F}(t) = (\mathrm{d}/\mathrm{d}t)^2\mathcal{G}(t)$ where $\mathcal{G}(t) = -m\boldsymbol{e}\cdot\boldsymbol{r}$. A similar relation holds for dilation $\boldsymbol{D} = \boldsymbol{r}$ with $\mathcal{G}(t) = -\frac{1}{2}m\boldsymbol{r}^2$. It follows that $\tilde{C}(\omega) = \omega^4\tilde{C}_G(\omega)$, where $\tilde{C}_G(\omega)$ is the power spectrum of $\mathcal{G}(t)$. If $\tilde{C}_G(\omega)$ is a bounded function (as it must be when correlations are short range), it immediately follows that $\tilde{C}(0) = 0$. Moreover since $\mathcal{G}(t)$ is a simple function of the particle position, we can assume it is a fluctuating quantity that looks like white noise on timescales $> t_{\mathrm{erg}}$. It follows that $\tilde{C}(\omega)$ is generically characterized by ω^4 behaviour for either translations or dilations.

We now turn to consider the case of rotations. This case is of particular interest because of its relation to the Drude conductance calculation in a uniform driving magnetic field (see the following section). For rotations we have $\boldsymbol{D} = \boldsymbol{e} \times \boldsymbol{r}$, and we can write $\mathcal{F}(t) = (\mathrm{d}/\mathrm{d}t)\mathcal{G}(t)$, where $\mathcal{G}(t) = -\boldsymbol{e}\cdot(\boldsymbol{r}\times\boldsymbol{p})$, is a projection of the particle's angular momentum vector[4].

[4] The cross-product form used here for $\boldsymbol{D}$ and $\mathcal{G}(t)$ is strictly valid in two and three dimensions only. For $d > 3$ the higher-dimensional generalization of a general rotation should be used.

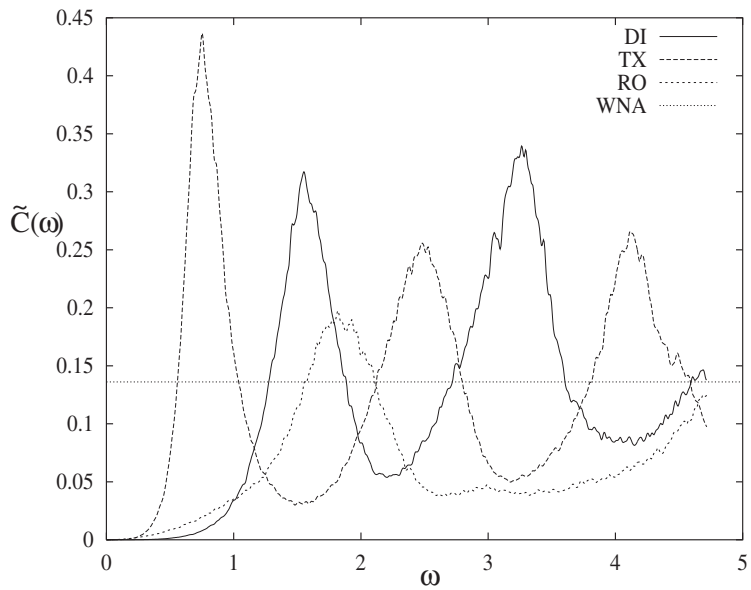


Figure 5. The WNA estimate compared to actual $\tilde{C}(\omega)$ noise power spectra for example 'special' deformation types: DI (dilation), TX (translation) and RO (rotation). See table 2 for definitions. The WNA fails to predict the vanishing in the small-ω limit.

Consequently $\tilde{C}(\omega) = \omega^2 \tilde{C}_G(\omega)$. Assuming the angular momentum is a fluctuating quantity that looks like white noise on timescales $>t_{\mathrm{erg}}$, it follows that $\tilde{C}(\omega)$ is generically characterized by ω^2 behaviour.

Thus we have predictions for the power laws in the regime $\omega < 1/t_{\mathrm{erg}}$ for special deformations (assuming hard chaos). These have been verified numerically in our previous paper [14], with a special emphasis on the case of dilation. The case of dilation plays a vital role in a highly successful numerical billiard diagonalization method that has been introduced recently [17].

For special deformations we have $\tilde{C}(\omega) = 0$ in the limit $\omega = 0$, and consequently the dissipation coefficient vanishes ($\mu = 0$). It should be noted that for the case of a general combination of translations and rotations this result follows from a simpler argument. Taking $\Omega \to 0$ while keeping $A\Omega$ constant corresponds to constant deformation velocity ($\dot{x} = \mathrm{const}$). Transforming the time-dependent Hamiltonian into the reference frame of the cavity (which is uniformly translating or rotating with constant velocity) gives a *time-independent* Hamiltonian. In the new reference frame the energy is a constant of the motion, which implies that the system cannot absorb energy (no dissipation effect), and hence we must indeed have $\mu = 0$.

6. Drude mesoscopic conductance for two-dimensional dot

Consider a two-dimensional quantum dot in a homogeneous (perpendicular) magnetic field (see figure 6(*b*)). The one-particle Hamiltonian is

$$\mathcal{H}(\boldsymbol{r}, \boldsymbol{p}; \Phi(t)) = \frac{1}{2m}[\boldsymbol{p} - e\boldsymbol{A}(\boldsymbol{r}; \Phi(t))]^2 + U(\boldsymbol{r}). \tag{15}$$

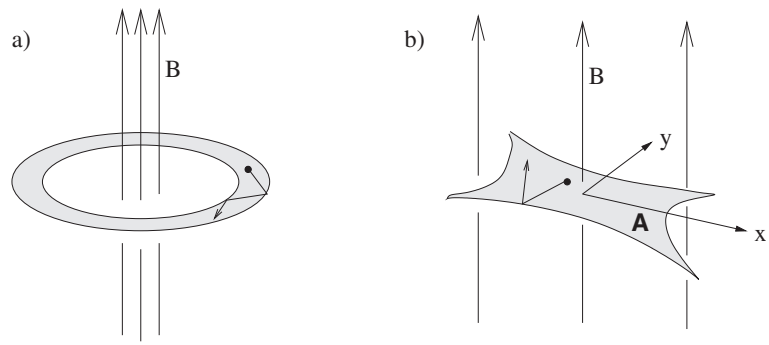


Figure 6. Two possible mesoscopic geometries which exhibit conductance when driven by a magnetic field: (*a*) conventional ring of perimeter L enclosing the time-dependent flux, (*b*) ballistic two-dimensional chaotic dot (cavity) of area A in a uniform time-dependent magnetic field.

The dot is defined by the confining potential $U(\boldsymbol{r})$, and we choose the magnetic field as the controlling (driving) parameter. Periodic driving means $\Phi(t) = A\sin(\Omega t)$. The vector potential is given by

$$\boldsymbol{A}(\boldsymbol{r};\Phi) = \frac{1}{2}\left(\frac{\Phi}{\mathsf{A}}\hat{\boldsymbol{z}}\right)\times\boldsymbol{r} \tag{16}$$

where A is the area of the dot, Φ/A is the magnetic field and $\hat{\boldsymbol{z}}$ is its (perpendicular) direction.

Referring to equation (2) one should realize that by Faraday's law $V = \dot{\Phi}$ is the induced electromotive force (measured in volts). Hence μ is just the conductance. The fluctuating quantity that is associated with Φ has the meaning of electric current:

$$\mathcal{I}(t) = -\frac{\partial\mathcal{H}}{\partial\Phi} = \frac{e}{2\mathsf{A}}(\hat{\boldsymbol{z}}\times\boldsymbol{r})\cdot\boldsymbol{v}. \tag{17}$$

In the conventional ring geometry (figure 6(*a*)) the current is just $\mathcal{I}(t) = (e/\mathsf{L})v$, where L is the perimeter, and v is the tangential velocity. In the general cavity case (figure 6(*b*)) $\mathcal{I}(t)$ can be thought of as the angular momentum of the charge.

The Drude mesoscopic conductance is given by the frequency-dependent version of the FD relation equation (A.6). With the one-particle density of states corresponding to a two-dimensional gas equation (A.6) becomes

$$\mu(\Omega) = \frac{N}{mv_{\mathrm{F}}^2}\tilde{C}_{\mathcal{I}}(\Omega) \tag{18}$$

where the Fermi velocity is related to the Fermi energy $E_{\mathrm{F}} = \frac{1}{2}mv_{\mathrm{F}}^2$. The power spectrum of the electric current $\tilde{C}_{\mathcal{I}}(\omega)$ is the Fourier transform of the current–current correlation function. In standard derivations of the Drude formula it is assumed that this correlation function is exponential:

$$C_{\mathcal{I}}(\tau) \sim \frac{e^2}{\mathsf{A}}v_{\mathrm{F}}^2\exp\left(-\frac{|\tau|}{\tau_{\mathrm{col}}}\right) \tag{19}$$

leading to the Lorentzian equation (6). However, for a given dot shape $C_{\mathcal{I}}(\tau)$ is not really an exponential, but rather reflects the system-specific geometry. Below we discuss two limits in which we can obtain approximations for $C_{\mathcal{I}}(\tau)$ and hence (via equation (18)) for the frequency-dependent conductance $\mu(\Omega)$.

The current $\mathcal{I}(t)$ is a piecewise constant function of time. It is constant between collisions with the walls because of conservation of angular momentum. The derivative of this quantity, $\mathcal{F}(t) = \dot{\mathcal{I}}$, is a train of spikes. It formally coincides (using (9)) with the $\mathcal{F}(t)$ of the deformation $\boldsymbol{D}(\boldsymbol{r}) = (e/(2m\mathsf{A}))\hat{\boldsymbol{z}} \times \boldsymbol{r}$, corresponding to rotation around the z axis. It follows that the current–current correlation $C_{\mathcal{I}}(\tau)$ is trivially related to the $\mathcal{F}(t)$ correlation function $C(\tau)$ as follows:

$$\tilde{C}_{\mathcal{I}}(\omega) = \frac{1}{\omega^2}\tilde{C}(\omega). \tag{20}$$

Thus we see that the calculation of 'conductance' is formally equivalent to a special case of deformation, namely a rotation.

There are two limits in which we can obtain an approximation for $\tilde{C}_{\mathcal{I}}(\omega)$. For small frequencies $\omega \ll (1/\tau_{\mathrm{col}})$ we may use the following simple estimate:

$$\tilde{C}_{\mathcal{I}}(\omega) = \langle \mathcal{I}^2 \rangle \times 2\tau_{\mathrm{col}} = \frac{1}{4}e^2\,\frac{\langle \boldsymbol{r}^2 \rangle}{\mathsf{A}^2}v_{\mathrm{F}}^2 \times \tau_{\mathrm{col}}. \tag{21}$$

The first equality can be taken as an operative definition of the correlation time τ_{col} in the context of this calculation. Obviously, up to a system-specific geometrical factor this result ($\sim\omega^0$) agrees with the standard Drude result. For ring geometry one should make the replacements $\langle \boldsymbol{r}^2 \rangle \mapsto (\mathsf{L}/(2\pi))^2$ and $\mathsf{A} \mapsto \pi(\mathsf{L}/(2\pi))^2$ where L is the length of the wire (perimeter of the ring). Thus one obtains $\tilde{C}_{\mathcal{I}}(\omega) = ((e/\mathsf{L})v_{\mathrm{F}})^2$ leading to the standard-looking Drude formula for a mesoscopic wire $\mu = (N/\mathsf{L}^2) \times (e^2/m) \times \tau_{\mathrm{col}}$.

In the limit $\omega \gg (1/\tau_{\mathrm{col}})$ we can get a much more satisfying result. The fluctuating quantity $\mathcal{F}(t) = \dot{\mathcal{I}}(t)$ is the same as (9) with $\boldsymbol{D}(\boldsymbol{r}) = (e/(2m\mathsf{A}))\hat{\boldsymbol{z}} \times \boldsymbol{r}$, corresponding to rotation. Using the WNA of equation (14), and dividing by ω^2 as in (20) we get

$$\tilde{C}_{\mathcal{I}}(\omega) = \left[\frac{2}{3\pi}\frac{e^2}{\mathsf{A}^3}v_{\mathrm{F}}^3 \oint |\boldsymbol{n} \times \boldsymbol{r}|^2\,\mathrm{d}s\right]\frac{1}{\omega^2}. \tag{22}$$

Again, up to a system-specific geometrical factor this result ($\sim\omega^{-2}$) agrees with the standard Drude result. The latter expression should become exact as we go to large frequencies, where the only significant contribution comes from the self-correlation of the $\mathcal{F}(t)$ spikes (see figure 2(*d*)).

Equation (21) leads (via (18)) to the small-frequency Drude result, while the WNA of equation (22) gives the Lorentzian tail of the Drude result. An exact result for the frequency-dependent conductance can be calculated numerically for a given geometrical shape. In figure 7 we display a plot of $\mu(\Omega) \propto \tilde{C}_{\mathcal{I}}(\Omega)$, which shows both the constant behaviour at small Ω and the convergence to the large-Ω WNA approximation. System-specific features are expressed by the deviation from a standard Lorentzian in the intermediate-frequency regime.

Finally we consider driving a quantum dot with homogeneous electric field in the x direction, in which case the Hamiltonian contains the interaction term $-e\mathcal{E}(t)x$. For calculation of the response in such a case one should evaluate the dipole–dipole correlation function $C_{\mathcal{P}}(\tau)$ where $\mathcal{P}(t) = ex$. The latter is related to translations, where the deformation field is $\boldsymbol{D} = \hat{\boldsymbol{x}}$. Consequently we get $C_{\mathcal{P}}(\tau) = (1/\omega^4)C(\tau)$. However, this result is not of great interest, because the screening effect leads to modification of the effective one-particle Hamiltonian, such that the actual electric field inside a quantum dot is much smaller than the applied field.

7. The white-noise assumption revisited

In section 4 we have assumed that generic fluctuating quantities such as $\boldsymbol{r}^2$ and $\boldsymbol{e} \cdot \boldsymbol{r}$ and $\boldsymbol{e} \cdot (\boldsymbol{r} \times \boldsymbol{p})$ have a white-noise power spectrum for $\omega \ll 1/\tau_{\mathrm{bl}}$. In section 8 we are going

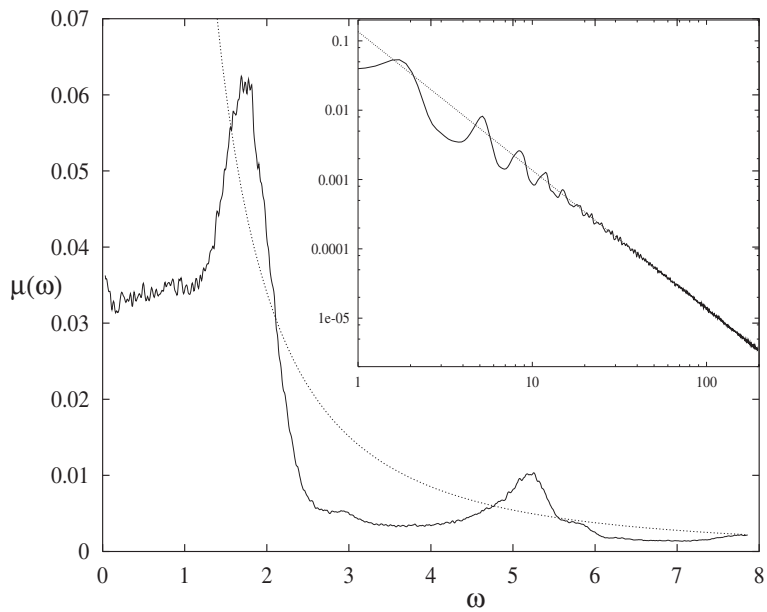


Figure 7. Calculation of dissipation coefficient $\mu(\Omega)$ (arbitrary units) for driving of a chaotic mesoscopic billiard system with a constant magnetic field at frequency Ω. The billiard chosen is the two-dimensional generalized Sinai of figure 1(*a*). The dotted curve WNA ($\mathcal{F}$) is the high-frequency estimate assuming $\mathcal{F}(t) \equiv \dot{\mathcal{I}}$ is white noise. The convergence to this ω^{-2} result is clearly visible in the log–log inset plot.

to suggest that this white-noise assumption is approximately true for any fluctuating quantity $\mathcal{F}(t)$ that comes from a normal deformation (the term 'normal' will be defined there).

Obviously, the goodness of the 'white-noise assumption' in the two cases mentioned is related to the chaoticity of the system, and should be tested for particular examples. This has been done for the cavity of figure 1 (see [14], and figures 4 and 9). This cavity is an example of a 'scattering billiard' and so exhibits strong chaos [10]. If the motion is *not* strongly chaotic we may get a $C(\tau)$ that decays like a power law (say $1/\tau^{1-\gamma}$ with $0 < \gamma \leqslant 1$) rather than an exponential [10], [21], [22], [23][5], [24]. In such a case the universal behaviour is modified: we get $\omega^{-\gamma}$ behaviour for $\tilde{C}_{\mathrm{E}}(\omega)$ at small frequencies (ν_{E} diverges), signifying faster-than-diffusive energy spreading in equation (A.2) [24]. The stadium is an example where such a complication may arise: an ergodic trajectory can remain in the marginally stable 'bouncing ball' orbit (between the top and bottom edges) for long times, with a probability scaling as t^{-1} [21–23]. Depending on the choice of $D(s)$ this *may* manifest itself in $C(\tau)$. For example, in figure 3 the deformation P involves a distortion confined to the upper edge, and the resulting sensitivity to the bouncing ball orbit leads to large enhancement of the fluctuation intensity $\tilde{C}(\omega = 0)$, and is suggestive of singular behaviour for small ω.

If the billiard has a mixed phase space (which is the generic case), then the integrable component does not contribute to diffusive energy spreading. Proposals have been made to account for this via a phase-space volume factor [9].

[5] The time of crossover to algebraic decay is discussed in [23].

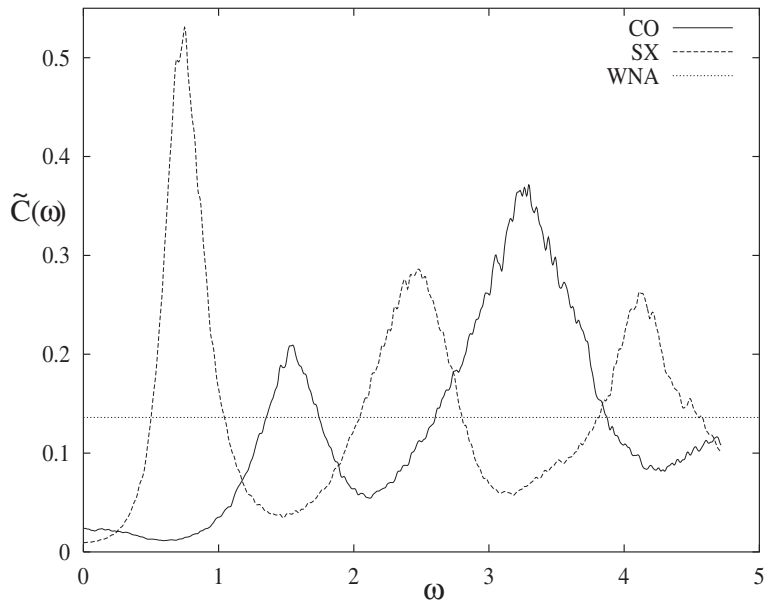


Figure 8. The failure of the WNA estimate for $\tilde{C}(\omega)$ for deformation types CO (similar to DI) and SX (similar to TX). The WNA is clearly a vast overestimate of the small-ω limit. See tables 1 and 2 for an explanation of deformation types.

8. Decomposition of general deformations

The failure of the WNA for 'special' deformations also extends to the much wider class of deformations which are *similar* to special. This is demonstrated in figure 8. It should be emphasized that this failure happens even if the cavity is strongly chaotic.

We seek an analytical estimate for $\tilde{C}(\omega)$, and in particular for its zero-frequency limit ν. This estimate should apply to any (general) deformation, including the case of 'close-to-special' deformations. It would be useful to regard any general deformation as a combination of a 'special' component and a 'normal' component. The formulation of this idea is the theme of the present section. Supporting numerical evidence is gathered in the next section.

The special deformations (for which we have $\nu = 0$) constitute a linear space, meaning that any sum of special deformations is also a special one. Now we would like to conjecture that there is also a linear space of 'normal' deformations. By definition, for 'normal' deformation $\mathcal{F}(t)$ looks like an uncorrelated random sequence of impulses, and consequently the WNA is a reasonable approximation. The notion of randomness can be better formulated as in appendix C leading to equation (C.4). However in practice (C.4) is not useful, because it cannot be applied as an actual classification tool. (Equation (C.4) is never satisfied exactly.) Still we are going to demonstrate that there is a *unique* way to identify the subspace of normal deformations, if we insist on a maximal (i.e. the most inclusive) definition of this subspace.

It is important to clarify the heuristic reasoning of having a linear space of normal deformations. The $\mathcal{F}(t)$ that corresponds to some normal deformation $D(s)$ looks like white noise. This means that only self-correlations of its spikes are statistically significant. If we have two such generic quantities, say $\mathcal{F}_1(t)$ and $\mathcal{F}_2(t)$, then we expect $\mathcal{F}_1(t) + \mathcal{F}_2(t)$ to share the same property.

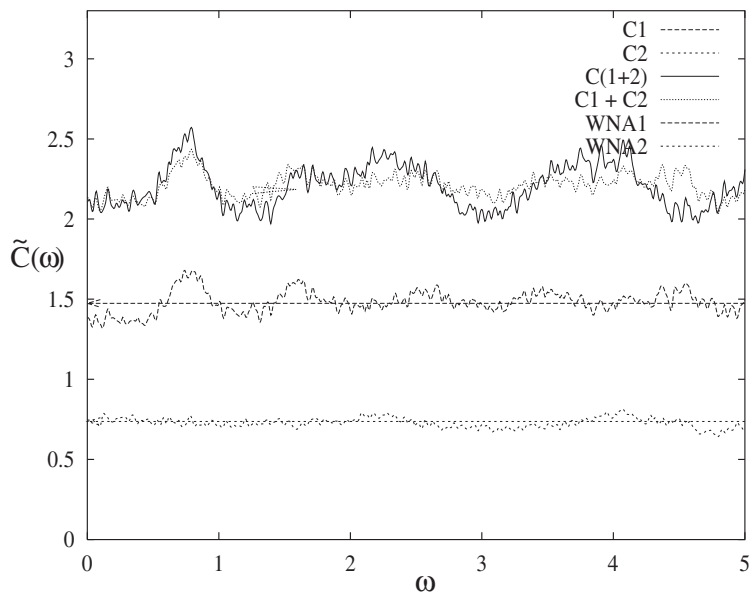


Figure 9. Addition of two 'good' normal deformations (1 = P2, 2 = WG). The two are orthogonal in the sense of (26). That they are 'good' can be seen by their good agreement with their WNA results (horizontal arrows). The power spectrum of the sum agrees well with the sum of the power spectra.

The correlation function of $\mathcal{F}(t) = \mathcal{F}_1(t) + \mathcal{F}_2(t)$ can be written formally as

$$C_{1+2}(\tau) = C_1(\tau) + C_2(\tau) + 2C_{1,2}(\tau) \tag{23}$$

where $C_{1,2}(\tau)$ is the *cross-correlation function*. In appendix B we argue the following:

$$\int_{-\infty}^{\infty} C_{1,2}(\tau)\,\mathrm{d}\tau = 0 \qquad \text{if} \quad 1 = \text{general} \quad 2 = \text{special}. \tag{24}$$

This result is exact, and does not involve any approximation. In appendix C we argue the following:

$$C_{1,2}(\tau) \approx c \times \left[\oint D_1(s)D_2(s)\,\mathrm{d}s\right]\delta(\tau) \qquad \text{if} \quad 1 = \text{normal} \quad 2 = \text{general} \tag{25}$$

where $c = 2m^2 v_{\mathrm{E}}^3 \langle|\cos\theta|^3\rangle/\mathsf{V}$. This result is an approximation, which is expected to be as good as our assumption regarding the 'normality' of the deformation $D_1(s)$. Consider now the case where $D_1(s)$ is normal and $D_2(s)$ is special. Both equations (24) and (25) should apply. But these equations are consistent if and only if $D_1(s)$ is orthogonal to $D_2(s)$. We say that $D_1(s)$ and $D_2(s)$ are orthogonal ($1 \perp 2$) using the following definition:

$$\text{orthogonality} \Leftrightarrow \oint D_1(s)D_2(s)\,\mathrm{d}s = 0. \tag{26}$$

Thus we have proved that normal deformations must be orthogonal (in the sense of (26)) to special deformations. Obviously we have proved here a necessary rather than a sufficient condition for 'normality'. However, if we insist on a maximal definition for the subspace of

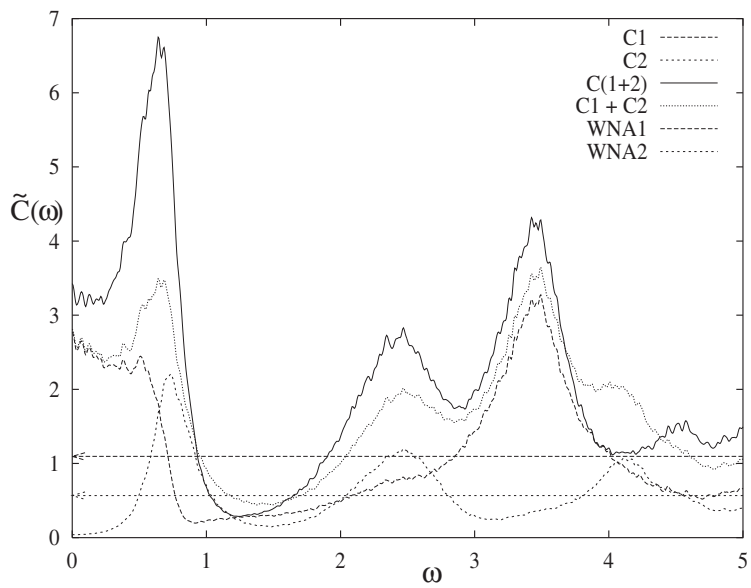


Figure 10. Addition of two 'bad' normal deformations (1 = FR, 2 = SX). The two are orthogonal in the sense of (26). That they are 'bad' is shown by a lack of agreement with their WNAs. The power spectrum of the sum is badly approximated by the sum of the power spectra (nonlinear addition).

normal deformations, then we get a unique identification. Namely, a deformation is classified as 'normal' if it is orthogonal to the subspace of special deformations.

The practical consequences of equations (24) and (25) are as follows:

$$\nu_{1+2} = \nu_1 \qquad \text{if} \quad 1 = \text{general} \quad 2 = \text{special} \tag{27}$$

and

$$\nu_{1+2} \approx \nu_1 + \nu_2 + 2c \oint D_1(s) D_2(s)\, \mathrm{d}s \qquad \text{if} \quad 1 = \text{normal} \quad 2 = \text{general}. \tag{28}$$

These results are tested in the next section.

9. Addition of deformations: numerical tests

On the basis of the discussion in the previous section we *define* normal deformation as those that are orthogonal to all special deformations, in the sense of equation (26). Obviously there are 'good' normal deformations for which the WNA is an excellent approximation (P1 and W8 in figure 4, for example), and there are 'bad' normal deformations for which the WNA is not a very good approximation (FR in figure 4, and the normal component in figure 14(*b*)). In this section we present numerical evidence that verifies the theoretical results of the previous section, and investigate how 'bad' a normal deformation has to be for them to break down.

From what we have claimed it follows that if $D_1(s)$ and $D_2(s)$ are orthogonal normal deformations, then $\nu_{1+2} = \nu_1 + \nu_2$. We could as well write

$$\tilde{C}_{1+2}(\omega) \approx \tilde{C}_1(\omega) + \tilde{C}_2(\omega) \qquad \text{if} \quad 1 = \text{normal} \quad 2 = \text{normal} \qquad \text{and} \qquad 1 \perp 2 \tag{29}$$

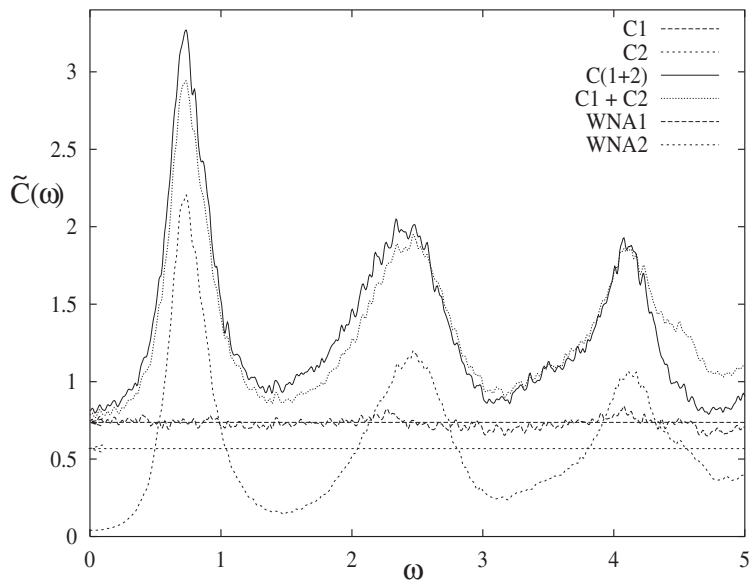


Figure 11. Addition of a 'good' normal deformation (1 = WG) to a general deformation (2 = SX). The two are orthogonal in the sense of (26). The power spectrum of the sum agrees well with the sum of the power spectra.

because by assumption the three correlation functions are approximately flat. We demonstrate this addition rule in the case of two 'good' deformations which are orthogonal in figure 9. We found that small 'pistons' (P2 is significant on only $\sim 1/50$ of the perimeter) were needed to achieve addition of the accuracy (a few %) shown. However, the restriction on the 'wiggle' type of deformation was somewhat more lenient (WG is $\sim$5 times wider than P2 yet obeys the WNA better than P2 does).

In general we observe that the quality of the addition rule is limited by the deviation from the WNA of the *better* of the two deformations. In figure 10 we see that if both $D_1(s)$ and $D_2(s)$ are bad, then also the addition rule (29) becomes quite bad. Figure 11 shows that the addition rule (29) is reasonably well satisfied also if *either* $D_1(s)$ or $D_2(s)$ is a 'good' normal deformation. We have chosen $D_1(s)$ as WG (good), and $D_2(s)$ as SX which is almost completely dominated by the special x-translation deformation. The addition rule (29) is obeyed at all ω. This proves that our assertions equation (25) about the vanishing of $C_{1,2}^{\text{non-self}}(\tau)$ is indeed correct. It holds here as a non-trivial statement ($D_2(s)$ is general and 'bad').

Finally, we consider the case where $D_1(s)$ is general and $D_2(s)$ is special. This is illustrated in figure 12. The addition rule (29) becomes exact in the limit of small frequency corresponding to the vanishing of $\tilde{C}_{1,2}(\omega \to 0)$ as implied by equation (24). In particular this implies that $\nu_{1+2} = \nu_1$. This will be the key to for improving over the WNA, which we are going to discuss in the next section.

In drawing the above conclusions it is important to note that symmetry effects can play a deceptive role if the cavity shape has symmetry (our example figure 1 is in the $C_{2\text{v}}$ symmetry group). In figure 13 we demonstrate that the addition rule (29) is very accurately satisfied at *all* ω if $D_1(s)$ and $D_2(s)$ belong to different symmetry classes of the cavity. Orthogonality of $D_1(s)$ and $D_2(s)$ is *not* sufficient to explain this perfect linearity of addition of $\tilde{C}_{\text{E}}(\omega)$. Rather,

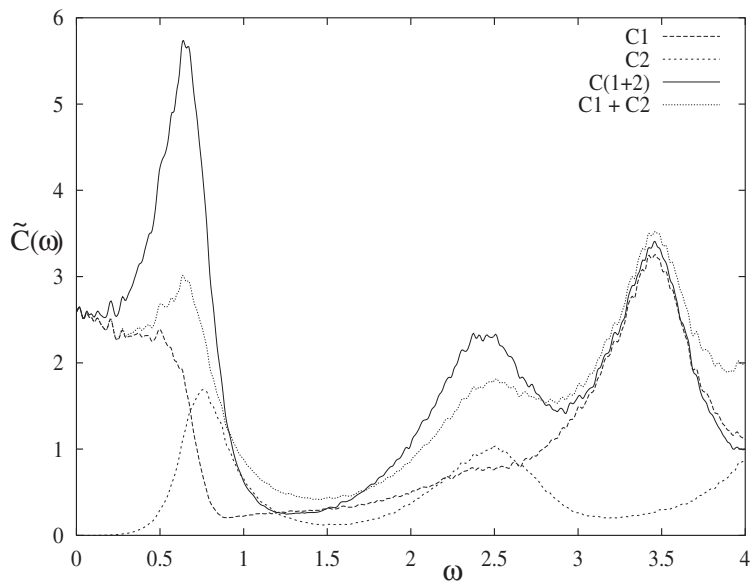


Figure 12. Addition of a general deformation (1 = FR) to 'special' deformation (2 = TX). The power spectrum of the sum coincides with the sum of the power spectra in the limit $\omega \to 0$, as implied by equation (27).

it follows from the symmetry of the kernel $\gamma_{\mathrm{E}}(s_1, s_2)$ of equation (12). The cross-terms in (12) rigorously vanish when such deformations are added. The consequence is that in order to non-trivially demonstrate the assertions of this and of the previous section, we had to choose deformations of the *same* symmetry class, or which break all symmetries of the cavity.

10. Beyond the WNA

It is possible now to consider the case of general deformation, and to go beyond the WNA. Given a general deformation $D(s)$ we should project out (subtract) all the special components, leaving the normal component, and only then apply the WNA. In figure 14 we demonstrate this decomposition for the deformation (CO + W16) and the deformation SX.

The special deformations constitute a linear space which is spanned by the basis functions: one dilation, d translations, and $d(d-1)/2$ rotations. (For $d = 2$ they are listed in table 2.) For a general cavity shape these basis functions are not orthogonal. However, because they are linearly independent, we can use standard linear algebra to build an orthonormal basis $\{D_i(s)\}$ of special deformations. The special ($\|$) and the normal ($\perp$) components of any given deformation $D(s)$ are therefore

$$\begin{aligned} D_{\|}(s) &= \sum_i \alpha_i D_i(s) \\ D_{\perp}(s) &= D(s) - D_{\|}(s) \end{aligned} \tag{30}$$

where the coefficients are

$$\alpha_i = \oint D(s) D_i(s)\, \mathrm{d}s. \tag{31}$$

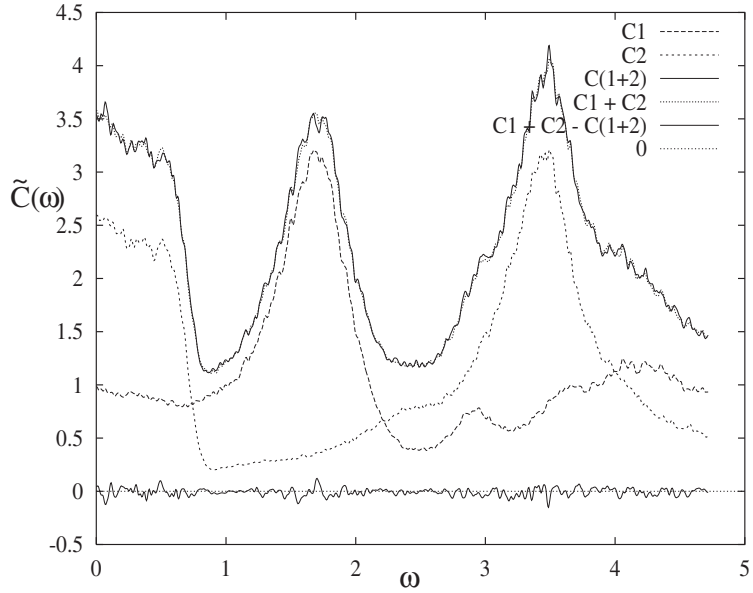


Figure 13. Addition of two 'bad' general deformations which come from different symmetry classes of the cavity (1 = W2, 2 = FR). The two must also be orthogonal, by symmetry. The deviation from linear addition (solid curve varying about zero) vanishes at all ω.

The improved approximation for ν applies the WNA only to the normal component, giving

$$\nu_{\mathrm{E}} \approx 2m^2 v_{\mathrm{E}}^3 \langle |\cos\theta|^3 \rangle \frac{1}{\mathsf{V}} \oint \mathrm{d}\boldsymbol{s}\, [D_\perp(\boldsymbol{s})]^2 \tag{32}$$

which we name the IFIF (improved fluctuation intensity formula). In the particular case of $d = 3$, substitution of this result into the microcanonical FD relation gives an 'improved wall formula' consisting of the replacement of $D(\boldsymbol{s})$ by $D_\perp(\boldsymbol{s})$ in equation (5).

In figure 14 we use the IFIF to estimate ν for two examples. The first is a deformation (CO+W16) whose normal component is 'good', due its oscillatory nature. The deviation from a flat white power spectrum is ∼20% for the normal component. The IFIF result equation (32) is accurate to a few per cent. It is a much better estimate of the actual ν compared with the naive WNA equation (14), which overestimates the correct value by a factor of 2.2. In the second example the deformation is SX. The resulting normal component is 'bad'. Its power spectrum fluctuates by a factor of about 10 in the ω range shown. Consequently the IFIF is limited in its accuracy, and the correct value for ν is underestimated by a factor of 2.5. However, it is still a great improvement over the naive result equation (14). In this second example we can extract another prediction about $\tilde{C}_{\mathrm{E}}(\omega)$. The special component is a factor ∼10 larger than the normal component. Therefore the ω^2 behaviour at small ω is almost entirely due to the 'rotation' component. The prefactor of the ω^2 behaviour need only be found once for each billiard shape (see section 6). This saves computation and gives extra information about the dissipation rate at finite driving frequency.

A few concluding remarks regarding the history of the wall formula are in order. It has been known since its inception that the naive wall formula gives unphysical answers in the case of *constant-velocity* translations and rotations. This was first regarded as a kinetic

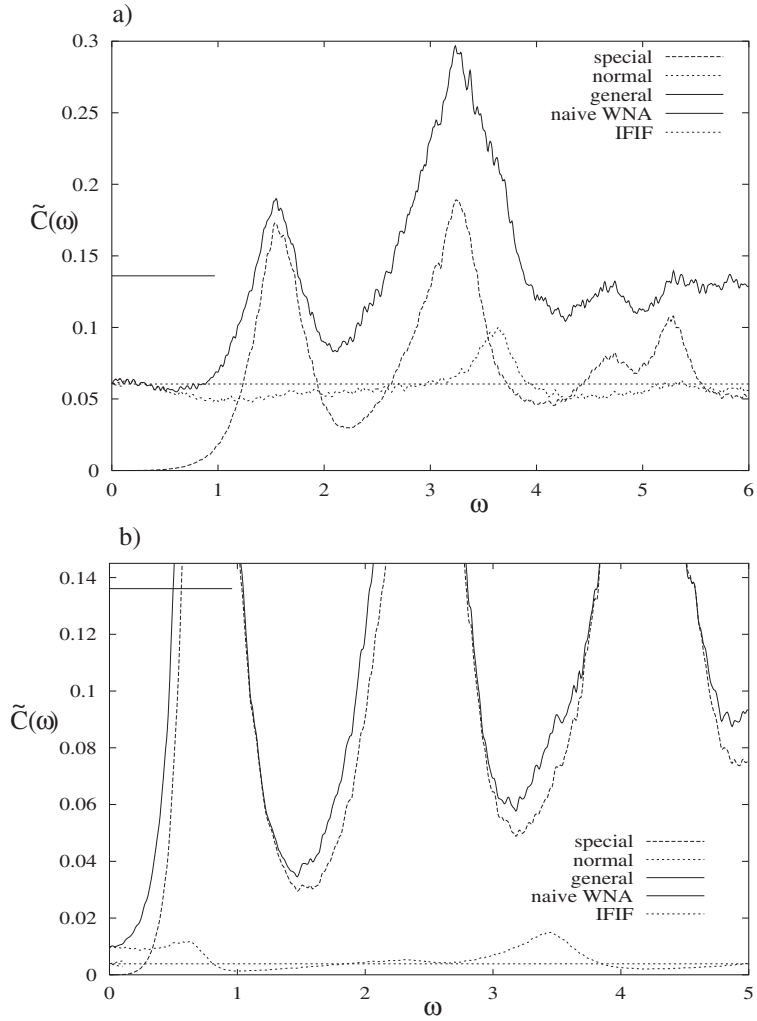


Figure 14. Decomposition of general deformations $D(s)$ into orthogonal 'normal' and 'special' components. The general deformation is CO + W16 in subfigure (*a*), and SX in subfigure (*b*). The naive WNA equation (14) is indicated by a short solid line. The improved (IFIF) result equation (32) is indicated by a long dashed arrow.

gas 'drift' effect [2]. It should be noted that the recipe presented in [2], namely to subtract this drift component, is equivalent in practice to the recipe (30) that we have presented here, *provided we ignore dilations*. It is also important to realize that the argumentation in [2] for this subtraction appears to be *ad hoc*, being based on a 'least-structured drift pattern' reasoning. A stated condition on this subtraction was that the resulting deformation preserve the location of the 'centre of mass' (centroid) of the cavity, for reasons particular to the nuclear application [2]. This condition seems to have become standard practice in numerical tests of the wall formula [9, 25–27]. However, as the system in figure 15(*a*) illustrates, this condition

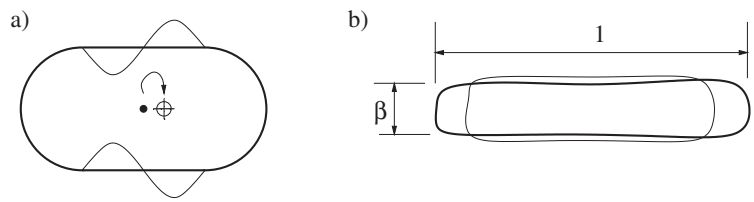


Figure 15. (*a*) A deformation of the stadium which moves the 'centre of mass' (centroid) of the cavity to the right (from the dot to the crosshairs symbol). This deformation is orthogonal (in the sense of (26)) to all special deformations, in particular, all translations. (*b*) An example volume-preserving deformation of an elongated approximately rectangular cavity ($\beta \ll 1$), which nevertheless has a large overlap with dilation. It can be shown that this results in an IFIF estimate of $\approx 4\beta$ times that of the naive WNA. In both diagrams the undeformed shape is shown as a heavy line, the deformed one as a thin line.

is generally *not* equivalent to the above subtraction of translation and rotation components[6]. This seems to invalidate the theorem presented in section 7.1 of [2]. Where the flaw in their reasoning lies we are not sure.

The consideration of the special nature of dilations is absent from the literature. Even if we restrict ourselves to volume-preserving deformations (the case for the nuclear application), then deformations of certain cavities can be found for which the dilation correction is significant; we illustrate this in figure 15(*b*). This correction can only be large if the cavity has a large variation in radius (i.e. is highly non-spherical). We suggest this as a possible reason why major discrepancies due to dilation have not emerged in the numerical tests of the wall formula until now. Such tests have generally been of shapes close to a 3D sphere [2, 9, 25–27].

Hence we believe that the recipe we have presented, along with the associated theory and in conjunction with the particular power-law dependences, is a significant step in the treatment of one-body dissipation.

Acknowledgments

In particular we would like to thank Eduardo Vergini for stimulating dialogue, which motivated us to consider the special nature of dilations. This paper was funded by ITAMP (at the Harvard-Smithsonian Center for Astrophysics and the Harvard Physics Department) and the National Science Foundation (USA).

Appendix A. Linear response theory of dissipation

Given a parametric Hamiltonian $H(Q, P; x(t))$, and given initial conditions, one defines the energy $\mathcal{E}(t) = H(Q(t), P(t); x(t))$ and the fluctuating quantity $\mathcal{F}(t) = -\partial\mathcal{H}/\partial x(Q(t), P(t); x(t))$. With no approximation we have

$$\mathcal{E}(t) - \mathcal{E}(0) = \int_0^t \mathcal{F}(t')\dot{x}(t')\,\mathrm{d}t'. \tag{A.1}$$

[6] The condition that a deformation $D(s)$ should not move the 'centre of mass' (centroid of the cavity volume) is $\oint D(s)\boldsymbol{r}(s)\,\mathrm{d}s = \mathbf{0}$. This is in general different from the condition for having zero overlap with translations, namely $\oint D(s)\hat{\boldsymbol{n}}(s)\,\mathrm{d}s = \mathbf{0}$.

Using the same steps as in [8] one obtains the following result for the variance of the energy spreading:

$$\delta E(t)^2 = \int_0^t \int_0^t C_{\mathrm{E}}(t'' - t') F(t'' - t')\, \mathrm{d}t'\, \mathrm{d}t'' \tag{A.2}$$

where $C_{\mathrm{E}}(\tau)$ is defined by equation (11). Microcanonical averaging has been taken over the initial conditions. The function $F(\tau) = \langle \dot{x}(t)\dot{x}(t+\tau)\rangle$ is the velocity–velocity correlation of the driving. For periodic driving $x(t) = A\sin(\Omega t + \text{phase})$ it is formally convenient to average over the initial phase and one obtains $F(\tau) = \frac{1}{2}(A\Omega)^2\cos(\Omega\tau)$.

For a chaotic system $C_{\mathrm{E}}(\tau)$ is characterized by some correlation time τ_{cl}. For $t \gg \tau_{\mathrm{cl}}$ one obtains diffusive spreading $\delta E(t)^2 = 2D_{\mathrm{E}}t$ where the diffusion rate is

$$D_{\mathrm{E}} = \tfrac{1}{2}\tilde{C}_{\mathrm{E}}(\Omega) \times \tfrac{1}{2}(A\Omega)^2 \tag{A.3}$$

which for small frequencies goes to $D_{\mathrm{E}} = \frac{1}{2}\nu_{\mathrm{E}}V^2$, where $\nu_{\mathrm{E}} \equiv \tilde{C}(0)$ as defined in the introduction. The picture to keep in mind is that of the fluctuating $\mathcal{F}(t)$ causing a *random walk* in energy space via equation (A.1) for times $t \gg \tau_{\mathrm{cl}}$. As explained in [4, 5, 8] the resulting diffusion in energy space implies systematic *growth* of the average energy. It is important to realize that this growth happens even if the random walk is locally unbiased: such is the case when changing the parameter x preserves the volume of a given energy-shell in phase space. (For a deforming billiard system this corresponds to preservation of the billiard volume.) The rate of energy growth is related to the diffusion as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle \mathcal{H}\rangle = -\int_0^\infty \mathrm{d}E\, g(E) D_{\mathrm{E}} \frac{\partial}{\partial E}\left(\frac{\rho(E)}{g(E)}\right) \tag{A.4}$$

where $\rho(E)$ is the energy distribution of the particles, and $g(E)$ is the one-particle density of states. The growth is therefore an effect of the E-dependence of both the diffusion rate and the density of states.

The rate of dissipation can be written as in equation (1) or as $\mathrm{d}\langle\mathcal{H}\rangle/\mathrm{d}t = \mu V^2$ in the small-frequency limit. Combining this with equation (A.4) implies a relation between the dissipation coefficient μ and the function $\tilde{C}_{\mathrm{E}}(\omega)$. The most familiar version of this FD relation is obtained for small frequency under the assumption of a canonical distribution $\rho(E) \propto g(E)\exp(-E/(k_{\mathrm{B}}T))$, leading to

$$\mu = \frac{1}{2k_{\mathrm{B}}T}\nu \tag{A.5}$$

where ν should be calculated for a canonical distribution. This result should be multiplied by the number of non-interacting classical particles.

The use of equation (A.4) can be justified also for non-interacting fermions [28]. This is because the effect of the Pauli exclusion principle cancels out (in analogy with the Boltzmann picture with elastic scattering). Substituting $\rho(E) = g(E)f(E - E_{\mathrm{F}})$, where $f(E - E_{\mathrm{F}})$ is the Fermi occupation function, one obtains

$$\mu = \tfrac{1}{2}g(E_{\mathrm{F}})\nu_{\mathrm{F}} \tag{A.6}$$

where ν_{F} should be calculated at the Fermi energy.

Finally, the microcanonical version of the FD relation is

$$\mu_{\mathrm{E}} = \frac{1}{2}\frac{1}{g(E)}\frac{\partial}{\partial E}(g(E)\nu_{\mathrm{E}}). \tag{A.7}$$

The subscript E indicates that both ν_{E} and μ_{E} are evaluated locally around some energy E.

Appendix B. Cross correlations I

In this appendix we introduce two proofs of equation (24). The first is a formal argument, while the second is a more physically appealing argument. The formal argument is as follows: equation (12) is an exact result which can be written using obvious abstract matrix notation as $\nu = D\gamma_{\mathrm{E}}D$. Let $D = D_1 + D_2$. If D_2 is a special deformation then by definition $D_2\gamma_{\mathrm{E}}D_2 = 0$. But this can be true only if D_2 belongs to the kernel (nullspace) of the matrix γ_{E}, hence we have $\gamma_{\mathrm{E}}D_2 = 0$. Therefore we have also $D_1\gamma_{\mathrm{E}}D_2 = 0$ for any D_1, which is precisely the statement of equation (24).

Now we present the alternative physically appealing argument. Consider two noisy signals $\mathcal{F}(t)$ and $\mathcal{G}(t)$. We assume that $\langle\mathcal{F}(t)\rangle = \langle\mathcal{G}(t)\rangle = 0$. The angular brackets stand for an average over realizations. The auto-correlations of $\mathcal{F}(t)$ and $\mathcal{G}(t)$ are described by functions $C_{\mathrm{F}}(\tau)$ and $C_{\mathrm{G}}(\tau)$ respectively. We assume that both auto-correlation functions are short range, meaning no power-law tails (this corresponds to the hard-chaos assumption of this paper), and that they are negligible beyond a time τ_{c}. We call a signal 'special' if the algebraic area under its auto-correlation is zero. The cross-correlation function is defined as

$$C_{\mathrm{F,G}}(\tau) \equiv \langle\mathcal{F}(t')\mathcal{G}(t'')\rangle \qquad \tau \equiv t' - t''. \tag{B.1}$$

We assume stationary processes so that the cross-correlation function depends only on the time difference τ. We also symmetrize this function if it does not have $\tau \mapsto -\tau$ symmetry. We assume that $C_{\mathrm{F,G}}(\tau)$ is short range, meaning that it becomes negligibly small for $|\tau| > \tau_{\mathrm{c}}$. We would like to prove that if either $\mathcal{F}(t)$ or $\mathcal{G}(t)$ is special then the algebraic area under the cross-correlation function equals zero.

Consider the case where $\mathcal{F}(t)$ is general while $\mathcal{G}(t)$ is special. The integral of $C_{\mathrm{F}}(\tau)$ will be denoted by ν. Define the processes

$$X(t) = \int_0^t \mathcal{F}(t')\,\mathrm{d}t' \tag{B.2}$$

$$Y(t) = \int_0^t \mathcal{G}(t'')\,\mathrm{d}t''. \tag{B.3}$$

From our assumptions it follows, disregarding a transient, that for $t \gg \tau_{\mathrm{c}}$ we have diffusive growth $\langle X(t)^2\rangle \approx \nu t$. However since $Y(t)$ is a stationary process [29], $\langle Y(t)^2\rangle \approx$ const. Therefore for a typical realization we have $|X(t)| \leqslant \text{const} \times \sqrt{\nu t}$ and $|Y(t)| \leqslant$ const. Consequently, without making any claims on the independence of $X(t)$ and $Y(t)$, we get that $\langle X(t)Y(t)\rangle$ cannot grow faster than const $\times \sqrt{\nu t}$. Using the definitions (B.1)–(B.3) we can write

$$\int_{-\infty}^{\infty} C_{\mathrm{F,G}}(\tau)\,\mathrm{d}\tau = \frac{\langle X(t)Y(t)\rangle}{t} \approx \frac{\text{const}}{\sqrt{t}} \to 0 \tag{B.4}$$

where the limit $t \to \infty$ is taken. Thus we have proved our assertion.

Appendix C. Cross correlations II

In this section we further discuss some features of the cross-correlation function. For the purpose of presentation we we would like to view the time as an integer variable $t = 1, 2, 3, \ldots$. One may think of each instant of time as corresponding to a bounce.

Let us assume that we have functions $f(s)$ and $g(s)$, and a time-sequence $(s_1, s_2, s_3, \ldots)$. This gives two stochastic-like processes $(\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \ldots)$ and $(\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \ldots)$. The cross correlation of these two processes is defined as follows:

$$C_{\mathrm{F,G}}(i - j) = \langle\mathcal{F}_i\mathcal{G}_j\rangle = \langle f(s_i)g(s_j)\rangle. \tag{C.1}$$

It is implicit in this definition that we assume that the processes are stationary, so the result depends only on the difference $\tau = (i - j)$. The angular brackets stand for an average over realizations of s-sequences.

If the sequences are ergodic on the s-domain, then it follows that

$$\begin{aligned} \langle \mathcal{F} \rangle &= \int f(s)\,\mathrm{d}s \\ \langle \mathcal{G} \rangle &= \int g(s)\,\mathrm{d}s \\ C_{\mathrm{F,G}}(0) &= \int f(s)g(s)\,\mathrm{d}s. \end{aligned} \tag{C.2}$$

The $\tau \neq 0$ cross correlation requires information beyond mere ergodicity. In the case where the s-sequence is completely uncorrelated in time we can factorize the averaging and we get $C_{\mathrm{F,G}}(\tau \neq 0) = \langle \mathcal{F} \rangle \times \langle \mathcal{G} \rangle$. If $\langle \mathcal{F} \rangle = 0$ then

$$C_{\mathrm{F,G}}(\tau \neq 0) = 0 \tag{C.3}$$

irrespective of $\langle \mathcal{G} \rangle$.

However, we would like to define circumstances in which equation (C.3) is valid, even if the s-sequence is *not* uncorrelated. In such a case either the $\mathcal{F}$ or the $\mathcal{G}$ may possess time correlations. (Such is the case if $\mathcal{G}$ is 'special'.) So let us consider the case where the $\mathcal{F}$ sequence *looks random*, while assuming nothing about the $\mathcal{G}$ sequence. By the phrase 'looks random' we mean that the conditional probability satisfies

$$\mathrm{Prob}(\mathcal{F}_i | s_j) = \mathrm{Prob}(\mathcal{F}_i) \qquad \text{for any} \quad i \neq j. \tag{C.4}$$

Equation (C.3) straightforwardly follows provided $\langle \mathcal{F} \rangle = 0$, irrespective of the $g(s)$ involved. Given $f(s)$, the goodness of assumption (C.4) can be actually tested. However, it is not convenient to consider (C.4) as a practical definition of a 'normal' deformation.

References

[1] For review and references, see Fishman S 1993 *Proc. Int. School of Physics Enrico Fermi, Course CXIX* ed G Casati, I Guarneri and U Smilansky (Amsterdam: North-Holland)

[2] Blocki J, Boneh Y, Nix J R, Randrup J, Robel M, Sierk A J and Swiatecki W J 1978 *Ann. Phys., NY* **113** 330

[3] Koonin S E, Hatch R L and Randrup J 1977 *Nucl. Phys.* A **283** 87
Koonin S E and Randrup J 1977 *Nucl. Phys.* A **289** 475

[4] Jarzynski C 1992 *Phys. Rev.* A **46** 7498

[5] Jarzynski C 1993 *Phys. Rev.* E **48** 4340

[6] Austin E J and Wilkinson M 1993 *J. Phys.: Condens. Matter* **5** 8461
See also Austin E J and Wilkinson M 1994 *J. Phys.: Condens. Matter* **6** 4153

[7] Cohen D 2000 *Proc. Int. School of Physics Enrico Fermi, Course CXLIII* ed G Casati, I Guarneri and U Smilansky (Amsterdam: IOS Press)

[8] Cohen D 2000 *Ann. Phys., NY* **283** 175–231

[9] Pal S and Mukhopadhyay T 1996 *Phys. Rev.* C **54** 1333–40
Mukhopadhyay T and Pal S 1997 *Phys. Rev.* C **56** 296–301

[10] Bunimovich L A 1985 *Sov. Phys.–JETP* **62** 842

[11] Wojtkowski M 1986 *Commun. Math. Phys.* **105** 391–414

[12] Cohen D 1999 *Phys. Rev. Lett.* **82** 4951

[13] Cohen D and Kottos T 2000 *Phys. Rev. Lett.* **85** 4839
(Cohen D and Kottos T 2000 *Preprint* cond-mat/0004022)

[14] Barnett A H, Cohen D and Heller E J 2000 *Phys. Rev. Lett.* **85** 1412

[15] Cohen D and Kottos T 2001 *Phys. Rev.* E at press
(Cohen D and Kottos T 2000 *Preprint* nlin.CD/0001026)

[16] Cohen D, Barnett A H and Heller E J 2001 *Phys. Rev.* E at press
(Cohen D, Barnett A H and Heller E J *Preprint* nlin.CD/0008040)

[17] Vergini E and Saraceno M 1995 *Phys. Rev.* E **52** 2204
[18] Barnett A H 2000 *PhD Thesis* Harvard University
[19] Vergini E 1995 *PhD Thesis* Universidad de Buenos Aires
[20] Hammel S M, Yorke J A and Grebogi C 1987 *J. Complexity* **3** 136
Grebogi C, Hammel S M, Yorke J A and Sauer T 1990 *Phys. Rev. Lett.* **65** 1527
[21] Friedman B and Martin R F Jr 1984 *Phys. Lett.* A **105** 23
[22] Dahlqvist P and Artuso R 1996 *Phys. Lett.* A **219** 212
[23] Dahlqvist P 1999 *Phys. Rev.* E **60** 6639
[24] Brown R, Ott E and Grebogi C 1987 *J. Stat. Phys.* **49** 511
[25] Blocki J, Brut F and Swiatecki W J 1993 *Nucl. Phys.* A **554** 107
[26] Blocki J, Shi Y-J and Swiatecki W J 1993 *Nucl. Phys.* A **554** 387
[27] Blocki J, Skalski J and Swiatecki W J 1995 *Nucl. Phys.* A **594** 137
[28] Wilkinson M 1990 *J. Phys. A: Math. Gen.* **23** 3603
[29] Gardiner C W 1983 *Handbook of Stochastic Methods* (Berlin: Springer)

Deformations and Dilations of Chaotic Billiards: Dissipation Rate, and Quasiorthogonality of the Boundary Wave Functions

Alex Barnett, Doron Cohen, and Eric J. Heller
Department of Physics, Harvard University, Cambridge, Massachusetts 02138
(Received 9 March 2000)

We consider chaotic billiards in d dimensions, and study the matrix elements M_{nm} corresponding to general deformations of the boundary. We analyze the dependence of $|M_{nm}|^2$ on $\omega = (E_n - E_m)/\hbar$ using semiclassical considerations. This relates to an estimate of the energy dissipation rate when the deformation is periodic at frequency ω. We show that, for dilations and translations of the boundary, $|M_{nm}|^2$ vanishes like ω^4 as $\omega \to 0$, for rotations such as ω^2, whereas for generic deformations it goes to a constant. Such special cases lead to quasiorthogonality of the eigenstates on the boundary.

PACS numbers: 05.45.Mt, 03.65.Sq

Chaotic cavities (billiards) in d dimensions are prototype systems for the study of classical chaos and its fingerprints on the properties of the quantum-mechanical eigenstates. As the properties of static billiards are beginning to be understood, questions naturally arise about deformations and their time dependence. It is perhaps not widely appreciated that certain deformations are very special, and that there is a close connection between the quantum and classical mechanics of such deformations in the case of ergodic systems. In this paper, which takes a fresh approach to these issues, we explore a special class of deformations which do not "heat" in the limit of small frequencies. We also establish a rather surprising relationship to a very successful numerical technique for finding billiard eigenfunctions.

We start with the one-particle Hamiltonian $\mathcal{H}_0(\mathbf{r},\mathbf{p}) = \mathbf{p}^2/2m + V(\mathbf{r})$, where m is the particle mass, $\mathbf{r}$ is the position of the particle inside the cavity, and $\mathbf{p}$ is the conjugate momentum. We will take the limit $V(\mathbf{r}) \to \infty$ outside the cavity, zero otherwise, corresponding to Dirichlet boundary conditions. In this limit, the Hamiltonian is completely defined by the boundary shape. We call the volume of the cavity $\mathsf{V} \equiv \mathsf{L}^d$. Upon quantization a second length scale $\lambda_{\mathrm{B}} \equiv 2\pi/k$ appears, where k is the wave number. For simple geometries the typical time between collisions with the walls is $\tau_{\mathrm{col}} \sim \mathsf{L}/v$, where v is the particle speed. The energy is $E = \frac{1}{2}mv^2$. Upon quantization the eigenenergies are $E_n = (\hbar k_n)^2/2m$.

A powerful tool for the classical analysis is known as the "Poincaré section." Rather than following trajectories in the full $(\mathbf{r},\mathbf{p})$ phase space, it is much more efficient to record only successive collisions with the boundary. This way we can deal with a canonical transformation (map) which is defined on a $2(d-1)$-dimensional phase space. A similar idea is used in quantum mechanics: By using Green's theorem, it is clear that all of the information about an eigenstate $\psi(\mathbf{r})$ is contained in the boundary normal derivative function $\varphi(\mathbf{s}) \equiv \mathbf{n}\cdot\nabla\psi$, where $\mathbf{s}$ is a $(d-1)$-dimensional coordinate on the boundary, and $\mathbf{n}(\mathbf{s})$ is the outward unit normal vector.

However, unlike the classical case, the reduction to the boundary is not satisfactory. One cannot define an associated Hilbert space that consists of the boundary functions. In particular, the orthogonality relation $\langle\psi_n \mid \psi_m\rangle = \delta_{nm}$ does not have an exact analog on the boundary. Still, the boundary functions "live" in an effective Hilbert space of dimension $\sim(\mathsf{L}/\lambda_{\mathrm{B}})^{d-1}$, and it has been realized [1] that the following quasiorthogonality relation holds. Define an inner product

$$M_{nm} \equiv \frac{1}{2k^2}\oint \varphi_n(\mathbf{s})\varphi_m(\mathbf{s})\,(\mathbf{n}\cdot\hat{\mathbf{D}})\,d\mathbf{s}\,, \qquad (1)$$

where $\mathbf{D}(\mathbf{s}) = \mathbf{r}(\mathbf{s})$ is the displacement field corresponding to dilation (about an arbitrary origin), and $k_n \approx k_m \approx k$ [2]. It is well known that the normalization condition $\langle\psi_n \mid \psi_n\rangle = 1$ implies $M_{nn} = 1$. We give a proof of this exact result in the Appendix. On the other hand, the off-diagonal elements are only approximately zero [3].

The main purpose of this Letter is to study the band profile of the matrix M_{nm} for a general displacement field $\mathbf{D}(\mathbf{s})$. In particular we want to understand why, for special choices of $\mathbf{D}(\mathbf{s})$ (notably dilations), we have quasiorthogonality. Later we will explain that M_{nm} can be interpreted as the matrix element of a perturbation $\delta\mathcal{H}$ associated with a deformation of the boundary, such that $(\mathbf{n}\cdot\mathbf{D})\delta x$ is the normal displacement of a wall element, given a control parameter δx. In the following two paragraphs we explain the main motivations for our study.

The matrix elements $|M_{nm}|^2$ determine the rate of irreversible energy absorption by the particle (i.e., dissipation) due to external driving. Here "external driving" means time-dependent deformation of the boundary. Having exceptionally small $|M_{nm}|^2$ for special choices of $\mathbf{D}(\mathbf{s})$, such as dilations, translations, and rotations, implies an exceptionally small dissipation rate ("nonheating" effect). This observation goes against the naive kinetic picture that the rate of heating should not depend on how we "shake" the boundary. The special nature of translations and rotations for $\omega = 0$ has been recognized in the context of nuclear

 0031-9007/00/85(7)/1412(4)$15.00 © 2000 The American Physical Society

dissipation [4,5]. Our present approach allows us to analyze the nonheating effect present for dilations as well, and provide the form of the low-frequency response of the system in all three cases (dilations, translations, and rotations).

There is another good motivation to study this issue. Recently, a powerful technique for finding clusters of billiard eigenstates and eigenenergies has been found by Vergini and Saraceno [1,6], with a speed typically $\sim 10^3$ greater than previous methods. This efficiency relies on the above quasiorthogonality relation, the associated numerical error being given by the deviation of M_{nm} from δ_{nm}. The authors tried to establish quasiorthogonality using the identity $M_{nm} = \delta_{nm} + [(k_m^2 - k_n^2)/2k^2]B_{nm}$, with $B_{nm} \equiv \langle \psi_n | \mathbf{r} \cdot \nabla | \psi_m \rangle$, and by assuming [7] that $|B_{nm}| \sim O(1)$. However, a naive random-wave argument would predict $|B_{nm}| \sim O(\mathsf{L}/\lambda_{\mathrm{B}})^{(d-1)/2}$.

Figure 2 (below) displays the band profile $|M_{nm}|^2$ for three choices of the displacement field $\mathbf{D}(\mathbf{s})$. The band profile can be regarded as a function of either $\kappa = k_n - k_m$ or, equivalently, $\omega = (E_n - E_m)/\hbar$, related via $\omega = v\kappa$. The three band profiles differ in their peak structure, and also in their $\omega \to 0$ limits: notably for dilations $|M_{nm}|^2$ vanishes in this limit. Our aim is to understand the overall ω dependence, and the small ω behavior in particular. For the calculation of band profile we used all 451 eigenstates of the 2D quarter stadium (see Fig. 1) lying between $398 < k < 402$, found by using the method of Vergini and Saraceno [1]. For this particular chaotic shape a remarkably good basis set (size of order $\mathsf{L}/\lambda_{\mathrm{B}}$) of real and evanescent plane waves has been devised [6], which allows the *tension* error (defined as the boundary integral of ψ^2) to be typically 3×10^{-11} in our calculation, (maximum 2×10^{-10} for any state). The resulting errors in φ manifest themselves only when $|M_{nm}|^2$ reaches its lowest reliable value $\sim 10^{-10}$, visible as bottoming out in the leftmost point of the inset of Fig. 2.

In order to understand the quantum-mechanical band profile, we can first assume that the eigenstates look like uncorrelated random waves. A lengthy but straightforward calculation [8] leads to the result

$$|M_{nm}|^2 \approx \frac{2\langle |\cos(\theta)|^3 \rangle}{\Omega_d} \frac{\lambda_{\mathrm{B}}^{d-1}}{\mathsf{V}^2} \oint (\mathbf{n} \cdot \mathbf{D})^2 \, d\mathbf{s}\,, \qquad (2)$$

where the geometric factors for $d = 2, 3, \ldots$ are $\Omega_d = 2\pi, 4\pi, \ldots$ and $\langle |\cos(\theta)|^3 \rangle = 4/(3\pi), 1/4, \ldots$. If the displacement field is normalized such that $|\mathbf{D}| \sim \mathsf{L}$, then we get $|M_{nm}|^2 \sim (\lambda_{\mathrm{B}}/\mathsf{L})^{d-1}$. Note that the above result implies that $|M_{nm}|^2$ is independent of ω.

To go beyond the random-wave estimate (2), we adopt a more physically appealing point of view. We include a parametric deformation of the billiard shape via the Hamiltonian $\mathcal{H}(\mathbf{r}, \mathbf{p}; x) = \mathbf{p}^2/2m + V(\mathbf{r} - x\mathbf{D}(\mathbf{r}))$, where x controls the deformation. Note that the displacement field $\mathbf{D}$ is regarded as a function of $\mathbf{r}$. The normal displacement of a wall element is $(\mathbf{n} \cdot \mathbf{D})x$. The position of a particle in the vicinity of a wall element is conveniently described by $Q = (\mathbf{s}, z)$, where $\mathbf{s}$ is a surface coordinate and z is a perpendicular "radial" coordinate. We set $V(\mathbf{r}) = V_0$ outside the undeformed billiard; later we take the limit $V_0 \to \infty$. We have $\partial \mathcal{H}/\partial x = -[\mathbf{n}(\mathbf{s}) \cdot \hat{\mathbf{D}}(\mathbf{s})]V_0 \delta(z)$. The logarithmic derivative with respect to z of an eigenfunction on the boundary is $\varphi(\mathbf{s})/\psi(\mathbf{s})$. For $z > 0$ the wave function $\psi(\mathbf{r})$ is a decaying exponential. Hence the logarithmic derivative of the wave function on the boundary should be equal to $-\sqrt{2mV_0}/\hbar$. Consequently one obtains $(\partial \mathcal{H}/\partial x)_{nm} = -[(\hbar k)^2/m]M_{nm}$. Thus the band profile of M_{nm} is equal (up to a factor) to the band profile of the perturbation $\delta \mathcal{H}$ due to a deformation of the boundary (see also Refs. [5,8,9]).

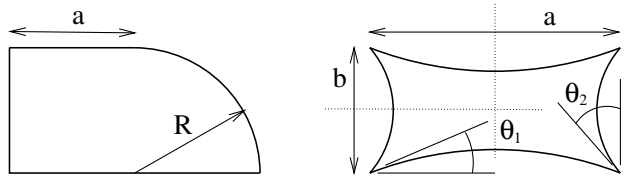


FIG. 1. The two-dimensional billiards under numerical study. Left: Bunimovich quarter stadium ($a/R = 1$). Right: generalized Sinai billiard ($a/b = 2$, $\theta_1 = 0.2$, $\theta_2 = 0.5$).

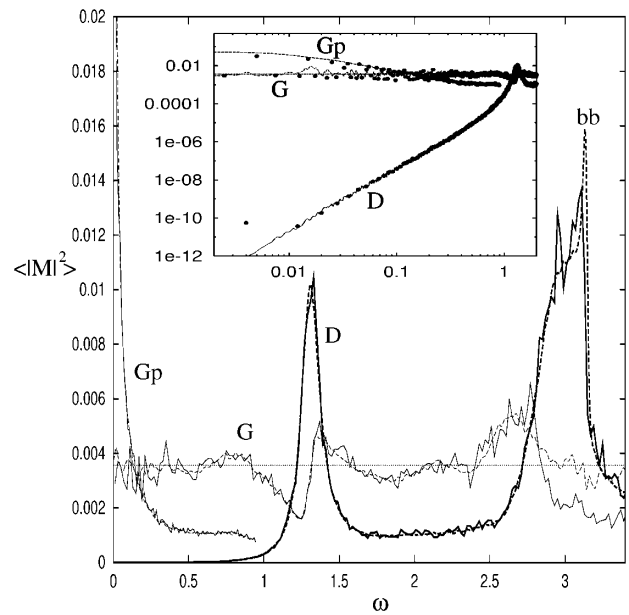


FIG. 2. The band profile in the two-dimensional quarter stadium at $k \approx 400$ for three choices of deformation field: dilation (D), a generic deformation (G), and a generic deformation restricted to parallel displacement of the stadium upper edge (Gp). G and Gp are chosen to be volume preserving. In each case, the solid line is the average $|M_{nm}|^2$ (estimation error 10%) versus $\omega = v(k_n - k_m)$, with $v \mapsto 1$, and the dashed line is the semiclassical estimate [Eq. (4)] (estimation error 3%). We normalized G and Gp so that they share the same random-wave estimate [Eq. (2)] as D; this is shown as a horizontal dotted line. The system-specific peak due to bouncing-ball orbits is labeled (bb). The inset is a log-log plot with average $|M_{nm}|^2$ shown as points.

We can now use semiclassical considerations [10]. The application to the cavity example has been introduced in [8]. Here we summarize the recipe. One should generate a very long (ergodic) classical trajectory, and define for it the fluctuating quantity $\mathcal{F}(t) = -\partial\mathcal{H}(\mathbf{r},\mathbf{p};x)/\partial x|_{x=0}$, where the time dependence of $\mathcal{F}$ is due to the trajectory $(\mathbf{r}(t),\mathbf{p}(t))$. Hence

$$\mathcal{F}(t) = \sum_{\mathrm{col}} 2mv\cos(\theta_{\mathrm{col}})D_{\mathrm{col}}\delta(t - t_{\mathrm{col}}), \qquad (3)$$

where t_{col} is the time of a collision, D_{col} stands for $\mathbf{n}\cdot\mathbf{D}$ at the point of the collision, and $v\cos(\theta_{\mathrm{col}})$ is the normal component of the particle's collision velocity. If the deformation is volume preserving, then $\langle\mathcal{F}(t)\rangle = 0$, otherwise it is convenient to subtract the (constant) average value. Now one can calculate the correlation function $C(\tau)$ of the fluctuating quantity $\mathcal{F}(t)$, and its Fourier transform $\tilde{C}(\omega) \equiv \int C(\tau)\exp(i\omega\tau)\,d\tau$. The semiclassical estimate for the matrix element is

$$\left\langle\left|\left(\frac{\partial\mathcal{H}}{\partial x}\right)_{nm}\right|^2\right\rangle \approx \frac{\Delta}{2\pi\hbar}\tilde{C}\left(\frac{E_n - E_m}{\hbar}\right), \qquad (4)$$

where Δ is the mean level spacing. In practice, it is convenient, without loss of generality, to work with units such that in (3) the time t is measured in units of length, and we make the replacements $m \mapsto 1$ and $v \mapsto 1$. Then (4) can be cast into the form $\langle|M_{nm}|^2\rangle \approx (\Delta_k/2\pi)\tilde{C}(\kappa)$, where Δ_k is the mean level spacing in k.

Figure 2 shows the excellent agreement between the actual band profile and that predicted by Eq. (4) for generic deformations and dilation. Note that there were no fitted parameters in this match. In all estimations of $\tilde{C}(\omega)$ we have used single trajectories of $\sim 10^6$ consecutive collisions.

Understanding the band profile of $|M_{nm}|^2$ has now been reduced to a matter of finding a classical theory for $\tilde{C}(\omega)$. If we assume that Eq. (3) is a train of uncorrelated impulses, then its power spectrum would be that of white noise, namely, $\tilde{C}(\omega) \approx$ const. A straightforward calculation [8] then leads to the random-wave result (2) already presented. However, in reality there are correlations in this train, and therefore we should expect $\tilde{C}(\omega)$ to have some structure on the frequency scale $\omega \sim 1/\tau_{\mathrm{col}}$. By looking at Fig. 2 we see that the white noise expectation is reasonably satisfied for one of the "generic" deformations (G), but not in the other two cases (D, Gp). We also see nonuniversal peaks at $\omega \sim 1/\tau_{\mathrm{col}} \sim 1$. We now explain that for $\omega \ll 1/\tau_{\mathrm{col}}$ there is total failure of the white noise result for dilations, as well as for translations and rotations, and discuss further complications that may arise if the billiard system is not strongly chaotic.

In Fig. 3 we display $\tilde{C}(\omega)$ for a different billiard shape, a generalized Sinai billiard (Fig. 1), chosen because it does not suffer from the nongeneric marginally stable orbits found in the quarter stadium. Here we see very convincing evidence that for small frequencies we have $\tilde{C}(\omega) \approx$ const

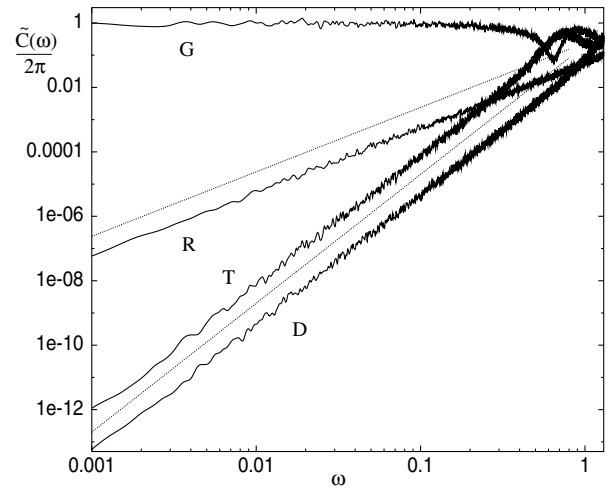


FIG. 3. The classical power spectrum $\tilde{C}(\omega)$ for $\mathcal{F}(t)$ corresponding to a generic deformation (G), dilation (D), translation (T), and rotation (R), in the case of the generalized Sinai billiard with $m = v = 1$. The estimation error is 13% for G and R, and 20% for D and T. The two dotted lines show ω^2 and ω^4 frequency dependence, for purposes of comparison.

for generic deformation, $\tilde{C}(\omega) \propto \omega^4$ for dilation and translation, and $\tilde{C}(\omega) \propto \omega^2$ for rotation. Thus the white noise expectation is indeed satisfied in the $\omega \ll 1/\tau_{\mathrm{col}}$ regime for generic deformations, but fails for dilations, translations, and rotations, for which $\tilde{C}(\omega) \to 0$ as $\omega \to 0$. This property is known (in the context of eigenvalue spectra) as "rigidity" [11]. It implies that the train of impulses is strongly correlated, a result which at first sight seems inconsistent with the assumption of chaotic motion. We will explain that there is no inconsistency here.

The quantity $\mathcal{F}(t) = -\partial\mathcal{H}/\partial x$ is related to $\dot{\mathbf{p}} = -\partial\mathcal{H}/\partial\mathbf{r} = -\nabla V$, the instantaneous force on the particle, by $\mathcal{F}(t) = -\mathbf{D}(\mathbf{r})\cdot\dot{\mathbf{p}}$. For translations we have $\mathbf{D} = \vec{\mathbf{e}}$, where $\vec{\mathbf{e}}$ is a constant vector that defines a direction in space. We can write $\mathcal{F}(t) = (d/dt)^2\mathcal{G}(t)$, where $\mathcal{G}(t) = -m\vec{\mathbf{e}}\cdot\mathbf{r}$. A similar relation holds for dilation $\mathbf{D} = \mathbf{r}$ with $\mathcal{G}(t) = -\frac{1}{2}m\mathbf{r}^2$. It follows that $\tilde{C}(\omega) = \omega^4\tilde{C}_G(\omega)$, where $\tilde{C}_G(\omega)$ is the power spectrum of $\mathcal{G}(t)$. Assuming that $\mathcal{G}(t)$, unlike $\mathcal{F}(t)$, is a generic fluctuating quantity that looks like white noise, it follows that $\tilde{C}(\omega)$ is generically characterized by ω^4 behavior for either translations or dilations. For rotations we have $\mathbf{D} = \vec{\mathbf{e}}\times\mathbf{r}$, and we can write $\mathcal{F}(t) = (d/dt)\mathcal{G}(t)$, where $\mathcal{G}(t) = -\vec{\mathbf{e}}\cdot(\mathbf{r}\times\mathbf{p})$ is a projection of the particle's angular momentum vector. Consequently $\tilde{C}(\omega) = \omega^2\tilde{C}_G(\omega)$, and we expect $\tilde{C}(\omega)$ to be generically characterized by ω^2 behavior in the case of rotations.

In the previous paragraph we assumed that generic fluctuating quantities such as $\mathbf{r}^2$ and $\vec{\mathbf{e}}\cdot\mathbf{r}$ and $\vec{\mathbf{e}}\cdot(\mathbf{r}\times\mathbf{p})$, as well as $\mathcal{F}(t)$ for any generic deformations, have a white noise power spectrum as $\omega \to 0$. Obviously, this "white noise assumption" should be verified for any particular example. If the motion is *not* strongly chaotic, meaning that $C(\tau)$ decays like a power law (say, $1/\tau^{1-\gamma}$ with

$0 < \gamma < 1$) rather than an exponential, then the universal behavior is modified: we may have $\omega^{-\gamma}$ behavior for small frequencies. For a generic system, for instance, the generalized Sinai billiard, we do not have this complication. The stadium example on the other hand is nongeneric: the trajectory can remain in the marginally stable "bouncing-ball" orbit (between the top and bottom edges) for long times, with a probability scaling as a power law in time. Depending on the choice of $\mathbf{D}(\mathbf{r})$, this *may* manifest itself in $C(\tau)$. For example, in Fig. 2 the deformation Gp involves a parallel displacement of the upper edge, and the resulting sensitivity to the bouncing-ball orbit leads to large enhancement of the fluctuation intensity $\tilde{C}(\omega = 0)$, and is suggestive of singular $\omega^{-\gamma}$ behavior for small ω.

Finally, consider the time-dependent problem which is described by the Hamiltonian $\mathcal{H}(\mathbf{r}, \mathbf{p}; x(t))$. It is well known that under quite general circumstances the dissipation is Ohmic ($\propto \dot{x}^2$) (see [8,12], and references therein). If $x(t) = A\sin(\omega t)$, linear response theory gives the long-time heating rate $d\langle\mathcal{H}\rangle/dt = \mu \cdot \frac{1}{2}(\omega A)^2$. The dissipation coefficient μ is determined by the matrix elements of (4) (which up to a factor equals $|M_{nm}|^2$), and therefore μ is proportional to $\tilde{C}(\omega)$. Our results imply that μ vanishes in the limit $\omega \to 0$ for translations. One should not be surprised [4], since this follows from Galilean invariance: One can view the limit $\omega \to 0$ as corresponding to the special case of constant $\dot{x}$. For constant nonzero $\dot{x}$ the particle(s) in the cavity accommodate their motion to the reference frame of the cavity, and there is no dissipation. A similar argument holds for rotations. On the other hand, it is somewhat surprising that the same conclusion holds for dilations (the only other shape-preserving deformation) as well. This observation, as far as we know, has not been introduced previously in the literature.

We thank Eduardo Vergini and Mike Haggerty for important discussions. This work was supported by ITAMP and the National Science Foundation.

Appendix.—There exist a couple of lengthy vector-identity proofs [9,13] of the normalization $M_{nn} = 1$ for the dilation case $\mathbf{D} = \mathbf{r}$, for $d = 2$. Here we present a physically illuminating alternative that works for arbitrary d. We use a phase-space-preserving definition of dilation operator $U(\alpha) \equiv \exp(i\alpha G/\hbar)$. It is generated by the Hermitian operator $G = \frac{1}{2}(\mathbf{r} \cdot \mathbf{p} + \mathbf{p} \cdot \mathbf{r})$. Applying this dilation to wave functions gives the expansion,

$$U(\alpha)\psi(\mathbf{r}) \approx \psi(\mathbf{r}) + \alpha[(d/2)\psi + \mathbf{r} \cdot \nabla\psi] + \mathcal{O}(\alpha^2). \tag{5}$$

The operator also has the effect $U^\dagger \mathbf{r} U = \mathrm{e}^{\alpha}\mathbf{r}$ and $U^\dagger \mathbf{p} U = \mathrm{e}^{-\alpha}\mathbf{p}$. Consider now any Hamiltonian $\mathcal{H}_0 = \mathbf{p}^2/(2m) + V(\mathbf{r})$. By defining the parameter-dependent version $\mathcal{H}(\mathbf{r}, \mathbf{p}; \alpha) = U(\alpha)\mathcal{H}_0(\mathbf{r}, \mathbf{p})U(\alpha)^\dagger$, it is straightforward to obtain

$$\left.\frac{\partial\mathcal{H}}{\partial\alpha}\right|_{\alpha=0} = \frac{\mathbf{p}^2}{m} - \mathbf{r} \cdot \nabla V, \tag{6}$$

whose matrix elements in the case of the billiard potential are $(\partial\mathcal{H}/\partial\alpha)_{nm} = [(\hbar k)^2/m][\delta_{nm} - M_{nm}]$. Thus the nondiagonal terms are the same as those of the deformation $\mathbf{D} = \mathbf{r}$. The diagonal elements can be calculated directly by taking the limit $\alpha \to 0$ of the expression $(\langle U\psi|\mathcal{H}_0|U\psi\rangle - \langle\psi|\mathcal{H}_0|\psi\rangle)/\alpha$. By using (5) and the fact that $\langle\psi|\mathbf{r} \cdot \nabla|\psi\rangle = -d/2$, one can easily show that the result equals zero. From here it follows that $M_{nn} = 1$.

[1] E. Vergini and M. Saraceno, Phys. Rev. E **52**, 2204 (1995).
[2] As is common in the "quantum chaos" literature, $k_n \approx k_m$ implies that $k_n - k_m$ is classically small, but still $n - m$ may be very large. For $n = m$, we define $k \equiv k_n$.
[3] For degenerate $k_n = k_m$, they are shown to be exactly zero in the dilation, translation, and rotation cases [9].
[4] J. Blocki, Y. Boneh, J. R. Nix, J. Randrup, M. Robel, A. J. Sierk, and W. J. Swiatecki, Ann. Phys. (N.Y.) **113**, 330 (1978).
[5] S. E. Koonin, R. L. Hatch, and J. Randrup, Nucl. Phys. **A283**, 87 (1977); S. E. Koonin and J. Randrup, Nucl. Phys. **A289**, 475 (1977).
[6] E. Vergini, Ph.D. thesis, Universidad de Buenos Aires, 1995.
[7] However, the behavior $B \sim |k_n - k_m|$ has since been observed numerically [E. Vergini (personal communication)], in agreement with our results.
[8] D. Cohen, Ann. Phys. (N.Y.) (to be published), cond-mat/9902168.
[9] M. V. Berry and M. Wilkinson, Proc. R. Soc. London A **392**, 15 (1984).
[10] M. Feingold and A. Peres, Phys. Rev. A **34**, 591 (1986); M. Feingold, D. Leitner, and M. Wilkinson, Phys. Rev. Lett. **66**, 986 (1991); M. Wilkinson, M. Feingold, and D. Leitner, J. Phys. A **24**, 175 (1991); M. Feingold, A. Gioletta, F. M. Izrailev, and L. Molinari, Phys. Rev. Lett. **70**, 2936 (1993).
[11] Namely, the sequence of times $\{t_{\mathrm{col}}\}$ is analogous to the sequence of energies $\{\epsilon_n\}$ of chaotic spectra. Rigidity means that $\tilde{C}(\omega)$, analogous to Berry's form factor $K(\tau)$, goes to zero at the origin. See M. V. Berry, Proc. R. Soc. London A **400**, 229 (1985).
[12] D. Cohen, Phys. Rev. Lett. **82**, 4951 (1999); in *Proceedings of the International School of Physics "Enrico Fermi," Course CXLIII,* edited by G. Casati, I. Guarneri, and U. Smilansky (IOS Press, Amsterdam, 2000).
[13] P. A. Boasman, Nonlinearity **7**, 485 (1994).

Substrate-based atom waveguide using guided two-color evanescent light fields

A. H. Barnett,* S. P. Smith, M. Olshanii, K. S. Johnson,† A. W. Adams,‡ and M. Prentiss
Lyman Laboratory, Harvard University, Cambridge, Massachusetts 02138
(Received 13 July 1999; published 18 January 2000)

We propose a dipole-force linear waveguide which confines neutral atoms up to $\lambda/2$ above a microfabricated single-mode dielectric optical guide. The optical guide carries far blue-detuned light in the horizontally-polarized TE mode and far red-detuned light in the vertically-polarized TM mode, with both modes close to optical cut-off. A trapping minimum in the transverse plane is formed above the optical guide due to the differing evanescent decay lengths of the two modes. This design allows manufacture of mechanically stable atom-optical elements on a substrate. We calculate the full vector bound modes for an arbitrary guide shape using two-dimensional non-uniform finite elements in the frequency-domain, allowing us to optimize atom waveguide properties. We find that a rectangular optical guide of 0.8 μm by 0.2 μm carrying 6 mW of total laser power (detuning ± 15 nm about the $D2$ line) gives a trap depth of 200 μK for cesium atoms ($m_F=0$), transverse oscillation frequencies of $f_x=40$ kHz and $f_y=160$ kHz, collection area ~ 1 μm^2 and coherence time of 9 ms. We discuss the effects of non-zero m_F, the $D1$ line, surface interactions, heating rate, the substrate refractive index, and the limits on waveguide bending radius.

PACS number(s): 03.75.Be, 32.80.Pj

There has been much recent progress in the trapping and cooling of neutral atoms, opening up new areas of ultra-low energy and matter-wave physics [1]. Waveguides for such atoms are of great interest for atom optics, atom interferometry, and atom lithography. Multimode atom waveguides act as incoherent atom pipes that could trap atoms, transport them along complicated paths or between different environments, or deliver highly localized atom beams to a surface. Single-mode waveguides (or multimode guides populated only by atoms in the transverse ground-state) could be used for coherent atom optics and interferometry [2,3], as well as a tool for one-dimensional physics such as boson-fermion duality [4–6] and low-dimensional Bose-Einstein condensation effects [7].

The optical dipole-force has long been used to trap and manipulate atoms [1] as well as dielectric particles [8]. The available intensity of lasers has allowed a multitude of such atom traps in the far-detuned regime, giving very low decoherence and heating rates, and storage times on the order of seconds [9].

Evanescent light waves have been popular in many atom mirrors, traps and guides [10,11] since they can provide potentials with high spatial gradients (decay lengths $\sim\lambda/2\pi$ where λ is the optical wavelength), and use rigid dielectric structures (prisms, fibers) to define the potential shape. For example, there has been a series of repulsive (blue-detuned) evanescent-wave (EW) traps which rely on gravity to provide the counteracting force [12] and recent experiments have shown that hollow optical fibers can guide atoms confined within the hollow core using a repulsive evanescent field guided by the fiber [13,14].

The idea of using an EW to provide both attractive and repulsive forces is due to Ovchinnikov *et al.* [15], who proposed the use of two colors (i.e., red and blue detunings) and differing evanescent decay lengths to achieve a trap with the potential minimum a distance $\sim\lambda$ from a prism surface. Until now, this design has been restricted to planar traps (weak confinement in the other two dimensions).

In this paper we discuss a two-color trap based on the EW fields above a single-mode, submicron optical ''channel'' waveguide. The trap provides tight confinement in *two* dimensions and allows free de Broglie wave propagation in the third, forming an atomic waveguide that could transport atoms a between $\lambda/4$ and $\lambda/2$ above the optical guide surface. Our proposal is to utilize the differing vertical evanescent decay lengths of the two *polarizations* carried in the single-mode optical guide (see Fig. 1). The physical origin of this decay length difference is the fact that the TM mode is closer to optical *cut-off* than the TE mode at the same frequency.

Our proposal is reminiscent of some existing resonant enhancement schemes for EW mirrors (demonstrated with surface plasmons [16] and dielectric waveguides [17]) but with a radical change from a planar geometry to a linear, the mechanism for exciting the guide, and the simultaneous guiding of a second frequency of opposite detuning. It also shares the feature of two guided colors with an atom trap proposal using microsphere whispering-gallery modes [18].

Our design has many desirable experimental features: (1) very little optical power is required to obtain large trapping intensities since the optical bound mode has very small cross sectional area (~ 0.3 μm^2), (2) the optical field is *non-divergent*, so can be maintained over distances orders of magnitude further than diffraction-limited propagation in free space allows, (3) the trapping potential is well-known, mechanically stable, and insensitive to experimental parameters other than the optical powers, since it is defined by single-mode intensity distributions fixed relative to a substrate, (4) fabrication of arrays of closely spaced atom waveguides is possible [19], for parallel lithography or mea-

*Electronic address: barnett@tornado.harvard.edu
†Present address: IGEN International Inc., 16020 Industrial Drive, Gaithersburg, MD 20877.
‡Present address: Department of Physics, University of California Berkeley, CA 94720.

 ©2000 The American Physical Society

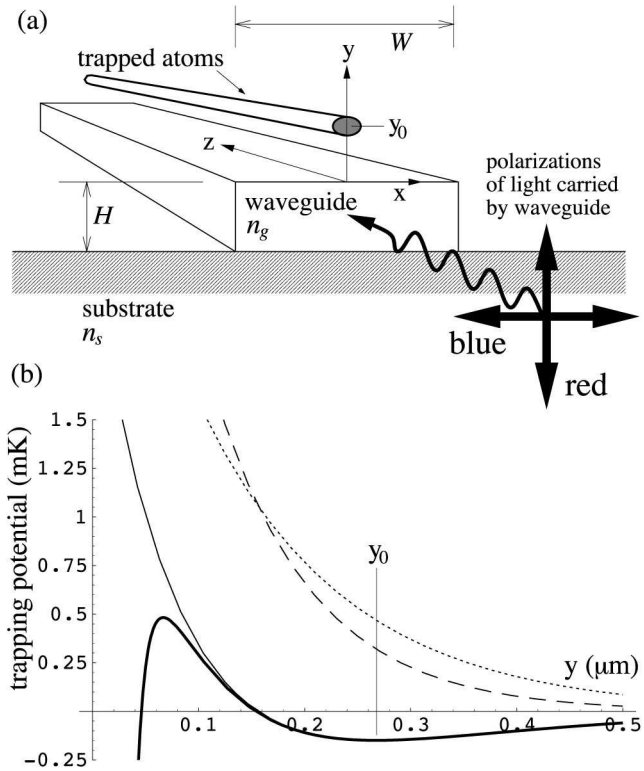


FIG. 1. (a) Shows trap geometry, dielectric guide dimensions, incoming laser polarizations, and the Cartesian axes; (b) shows the trapping potential above the dielectric along a vertical slice at $x=0$. The component due to red-detuned light (absolute value shown as dotted line) subtracts from that of blue-detuned light (dashed line) to give the total dipole potential U_{dip} (thin solid line). This is modified by the Casimir surface interaction (Sec. IV A), giving the final potential (thick line). Here the trap depth of 150 μK and coherence time of 12 ms is generated in our design by 2 mW total guided laser power detuned by ± 15 nm from the cesium $D2$ line.

surement, creating ''on-chip'' integrated atom-optical elements, (5) the atoms are exposed providing additional optical and physical access (a feature not shared with hollow-fiber designs), and (6) the velocity of the atoms along the direction of the waveguide could be controlled by standing waves in the light carried by the optical guide [20].

Compared to a Zeeman-effect magnetic trap for neutral atoms, far-detuned optical dipole-force traps can have comparable trapping times, but typically an order of magnitude less depth and transverse mode spacings than recent magnetic traps [21,22,4]. However, in microfabricated applications the stray magnetic fields decay as a power law with distance, whereas evanescent light fields decay exponentially (ignoring for now any scattering into free space caused by optical defects). We believe this could give guided optical traps a distinct advantage in terms of achievable density of independent atom-optical elements on a single substrate.

Also, optical traps have the advantage that there is no significant loss mechanism which can remove atoms from the trap (assuming the thermal energy is much less than the trap depth): spontaneous events cause a small heating rate, and non-adiabatic changes in m_F can change the optical potential but not the fact that the atom remains trapped. This is to be contrasted with a non-adiabatic spin-flip event in a magnetic trap, which results in loss of the atom. This makes optical waveguides particularly attractive for incoherent transport, when the loss of coherence due to the spontaneous events is unimportant. Finally, optical manipulation has the advantage over magnetic manipulation in terms of high possible switching speeds.

This paper is organized as follows. In Sec. I we describe the dipole potential, the exponential approximation for the EW fields, and the mechanism for the difference in decay length. We show how we optimized the optical guide dimensions, in the case of a rectangular guide on a substrate of unity refractive index (for $m_F=0$), and discuss some design objectives and implementation issues. In Sec. II we give simulated results for cesium atoms: trap depth, coherence time, transverse mode spacing and Q factor, and spontaneous heating rate. We also show how depth and coherence time are generally limited by only two parameters (the detuning and the normalized decay length difference). We study both the case of a substrate refractive index of unity, and in Sec. II C the more realistic index of 1.32. We describe the numerical electromagnetic finite element technique in Sec. III, including the accuracy achieved. Section IV is an investigation of two potential causes of loss or decoherence of atoms, namely interactions with the dielectric surface and bending of the waveguide. Finally in Sec. V we conclude and give some future prospects for this proposal.

I. TRAP CONCEPT

A. Theory of the light potential

An atom in a near-resonant light field of frequency ω experiences both a conservative force (due to stimulated photon exchange) and a dissipative force (due to spontaneous photon emission) [23,24]. The conservative force is the gradient of a spatially-dependent potential $U_{\mathrm{dip}}(\mathbf{r})$ which can be viewed as the time-averaged induced dipole interaction energy (proportional to the real part of the classical polarizability) in the electric field, or equally well as the ''light shift'' (that is, energy level shift due to the ac Stark effect) of the atomic ground state [3,9].

We assume that we apply a monochromatic light field of detuning $\Delta \equiv \omega - \omega_0$ to an alkali atom (with the $ns \rightarrow np$ transition resonant at ω_0), in the *far-detuned* regime (Δ greater than the excited state hyperfine splitting, but much less than ω_0) and the *low saturation* regime ($\Omega_{\mathrm{rabi}} \ll \Delta$, where the Rabi flopping rate Ω_{rabi} is defined [25] by $\hbar\Omega_{\mathrm{rabi}} \equiv \mu E_0$, the dipole matrix element multiplied by the electric field amplitude). The dipole potential has both a scalar and a magnetic part:

$$U_{\mathrm{dip}}(\mathbf{r}) = \beta_s \frac{\hbar\Gamma}{8}\frac{\Gamma}{\Delta} s(\mathbf{r}) + \mu_{\mathrm{Bohr}} g_{\mathrm{nuc}}(L,S,J,i,F) m_F |\mathbf{H}(\mathbf{r})|, \tag{1.1}$$

where Γ is the spontaneous decay rate, $s(\mathbf{r})$ is the spatially-dependent saturation parameter, and the potential is taken to be much less than the ground state hyperfine splitting. Only the magnetic part is affected by m_F, which is defined as the projection of the total angular momentum F on the direction of the local effective magnetic field $\mathbf{H}(\mathbf{r})$. The constant g_{nuc} is the nuclear Landé g-factor appropriate for the F of the ground state. The scalar potential is identical to the case of a two-level atom, apart from the strength factor β_s which is $\frac{1}{3}$ for detuning from the $D1$ line, $\frac{2}{3}$ for the $D2$ [9].

It is important to be precise with the definition of the saturation parameter. We write

$$s(\mathbf{r}) \equiv \frac{E_0^2(\mathbf{r})}{E_{0,\mathrm{sat}}^2}, \tag{1.2}$$

where $E_{0,\mathrm{sat}}^2 \equiv 2I_{\mathrm{sat}}/\epsilon_0 c$ (in the MKSA system) is the squared electric field amplitude in a plane wave of intensity I_{sat} [25]. This avoids the ambiguities present with the usual definition $s(\mathbf{r}) \equiv I(\mathbf{r})/I_{\mathrm{sat}}$ in the case of a general monochromatic light field (composed of an arbitrary coherent sum of travelling and evanescent waves), and emphasizes that it is the local electric field alone that causes the dipole potential.

The effective magnetic field has a strength and direction given by the circularly polarized component of the electric field [9], which can be written thus:

$$\mu_{\mathrm{Bohr}}\mathbf{H}(\mathbf{r}) = \beta_m \frac{\hbar\Gamma}{8}\frac{\Gamma}{\Delta}\frac{\mathrm{Re}[\mathbf{E}_0^*(\mathbf{r})]\times \mathrm{Im}[\mathbf{E}_0(\mathbf{r})]}{E_{0,\mathrm{sat}}^2}, \tag{1.3}$$

where the strength factor β_m is $-\frac{2}{3}$ for $D1$, $\frac{2}{3}$ for $D2$, and we take the physical electric field [with amplitude $E_0(\mathbf{r}) \equiv |\mathbf{E}_0(\mathbf{r})|$] to be the real part of a complex field $\mathbf{E}(\mathbf{r},t) \equiv \mathbf{E}_0(\mathbf{r})\exp(-i\omega t)$. The reason for the ''dummy'' constant μ_{Bohr} is aesthetic, so that (1.1) can be expressed in a standard magnetic form. Note that for nonzero m_F the magnetic and scalar contributions to the potential are of similar order, if the fields have significant circular polarization (this will be true for our trapping fields, for the reason that the optical guide is close to cut-off).

The fact that $U_{\mathrm{dip}}(\mathbf{r})$ has its sign controlled by the sign of the detuning allows both attractive (red-detuned) and repulsive (blue-detuned) potentials to be created. The potential scales as I/Δ but the spontaneous emission rate scales as I/Δ^2; from this follows the well-known result that, if coherence time is an important factor, it is best to be far off-resonance and use high intensities in order to achieve the desired trap depth [9,3].

For simplicity, in this paper we will restrict our further analysis and simulations to $m_F=0$, although our initial calculations suggest that the effect of the magnetic part of our potential when trapping in other m_F states will not pose major problems [assuming the spin axis adiabatically follows the $\mathbf{H}(\mathbf{r})$ field direction], and can even be used to our advantage by increasing the depth and the transverse oscillation frequency ω_x in the case $m_F>0$ [26]. Also, we will consider the effect of only a single resonance (choosing $D2$ because it has a larger β_s than $D1$), which is a valid approximation when the detunings from this resonance are much less than the alkali atom fine structure splitting Δ_{fs}. Even when it becomes advantageous to use a large detuning of the order of Δ_{fs}, it is possible to cancel the effect of the other resonance by a simple shift in the two detunings (as we will see at the end of Sec. II A).

If we now have two light fields of differing frequency, the atomic potentials add [15,23], as long as we assume that the timescale of atomic motion is much slower than the beating period (that is, the inverse of the frequency difference). In our case, atomic motion occurs at $\sim 10^5$ Hz and our light field frequency difference is $\sim 10^{15}$ Hz, so this assumption is valid. Choosing equal but opposite detunings $\pm\Delta$ about the $D2$ line, the trapping potential for $m_F=0$ is

$$U_{\mathrm{dip}}(\mathbf{r}) = \frac{2}{3}\frac{\hbar\Gamma}{8}\frac{\Gamma}{\Delta}[s_{\mathrm{blue}}(\mathbf{r}) - s_{\mathrm{red}}(\mathbf{r})], \tag{1.4}$$

written in terms of the saturation parameters for the two colors. The spontaneous scattering rate [3] is a factor Δ/Γ smaller than $U_{\mathrm{dip}}(\mathbf{r})/\hbar$ (in fact this relation applies for any m_F state), but it depends on the *sum* of the saturation parameters rather than the *difference*. It is also spatially dependent, and has the form

$$\Gamma_{\mathrm{scatt}}(\mathbf{r}) = \frac{2}{3}\frac{\Gamma}{8}\left(\frac{\Gamma}{\Delta}\right)^2[s_{\mathrm{blue}}(\mathbf{r}) + s_{\mathrm{red}}(\mathbf{r})]. \tag{1.5}$$

B. Design of the light fields

Our basic task is to create intense evanescent light fields with a potential minimum sufficiently far from a dielectric surface to make the surface interaction potential and heating mechanisms negligible (discussed in Sec. IV A). The main difficulty arises because the evanescent fields have a typical exponential decay length $\sim\lambda/2\pi$, so if we are to have a trap of useful depth, we are restricted to keep it within roughly λ of the surface (less than a micron).

A potential minimum in one dimension can be obtained using a blue (repulsive) light field of higher intensity at the dielectric surface than the red (attractive) light field, and ensuring the decay lengths obey $L_{\mathrm{red}} > L_{\mathrm{blue}}$, giving a potential of the form

$$U_{\mathrm{dip}}(y) = A_{\mathrm{blue}}e^{-y/L_{\mathrm{blue}}} - A_{\mathrm{red}}e^{-y/L_{\mathrm{red}}}. \tag{1.6}$$

This gives a repulsive force at short range, which becomes attractive at long range [see Fig. 1(b)], and is the scheme for the planar trap of Ovchinnikov *et al.* [15]. A large amount of insight into our proposed trap can be gleaned from this simple one-dimensional model (which we call the *exponential approximation*), because the squared electric fields above the guide will turn out to approximate exponential forms in the vertical direction quite closely.

If we define a normalized decay length difference $\alpha_L \equiv (L_{\mathrm{red}} - L_{\mathrm{blue}})/L_{\mathrm{blue}}$, then we can give two reasons why increasing α_L is a vital design objective. Firstly, it is easy to show that for small α_L the deepest available trap depth (found by optimizing the ratio of surface intensities $A_{\mathrm{red}}/A_{\mathrm{blue}}$) scales as α_L. Secondly, a larger α_L is beneficial

for trap coherence, (giving a smaller spontaneous decay rate at a given trap depth and detuning), because the sum of the intensities can be kept lower [see Eq. (1.5)] for a given intensity difference [Eq. (1.4)]. We will quantify this latter connection in Sec. II A.

Our two key differences from the proposal of Ovchinnikov *et al.* are as follows. Firstly, we create a non-zero α_L by using two orthogonally-polarized bound modes of a dielectric slab guide, which have different evanescent decay lengths at the same frequency [27]. This contrasts with Ovchinnikov *et al.* who suggest varying the decay lengths by varying the reflection angles from the inside surface of a glass prism. Secondly, horizontal confinement is achieved by limiting the width of the slab guide to approximately λ (forming what is called a *channel guide* [28]), which automatically creates a maximum in each light intensity field in the horizontal direction. This results in a tight horizontal confinement in the atomic potential, of similar size to the vertical confinement, and is something very hard to achieve in a prism geometry.

A schematic of our design is shown in Fig. 1(a). The optical guide height H and width W are kept small enough to guarantee that there are exactly two bound modes, differing in polarization but not in nodal structure (in optical terminology this is called *single-mode*): E_{11}^x has an electric field predominantly in the x direction [29], and is to be excited by blue-detuned laser light, and E_{11}^y has electric field predominantly in the y direction and is to be excited by red-detuned laser light. We can see why their vertical decay lengths differ by considering the case of the slab (i.e., taking the width $W\rightarrow\infty$), where these modes are simply the slab TE and TM modes respectively. For both these slab modes the purely transverse field obeys the differential equation

$$\frac{\partial^2\phi}{\partial y^2}=k_0^2[(n_{\rm eff}^{(i)})^2-n(y)^2]\phi, \tag{1.7}$$

where $\phi=E_x(H_x)$ for the TE (TM) mode, the eigenvalue $n_{\rm eff}^{(i)}\equiv k_z^{(i)}/k_0$ is the effective refractive index for the $i^{\rm th}$ mode (k_z being the wavenumber in the propagation direction and k_0 the free space wavenumber), and $n(y)$ is the spatially-dependent refractive index [28]. (This is equivalent to a one-dimensional quantum problem in the direction normal to the slab, in a potential $-k_0^2n(y)^2$ with $\hbar^2/2m=1$). However the boundary conditions on the slab surfaces differ for the two mode types: ϕ is always continuous, but $\partial\phi/\partial y$ is continuous for TE as opposed to $n^{-2}\partial\phi/\partial y$ continuous for TM. This asymmetry exists because the permittivity $\epsilon=n^2$ varies in space but the permeability μ is assumed to be constant. This discontinuity in the gradient for the TM mode forces it to have a lower $n_{\rm eff}$ than that of the TE mode, which means it is less tightly bound so has a longer evanescent decay length. This effect becomes more pronounced as the slab index increases or as optical cut-off is approached (which happens when $n_{\rm eff}$ decreases until it reaches n_s and the mode becomes unbound). This tendency is preserved even as the width is decreased to only a few times the height, as in our scheme.

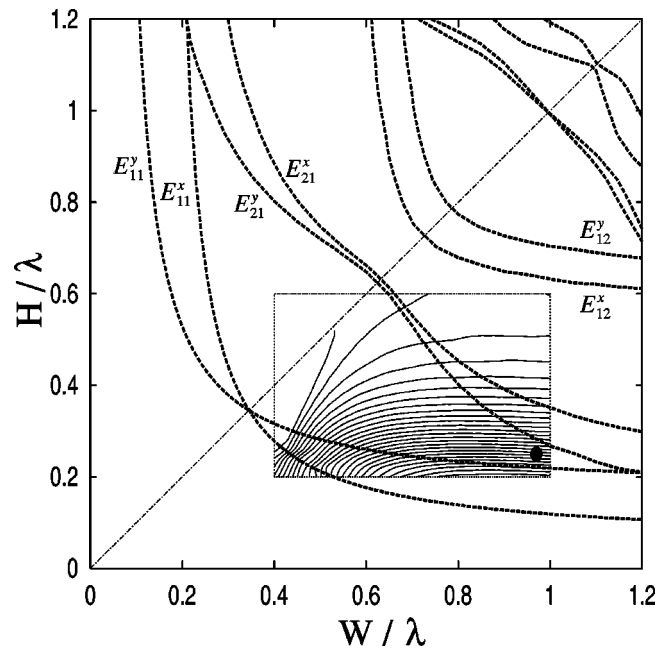


FIG. 2. Numerically solved cut-off curves for a dielectric waveguide of $n_g=1.56$ (with $n_s=1$) as a function of its width and height (dashed curves) and, on the same axes, contours of the maximum trapping potential depth achievable at fixed total laser power (thin solid curves in rectangular overlayed box region). Also shown is the symmetry line $W=H$ (thin dash-dotted line). Note that the contours show depth *increasing* as H decreases, almost independent of W, near the suggested operating dimensions (shown as a solid ellipse). Cut-off is defined as reaching an effective refractive index $n_{\rm eff}\equiv k_z/k_o=1.05$, except for E_{21}^x and E_{21}^y (thick dashed lines) which we show cut-off at $n_{\rm eff}=1.02$.

No analytic solution exists for the general rectangular guide, so we used the finite element method discussed in Sec. III to solve for the bound mode $n_{\rm eff}$ values and fields as a function of guide dimensions. Figure 2 shows the resulting ''cut-off curves,'' that is, contours of constant $n_{\rm eff}$ in the parameter space (W,H). In this example we chose a guide index $n_g=1.56$ (typical for a polymer dielectric) and, as a preliminary case, a substrate index $n_s=1$.

The single-mode region, in which we wish to remain, is bounded below by the E_{11}^y and E_{11}^x curves and above by the E_{21}^y curve. Note that, as in any dielectric guiding structure uniform in the z axis, the lowest two modes (E_{11}^y and E_{11}^x in our case) *never* truly reach cut-off, rather, they approach it exponentially as the guide cross-section is shrunk to zero. For this reason, we chose the practical definition of cut-off for these modes to be $n_{\rm eff}=1.05$, which corresponds to only about 20% of the power being carried inside the guide. In contrast, higher modes do have true cut-offs [28,30] (this distinction is illustrated by the dispersion curves of Fig. 4), and for the E_{21}^y mode our (numerically limited) contour choice of $n_{\rm eff}=1.02$ falls very close to the true cut-off curve.

Using the numerically calculated electric field strengths of the E_{11}^y and E_{11}^x modes, we found the red and blue guided laser powers which gave the deepest trap, subject to the constraint of fixed total power (keeping the detuning constant

[31]). We also imposed the restriction that the zero of trapping potential come no closer than 100 nm along the vertical line $x=0$, which kept the trap minimum a reasonable distance from the surface (see Sec. IV A). Performing this optimization over a region of the parameter space covering the single-mode region gave a contour plot of maximum achievable depth for a given total power, shown within the rectangle overlayed on Fig. 2. This depth increases from negligible values in the top left to the largest depths in the lower right, indicating that choosing W and H to be in this latter corner of the single-mode region is best for depth. The depth shows very little variation with W in this corner, rather it is clear that varying H to stay within our definition of the single-mode region has become the limiting factor on achievable depth. We indicate a practical choice of $W=0.97\lambda$ and $H=0.25\lambda$ as a small marker on Fig. 2. Example trapping potentials shapes possible with these parameters are shown in Fig. 3; we discuss their properties in Sec. II.

In Fig. 1(a) the direct excitation of the optical guide by the two laser beams is shown only schematically. In a realistic experimental setup this coupling into the guide would happen on the order of a centimeter from the atom guiding region, and could involve tapered or Bragg couplers [28] from beams or from other fibers. At this distance we estimate that isotropic stray light due to an insertion loss of 0.5 would have 8 orders of magnitude less intensity than the EW fields in the guiding region. Assuming the light is coherent, this limits the fractional modulation of the guiding potential to 10^{-4}. More improvements are possible, including the use of absorbing shields, bending the guide through large angles away from the original coupling direction, and reducing the coherence length.

C. Discussion of optical cut-off and substrate choice

We have shown in Fig. 2 that parameters optimized for trap depth are near optical cut-off. It is worth gathering together the physical reasons for this. Firstly, on general geometric grounds, the typical available intensities in a guide scale inversely with the effective cross-sectional area of the bound mode, which, far from cut-off (that is, when $n_g-n_{\mathrm{eff}}\ll n_g-n_s$) follows very closely the cross-sectional area of the guide. This makes it favorable to shrink guiding structures to areas less than a square wavelength, where they generally become single-mode. Secondly, once we're in the single-mode regime, as we approach cut-off the mode power is carried increasingly outside the guide, increasing the ratio of surface intensity to guide center intensity. Thirdly, the evanescent decay length in the vacuum is longer as we approach cut-off (recall that in the case of the slab, this is exactly expressed by $L^{-1}=2k_0\sqrt{n_{\mathrm{eff}}^2-1}$ [28], where the factor of 2 arises because we are considering intensity decay length rather than amplitude). In the special case of $n_s=1$, the decay length diverges to infinity as we approach cut-off. Finally, the ratio α_L becomes larger as we approach cut-off, with corresponding beneficial effects on depth and coherence (due to increasing the ''goodness factor'' we will introduce in Sec. II A).

Unfortunately, these purely theoretical reasons for approaching cut-off are in opposition to more practical ones. The closer to cut-off a guide is, the more sensitive it is to manufacturing variations in cross-section: in our case this will be predominantly a sensitivity to H. The result is that small variations in H cause large variations in mode size, or, at worst, complete cut-off. If the mode size change is rapid (nonadiabatic) along the z axis, (for instance if this change is due to surface roughness or refractive index inhomogeneities) the resulting mismatches will be a source of scattering of the guided power. Any coherent scattering back down the guide will set up periodic modulations of the light field over long distances. (One way to reduce the distance over which coherent addition is possible is to use very broad line-width light sources, which would dramatically reduce modulations due to both guided and stray scattered light). The consequence for the atoms would be a z-dependent trap depth and shape, and this could lead to partial reflection or even localization of the matter waves. In general we expect scattering to limit how close to cut-off we can reliably operate.

With regard to the substrate, further practical issues arise. In the above cut-off calculation we chose the simplest case of $n_s=1$, corresponding to a guide surrounded by vacuum. A real substrate with $n_s>1$ has the unfortunate effect of limiting the propagation constant k_z of strictly bound modes to be larger than the freely propagating wavevector in the substrate; in other words, $n_{\mathrm{eff}}>n_s$ must hold or the light field will rapidly tunnel into the ''attractive potential'' of the substrate. This in turn limits the decay lengths and α_L that can be achieved. The trap properties quoted in the abstract and in Secs. II A and II B paper rely on very low n_{eff} values (1.07 for the E_{11}^y mode, 1.18 for E_{11}^x) for the reason that a low n_{eff} is the only way to create long evanescent decay lengths in the vacuum. (This is equivalent to Ovchinnikov *et al.* choosing reflection angles very close to critical [15].) However, in Sec. II C we present preliminary results for a substrate of sodium fluoride (the lowest-index common optical mineral, at $n=1.32$), and do not believe the substrate alters the basic feasibility of our waveguide.

For completeness, here we list some other possible approaches to the substrate issue. (1) Use an aerogel substrate, which can have exceptionally low refractive indices and low loss (films of several μm thickness with indices of about 1.1 can be produced [32]). (2) Use a dielectric multilayer substrate with an effective index of unity or less (very low loss multilayer mirrors [33] with effective indices less than unity can be created). (3) Investigate if there exist guide shapes which have sufficiently small tunnelling rate into a conventional substrate that the fact that the modes are not strictly bound becomes irrelevant (for instance, a wedge shape with the smallest face in contact with the substrate). (4) Unsupported guiding structures could be produced over short distances [19]. Finally, it is important to note that the idea of replacing the substrate by a metallic reflective layer is not practical because they are too lossy.

Ultimately, the best values of W and H, the best guide cross-sectional shape, and the substrate choice will depend on many of the above factors and is an area for further research.

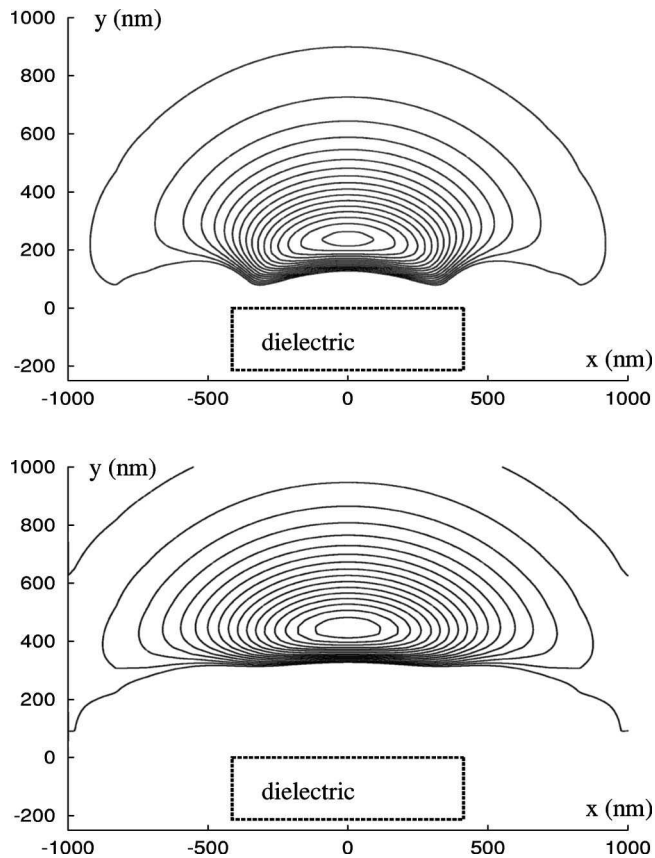


FIG. 3. Trapping potential shapes in the xy plane, with guide dimensions of 0.83 μm by 0.21 μm and an index of 1.56. In each case the maximum depth is 100 μK and the coherence time of atoms in the ground state is 19 ms, achieved with detuning ± 15 nm from the cesium $D2$ line. The outer contour shows a depth of 3 μK, the cesium MOT temperature. Subsequent contours are spaced by 6 μK. The plots illustrate the range of trapping distances acheivable: the upper trap ($p=0.4$, using a total guided power of 1 mW) has a minimum 0.24 μm from the surface; the lower trap ($p=0.2$, total guided power of 7.5 mW) has a minimum 0.44 μm from the surface.

II. TRAP PROPERTIES

In the bulk of this section we will examine the atomic waveguide properties for light nearly resonant with the $D2$ line of cesium, using an optical guide of index 1.56 of the dimensions $W=0.97\lambda$ and $H=0.25\lambda$ from Sec. I B, and a substrate of unity index. The saturation intensity for cesium is 11.2 W/m^2 [3], and its resonant wavelength of 852 nm requires that the physical guide size is 0.83 μm by 0.21 μm. (At the end of the section we present preliminary calculations for a $n_s=1.32$ substrate and a different guide, and discuss how the atom waveguide properties are changed).

Given the guide, we are free to choose three experimental parameters, namely the optical powers carried in the two modes, and the detuning Δ (assumed to be symmetric, that is, to be of equal magnitude for red and blue beams, because little advantage can be gained with an unsymmetric detuning). The first two of these can usefully be reexpressed as total power $P_{\rm tot}\equiv P_{\rm red}+P_{\rm blue}$, and the power ratio $p\equiv P_{\rm red}/P_{\rm blue}$. The trap shape will be affected by p alone: we show the trapping potential shapes achievable at the two practical extremes of $p=0.4$ and $p=0.2$ in Fig. 3, where we have chosen $P_{\rm tot}$ and Δ to give identical trap depths and coherence times. Smaller p values cause the trap minimum to move further from the surface (a distinct advantage), to be less ''bean'' shaped (i.e., to have smaller cubic deviations from a 2D harmonic oscillator), and to cause a slight increase in collection area. It is possible to achieve a trap minimum as distant as $y_0=0.52\lambda$ from the surface when $p=0.2$. The only disadvantage to implementing these smaller p values is that a higher $P_{\rm tot}$ is required to achieve the same trap depth and coherence time (for instance a factor of 7.5 increase is required as we take p from 0.4 to 0.2). This can be quantified within the exponential approximation, and it can be found that the total power required to maintain a given depth and coherence time with a fixed trap geometry scales as $P_{\rm tot}\sim(1+p)/p^{1+1/\alpha_L}$.

If we were purely interested in maximizing trap depth at a given $P_{\rm tot}$ and detuning, it would be best to make p as large as possible, however if we take p much larger than 0.4 the trap is brought so close that the corners of the ''bean'' shape touch the dielectric surface (see Fig. 3, upper plot) and we will lose effective collection area due to sticking of atoms onto this surface.

A. Depth, coherence time, and Q factor

We may ask what trade-offs are necessary between trap depth and coherence time. It turns out that, within the exponential approximation (1.6), this is elegantly quantifiable. We can define a ''goodness factor''

$$G\equiv\frac{s_{\rm blue}(\mathbf{r}_0)-s_{\rm red}(\mathbf{r}_0)}{s_{\rm blue}(\mathbf{r}_0)+s_{\rm red}(\mathbf{r}_0)}=\frac{\Gamma}{\hbar|\Delta|}U_{\rm max}\tau_{\rm coh}, \tag{2.1}$$

where the trap minimum position $\mathbf{r}_0$ is at $(x=0,y=y_0)$, and the second equality is verified by substitution of (1.4) and (1.5), and defining $U_{\rm max}\equiv|U_{\rm dip}(\mathbf{r}_0)|$ and $\tau_{\rm coh}\equiv\Gamma_{\rm scatt}^{-1}(\mathbf{r}_0)$. We use this latter definition because we are interested in the coherence time of atoms spending time close to the trap minimum (which will certainly be true for the transverse ground state). Using (1.6) to solve for y_0 and evaluate the ''goodness factor,'' it turns out that the factor is independent of either laser power (i.e., of either $A_{\rm red}$ or $A_{\rm blue}$), giving

$$G=\frac{L_{\rm red}-L_{\rm blue}}{L_{\rm red}+L_{\rm blue}}=\frac{\alpha_L}{2+\alpha_L}. \tag{2.2}$$

Combining (2.1) and (2.2) gives

$$U_{\rm max}\tau_{\rm coh}=\frac{\alpha_L}{2+\alpha_L}\frac{\hbar|\Delta|}{\Gamma}, \tag{2.3}$$

fixing the product of achievable depth and coherence time as a constant multiple of the detuning. This is a remarkable result since it shows that increasing α_L is really *the* only objective in the field design of two-color EW traps. We can write this in units more convenient for cesium trap design, thus,

$$\left(\frac{U_{\max}}{1\ \mu\mathrm{K}}\right)\left(\frac{\tau_{\mathrm{coh}}}{1\ \mathrm{ms}}\right)=(644.2)\frac{\alpha_L}{2+\alpha_L}\left(\frac{|\Delta|}{1\ \mathrm{nm}}\right)=(122)\left(\frac{|\Delta|}{1\ \mathrm{nm}}\right), \tag{2.4}$$

where the value $\alpha_L=0.47\pm0.02$ (taken from best-fit exponentials to the numerically-found squared electric fields for the guide dimensions of Sec. I B with $n_s=1$) has been substituted to give the the final form. This design expression does not give the P_{tot} required to reach a desired balance between $U_{\max}$ and τ_{coh}, however, the total laser power is usually in the mW range, several orders of magnitude less than in most free-space trap designs. For instance, with $P_{\mathrm{tot}}=20$ mW, $p=0.4$, and $\Delta=\pm15$ nm we could generate a trap of 2 mK depth with the relatively short coherence time of 0.9 ms. The transverse oscillation frequencies in this trap would be $\omega_x/2\pi=116$ kHz and $\omega_y/2\pi=490$ kHz (the field shapes fix this ratio at about 1:4), giving an atomic mode spacing due to the x motion of 5.6 μK, roughly twice the cesium MOT temperature, and a characteristic ground-state size of 26 nm by 12 nm.

For coherent guiding, we can define a more physically meaningful figure of merit, $Q\equiv\omega_\perp\tau_{\mathrm{coh}}$, which tells us the typical number of coherent transverse oscillations we can expect multiplied by 2π (i.e., it is the Q factor of the transverse oscillations). We should choose $\omega_\perp=\omega_x$ since this is the smaller of the transverse frequencies in our case. For $Q\gg1$ the transverse atomic modes will be well resolved, and our guide can be a useful interferometric device. Using (2.3), in conjunction with the fact that when the trapping potential shape is fixed then ω_x is proportional to the square root of the depth, tells us that for a given trap and detuning, $Q\propto1/\omega_x$. For a higher Q we should choose smaller transverse oscillation frequencies, that is, shallower traps. For example, the 2 mK trap discussed above has $Q\approx650$, but if we reduce it to a 20 μK trap of the same Δ (by changing p or the laser powers), the Q is 10 times larger. Increasing Δ would allow even higher Q to be realized.

The dependence on detuning in (2.4) is another way of expressing the advantages already known about using far off-resonant beams [3,9]. However, our single-resonance approximation will break down if the detuning is too large: we have (somewhat arbitrarily) chosen a detuning limit of 15 nm, as compared to $\Delta_{\mathrm{fs}}=43$ nm for cesium. At this limit, the additional dipole potential created due to the detunings from the $D1$ line is very significant. However, by removing the detuning symmetry (changing Δ_+ from $+15$ nm to $+12.07$ nm and Δ_- from -15 nm to -17.14 nm), the desired $D2$ single-resonance approximation potential is recovered in the true physical situation of both resonances present. (These required shifts, which are of order $\Delta^2/\Delta_{\mathrm{fs}}$, can easily be found using the expression for the sum of dipole potentials from the two lines.) An additional necessity for our limit is the fact that any larger detunings start to demand separate bound-mode calculations for the two colors, a treatment we reserve for the future. This detuning limit in turn limits the depths, coherence times and Q-factors we quote here, but we anticipate similar future EW atom waveguide designs which explore the region $|\Delta|>\Delta_{\mathrm{fs}}$ (or even $|\Delta|\sim\omega_0$), and achieve much better coherence.

B. Other properties

We estimate the collection area of the trap as the cross-sectional region within which the potential is deeper than the typical cesium MOT energy k_BT_{MOT}. For our example 100 μK traps of Fig. 3 this area is about 1 $\mu\mathrm{m}^2$. However, it is not possible to do much better than this with our design: if one tries to increase the area by increasing the overall trap depth, the k_BT_{MOT} contour touches the substrate, indicating that atoms at this energy can reach the substrate surface, where they will stick, limiting the effective collection area.

To investigate the lifetime of atoms transported incoherently (the multi-mode regime), we can calculate the heating rate along similar lines as Grimm *et al.* [9]. We start with their Eq. (23) which gives the rate of change of the average of the total energy of atomic motion $E=E_{\mathrm{kin}}+E_{\mathrm{pot}}$ as

$$\dot{\overline{E}}=k_BT_R\overline{\Gamma}_{\mathrm{scatt}}, \tag{2.5}$$

T_R being the recoil temperature, and use the assumption that in an equilibrated 3D trap $\overline{E}_{\mathrm{kin}}=\frac{3}{2}k_BT$. Since there is harmonic motion in two directions but free motion in the third, the virial theorem gives us $\overline{E}_{\mathrm{pot}}=\frac{2}{3}\overline{E}_{\mathrm{kin}}$. Combining this with $\overline{\Gamma}_{\mathrm{scatt}}=U_{\max}\Gamma/G\hbar\Delta$ from (2.1) gives the heating rate

$$\dot{T}=\frac{2}{5G}\frac{\Gamma}{\Delta}T_R\frac{U_{\max}}{\hbar}, \tag{2.6}$$

which is of the order of one recoil temperature per coherence time. For our 100 μK depth trap at $\Delta=15$ nm the rate is 4.4 $\mu\mathrm{K\ s}^{-1}$, implying that storage and transport for many seconds is possible. For simplicity, we have ignored the fact that there may be distinct longitudinal and transverse temperatures which do not equilibrate over the trapping timescales.

C. Effect of a realistic substrate

In this section we present calculations, performed using the method of Sec. III, for a practical substrate choice of sodium fluoride (the lowest refractive index common mineral, with $n_s=1.32$ at a wavelength of 852 nm), and investigate how this changes the atom waveguide properties from those presented above. We increased n_g to 1.7 (dense flint glass, e.g., BaSF type) in order to provide sufficient index difference from the substrate.

Fixing the width at $W=1.00\lambda$, we found that a height $H=0.34\lambda$ gave the largest α_L of 0.20 ± 0.01, and allowed both modes to be sufficiently far from cut-off (greater than half the power being carried inside the guide for both modes). The result is a goodness factor G which is approximately half that of the $n_s=1$ case, with a corresponding halving of the achievable product of depth and coherence time according to (2.4), and doubling of the heating rate at a given $U_{\max}$ and Δ according to (2.6). The shorter decay lengths of 56 nm and 68 nm (compared to 93 nm and 137 nm for $n_s=1$) cause the typical trapping distance y_0 to be reduced by a factor of roughly 1.8.

We found that in order to reproduce the depth of 100 μK and $y_0 = 0.24$ μm of the upper trap of Fig. 3 (with Δ unchanged) we needed $P_{\rm tot} = 22$ mW, giving $\tau_{\rm coh} = 9$ ms. The large power increase over the 1 mW required for $n_s = 1$ is explained by the fact that this y_0 is now towards the upper limit practically achievable rather than the lower. (If y_0 is instead scaled in proportion to the new decay lengths, the required increase in $P_{\rm tot}$ is only a factor 1.7). In this example, we find the transverse oscillation frequencies have increased to $\omega_x/2\pi = 81$ kHz and $\omega_y/2\pi = 202$ kHz, compared to the original $\omega_x/2\pi = 26$ kHz, and $\omega_y/2\pi = 109$ kHz. The increase in ω_y is explained entirely by the shorter decay lengths, and the increase in ω_x (by a factor of over three) is attributed to tighter optical mode shapes. It is clear that this latter effect outweighs the decrease in $\tau_{\rm coh}$, implying that the inclusion of the substrate has actually *increased* Q by 50%.

In summary, the effects of including a realistic substrate limit the maximum trapping distance y_0 that can be achieved (because of the reduction in decay lengths), lower the goodness factor, increase the heating rate and the required optical power, but also increase the oscillation frequencies. For our substrate choice, each of these changes was approximately a factor of two, and we believe that they do not alter the basic practicality of implementing our proposed waveguide.

III. NUMERICAL SOLUTION OF THE LIGHT FIELDS

The detailed electric field distribution is very important in calculating the trapping potential above the waveguide. An approximation to the form of the fields in the y direction is given by the analytically-known solution for the slab waveguide, but to get more accuracy and knowledge of the full potential shape in the xy plane, we used a full-vector finite element calculation.

The technique represents the electric and magnetic fields as simple piecewise functions over many ''elements'' (regions) subdividing a slice through the guide and surrounding media in the xy plane, therefore by a finite number of degrees of freedom. Each element has a dielectric constant associated with it, allowing arbitrary stepwise refractive index distributions in the xy plane to be modelled. Maxwell's equations for propagating solutions of the form $\exp(ik_z z - i\omega t)$ are reduced to a generalized sparse eigenvalue equation with k_z^2 as the eigenvalue and the bound mode field distributions as the eigenvectors [34]. Specifically, we used the technique of Fernandez and Lu [35], with H_x and H_y as the field degrees of freedom, for simplicity using first-order (bilinear) functions to represent these fields over a non-uniform but separable rectangular grid of elements. This required a generalization of the Fernandez and Lu implementation, and careful consideration of their line-integral terms (which are non-standard for a finite element formulation) [26]. Of the many available finite element approaches to dielectric waveguide mode solving, this frequency-domain method was chosen for its absence of ''spurious modes,'' its ability to handle index step discontinuities, its small number of required degrees of freedom and its matrix sparsity [36].

Rather than emulating a radiative boundary condition (a notoriously hard task usually requiring an iterative procedure due to the k_z dependence), we enclosed the problem in a large, perfectly-conducting box of sufficient size that the bound mode evanescent fields were negligible on its walls, making the nature of the boundary condition irrelevant. However, the average level spacing of the unbound modes (the ''continuum'') decreases with increasing box size, and especially near cut-off this increases the number of iterations required to solve the eigenvalue problem to a given accuracy (we used the well-known ARPACK solver to find the 11 lowest eigenmodes of the sparse matrix). We found that a box size of 6λ to 7λ gave the best compromise between accuracy and speed. Our non-uniform elements allowed us to have a high element density across the waveguide and in the trapping region, but a low density over the much larger box area, keeping the total number of degrees of freedom manageable.

The fractional error ϵ in the propagation constant k_z was less than 1%, and the accuracy of the electric field strengths in the trapping region $\approx 3\%$, when we used $N \sim 2000$ elements. This was sufficiently accurate for the present work. Finding the bound modes of each waveguide parameter choice typically took between 3 and 20 minutes of computing time on a Silicon Graphics R8000 processor, depending on how close to cut-off the guide was. We tested the accuracy of the method by solving a cylindrical guide in an identical fashion with the same N and a very similar non-uniform grid, for which there are known field solutions [30]. Figure 4 shows the propagation constant agrees with the analytics to within 1%, even close to cut-off. The convergence with N was measured for the rectangular guide case, and found to be $\epsilon \sim N^{-\gamma}$ with $0.55 < \gamma < 0.7$. This is less than optimal for first-order elements (which have a maximum possible convergence of $\gamma = 1$), and is believed to be due to an inability of the bilinear functions to represent physical in-plane E and H components at dielectric steps, or the weak field singulari-

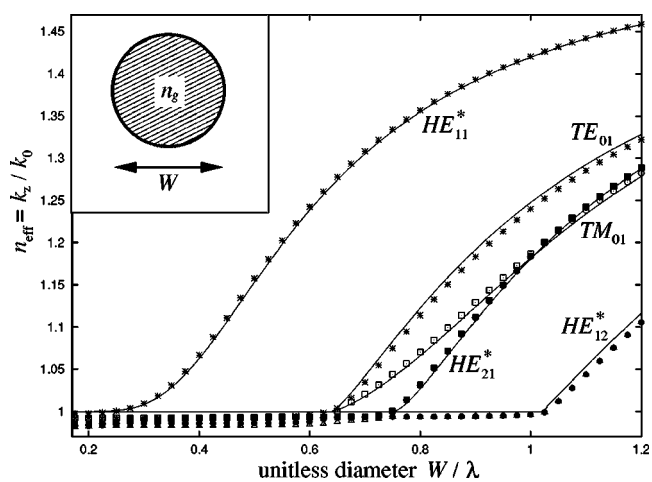


FIG. 4. Comparison of our optical guide bound-mode numerical implementation against known analytic solutions, in the case of a free-standing dielectric cylinder of $n_g = 1.56$. We used discretization and box-size identical to the rectangular guide case, and observe typical errors of $\pm 0.5\%$ in propagation constant for the first two modes. The mode naming convention and analytic calculation follow Snitzer [30]; an asterisk indicates a doubly degenerate mode.

ties which can physically occur at any exterior dielectric sharp edges (regardless of whether acute or obtuse) [26].

Future improvements to the method, which would increase the accuracy or the convergence rate γ, include using higher order elements (if done carefully, this could correctly represent physical E and H components at dielectric steps), and explicit modelling of the field singularities at guide corners using specialized elements.

IV. FURTHER DECOHERENCE AND LOSS MECHANISMS

A. Effects of surface interactions

The EW trap has the benefit of creating high field gradients near a surface, but along with this comes the disadvantage that interactions with that surface that can alter the trap potential and even cause heating and loss of trapped atoms.

An atom's change in potential near a surface is known as the van der Waals interaction ($l \ll \lambda$) or the Casimir interaction ($l \gg \lambda$), depending on the distance l from the surface compared to λ, the dominant wavelength responsible for the polarizability of the atom (in our case of Cs this is the D line resonance, the same as our trapping resonance). There is a smooth cross-over from van der Waals ($U \sim l^{-3}$, which can be viewed as the atom's electrostatic interaction with the image of its own fluctuating dipole) to Casimir ($U \sim l^{-4}$, which can be viewed as a retarded van der Waals attraction or equally well as an atomic level shift due to a cavity QED effect) at $l \approx \lambda/10$ [37]. In the case of a perfect mirror surface, the full form is known for any l, but for a dielectric surface, the expression becomes much more complicated to evaluate [38].

Since our trapping distances are larger than this cross-over point, we will use the Casimir form, which is correct for asymptotically large l, and is always an overestimate of the true potential [38]. The dependence of the coefficient with dielectric constant is complicated [39,40], but we will use the simpler approximate form given by Spruch and Tikochinsky [38], to give

$$U_{\mathrm{Cas}}(l) = -\frac{3}{8\pi}\frac{\hbar c \alpha(0)}{4\pi\epsilon_0 l^4}\frac{\epsilon - 1}{\epsilon + (30/23)\epsilon^{1/2} + 7/23} \tag{4.1}$$

(in the MKSA system). This approximate form is known to be within 6% of the exact expression for any dielectric constant ϵ [40]. Substituting the recently calculated [41] static polarizability of cesium, $\alpha(0) = 399.9$ a.u., gives a Casimir interaction coefficient of 4.9 nK μm^4 for $n_g = 1.56$.

Figure 1(b) shows the effect of this potential on a typical trap of depth 150 μK and distance $y_0 = 270$ nm. It is clear that the change is negligible further than 100 nm from the surface, and a WKB tunneling calculation along this straight-line path (at $x = 0$) shows that even if all atoms that reach the surface stick, the loss rate from the first few transverse modes is entirely negligible. However, care should be taken with the multi-mode regime, or in the case of high-p traps, since the tunneling via the corners of the ''bean'' shape may dominate for $p > 0.4$ (Fig. 3).

The issue of energy transfer to trapped atoms due to a finite (and possibly room) temperature nearby surface is far less well understood, and may be a problem with many surface-based particle traps, as discussed by Henkel and Wilkens [42]. However, since we are trapping neutral particles and the conductivity of our surface is low, we expect a decoherence rate negligible compared to that already present from spontaneous absorption and emission cycles.

B. Bending the waveguide

It would be very useful to be able to carry atom beams along curved paths, by bending our atom waveguide in the plane of the substrate, without significant atom loss. Here we briefly estimate three limitations on the waveguide minimum bending radius (in decreasing order of leniency): (1) the limit imposed by optical radiation leakage, (2) the limit needed for incoherent atom transport, and (3) the limit needed for coherent atom transport in the transverse groundstate. This will give us an idea of the practicality of curved atomic guides.

Firstly, whenever a dielectric optical guide has curvature, there is a loss rate (exponentially small in the curvature radius R), which can be viewed as tunneling out of the guide's ''potential well'' induced by the addition of an effective centrifugal potential. In the limit $R \gg W$, the effective potential is linear with x (the radial coordinate), and the fractional loss per radian of curvature can be estimated [43], for instance using the one-dimensional WKB formula, to be

$$\alpha = C\frac{R}{\lambda}\exp\left(-\frac{1}{6\pi^2}\frac{\lambda^2 R}{L_x^3}\right), \tag{4.2}$$

where L_x is the evanescent decay length in the radial direction, and C is a constant of order unity. Therefore for negligible light loss at a $\pi/2$ bend we need $R > 60\pi^2 L_x^3/\lambda^2$, typically a couple of tens of microns. This is so small chiefly because we are using an optical guide with a large refractive index step [28].

Secondly, we consider atom loss from an incoherent beam with a transverse temperature $k_B T_\perp$ (assumed small compared to the trap depth magnitude $U_{\max}$), and a longitudinal kinetic energy $E_\parallel$. We call the approximate spatial extent of the trap potential in the x direction 2ξ, and restrict ourselves to one-dimensional classical motion in this direction. When in a region of radius of curvature $R \gg \xi$, an effective centrifugal term adds to the trapping potential giving $U(x) = U_{\mathrm{dip}}(x) - 2E_\parallel(x/R)$. This causes the atoms to ''slosh'' towards positive x, only ever returning if there exists a point where $U(x) > -U_{\max}$ for $x > 0$. We can estimate that this will happen if $U_{\max} > 2E_\parallel(\xi/R)$, giving our lower limit on R as $2\xi E_\parallel/U_{\max}$. In our design $\xi \approx 0.5$ μm, so if we choose $R = 1$ mm we can expect loss-free transport of a beam at a longitudinal kinetic energy up to 10^3 times the trap depth.

Thirdly, to model coherent matter-wave propagation along a curved guide, we consider the amplitude for remaining in the transverse ground-state, having passed into a curved section and back into a straight section. If again we assume one-dimensional x motion, and assume a harmonic potential $U(x) = \frac{1}{2}M\omega_x^2(x - x_0)^2$ around the trap minimum,

then the effect of curvature is to shift the minimum position from $x_0=0$ to $x_0=2E_\parallel/M\omega_x^2R$. If this shift is much less than the characteristic ground-state size $(\hbar/M\omega_x)^{1/2}$ then the projection at each transition will be high, resulting in high flux transmission coefficient. This gives $R \gg 2E_\parallel/(\hbar M\omega_x^3)^{1/2}$ as our condition, which for $E_\parallel/k_B = 10$ mK (that is, $v_\parallel=1.1$ ms^{-1}) and $\omega_x/2\pi=40$ kHz corresponds to $R\gg 0.45$ mm. This limit is very conservative since we have not yet made use of the adiabatic condition $\Omega \ll \omega_x$ (where $\Omega\equiv v_\parallel/R$ is the rate of change of direction of the guided atom), to design a waveguide path without discontinuities in the curvature.

In conclusion, we have shown that it is possible to bend atoms both incoherently and coherently through large angles on a compact substrate structure of a few millimeters in size.

V. CONCLUSION

We have presented a novel substrate-based neutral atom waveguide, based on the optical dipole force, which combines the features of a planar far-detuned two-color evanescent trap [15] with the ability to confine strongly along two axes. We utilized the differing vertical decay lengths of the two bound-mode *polarizations* of a submicron-sized optical waveguide near cut-off. We have shown that only a few milliwatts of guided laser power can create atomic potential depths ~ 100 μK with transverse oscillation frequencies ~ 100 kHz, a coherence time ~ 10 ms and a trap minimum 200–400 nm above the optical guide surface, for Cs atoms in the $m_F=0$ state. Laser powers greater than ten milliwatts can give a transverse mode spacing greater than the temperature of a Cs MOT, opening up the single-mode waveguide regime. The advantages of guiding optical trapping fields on a substrate include mechanical stability and reliability, mass production and the potential for transport along complicated paths.

We have given some design criteria for guided-lightwave two-color atom waveguides (chiefly the maximization of the evanescent decay lengths, and of their normalized difference α_L), and shown that a substrate of low refractive index can be very beneficial. We modeled in detail the trapping potentials for a general rectangular guide of index 1.56 above a unity-index substrate, and have shown that a realistic substrate choice of index 1.32 poses few problems to the viability of the device. We predict that the effect of the surface interaction is generally small, and that coherent guiding is possible around corners of radii ~ 1 mm for a longitudinal velocity ~ 1 ms^{-1}. We also believe that the magnetic part of the potential felt by nonzero m_F atoms could be used to increase the depth and oscillation frequencies further.

This preliminary work [specifically Eq. (2.3)] indicates that utilizing detunings much larger than the 15 nm we limit ouselves to here will be very advantageous for coherent guiding. We have only scratched the surface of the design variations possible; for instance, equalizing the horizontal and vertical oscillation frequencies is yet to be attempted. The use of two polarizations is our solution to the problem of maximizing α_L when the detuning is very small compared to the wavelength, but we suspect that there will exist other fruitful schemes where these are comparable ($\Delta\sim\omega_0$, very far-detuning) and where a large α_L is caused simply by the different optical cut-off conditions at the two wavelengths. We have reserved investigation of cooling schemes for future work (although this has already been demonstrated in an EW mirror [44] and proposed in EW traps [45]). We believe that the potential shapes capable of being produced by guided waves on a substrate also include the possibility of funnel-type loading schemes and coherent atom couplers, allowing for a complete ''integrated'' atom-optical experiment on a substrate.

ACKNOWLEDGMENTS

It our pleasure to thank J. Thywissen, J. Babb, N. Dekker, V. Lorent, and the members of the Heller Group for fruitful discussions. A.H.B. would also like to thank Professor E. J. Heller for computing resources. This work was supported by the National Science Foundation, Grant Nos. CHE-9610501 and PHY-9732449.

[1] S. Chu, Rev. Mod. Phys. **70**, 685 (1998); C. Cohen-Tannoudji, *ibid.* **70**, 707 (1998); W. D. Phillips, *ibid.* **70**, 721 (1998).
[2] M. A. Olshanii, Laser Phys. **4**, 995 (1994).
[3] C. S. Adams, M. Sigel, and J. Mlynek, Phys. Rep. **240**, 143 (1994).
[4] J. H. Thywissen, M. Olshanii, G. Zabow, M. Drndić, K. S. Johnson, R. M. Westervelt, and M. Prentiss Eur. Phys. J. D **7**, 361 (1999).
[5] M. Olshanii, Phys. Rev. Lett. **81**, 938 (1998).
[6] T. Cheon and T. Shigehara, Phys. Rev. Lett. **82**, 2536 (1999).
[7] W. Ketterle and N. J. van Druten, Phys. Rev. A **54**, 656 (1996).
[8] A. Ashkin, Phys. Rev. Lett. **24**, 156 (1970); S. Kawata and T. Sugiura, Opt. Lett. **17**, 772 (1992); for a review see A. Ashkin, Proc. Natl. Acad. Sci. USA **94**, 4853 (1997).
[9] For a review of dipole-force atom trapping, see: R. Grimm, M. Weidemüller, and Yu. B. Ovchinnikov, Adv. At. Mol. Opt. Phys. **42**, 95 (2000).
[10] For a review of evanescent-wave atom mirrors, see J. P. Dowling and J. Gea-Banacloche, Adv. At. Mol. Opt. Phys. **37**, 1 (1996).
[11] R. J. Cook and R. K. Hill, Opt. Commun. **43**, 258 (1982); V. I. Balykin, V. S. Letokhov, Yu. B. Ovchinnikov, and A. I. Sidorov, Phys. Rev. Lett. **60**, 2137 (1988).
[12] C. G. Aminoff, A. M. Stearne, P. Bouyer, P. Desbiolles, J. Dalibard, and C. Cohen-Tannoudji, Phys. Rev. Lett. **71**, 3083 (1993); Yu. B. Ovchinnikov, I. Manek, and R. Grimm, *ibid.* **79**, 2225 (1997).
[13] M. J. Renn, D. Montgomery, O. Vdovin, D. Z. Anderson, C. E. Wieman, and E. A. Cornell, Phys. Rev. Lett. **75**, 3253 (1995).
[14] H. Ito, K. Sakaki, M. Ohtsu, and W. Jhe, Appl. Phys. Lett. **70**, 2496 (1997); H. Ito, T. Nakata, K. Sakaki, and M. Ohtsu, *ibid.* **76**, 4500 (1996).

[15] Yu. B. Ovchinnikov, S. V. Shul'ga, and V. I. Balykin, J. Phys. B **24**, 3173 (1991).

[16] T. Esslinger, M. Weidemüller, A. Hemmerich, and T. W. Hänsch, Opt. Lett. **18**, 450 (1993).

[17] R. Kaiser, Y. Levy, N. Vansteenkiste, A. Aspect, W. Seifert, D. Leipold, and J. Mlynek, Opt. Commun. **104**, 234 (1994); G. Labeyrie, A. Landgragin, J. Von Zanthiert, R. Kaiser, N. Vansteenkiste, C. Westbrook, and A. Aspect, Quantum Semiclassic. Opt. **8**, 603 (1996).

[18] H. Mabuchi and J. H. Kimble, Opt. Lett. **19**, 749 (1994).

[19] T. A. Savas, S. N. Shah, M. L. Schattenburg, J. M. Carter, and H. I. Smith, J. Vac. Sci. Technol. B **13**, 2732 (1995).

[20] S. Kawata and T. Tani, Opt. Lett. **21**, 1768 (1996).

[21] For an introduction, see T. Bergeman, G. Erez, and H. J. Metcalf, Phys. Rev. A **35**, 1535 (1987).

[22] J. D. Weinstein and K. G. Libbrecht, Phys. Rev. A **52**, 4004 (1995); E. A. Hinds, M. G. Boshier, and I. G. Hughes, Phys. Rev. Lett. **80**, 645 (1998); J. Fortagh, A. Grossmann, C. Zimmermann, and T. W. Hänsch, *ibid.* **81**, 5310 (1998); J. Denschlag, D. Cassettari, and J. Schmiedmayer, *ibid.* **82**, 2014 (1999).

[23] J. P. Gordon and A. Ashkin, Phys. Rev. A **21**, 1606 (1980).

[24] C. Cohen-Tannoudji, in *Fundamental Systems in Quantum Optics*, Proceedings of the Les Houches Summer School 1990, edited by J. Dalibard, J.-M. Raimond, and J. Zinn-Justin (North-Holland, Amsterdam, 1992).

[25] Our definitions of Ω_{rabi} and I_{sat} correspond to those of C. S. Adams *et al.* [3].

[26] A. H. Barnett (unpublished).

[27] Even though the frequencies will actually differ by a couple of percent because of their detunings, and cause a difference in decay lengths of the correct sign, this effect alone is not large enough to create large α_L values. Future designs with larger detunings, possibly much larger than the fine splitting, may be able to use this effect alone.

[28] H. Kogelnik, in *Guided-Wave Optoelectronics*, edited by T. Tamir, 2nd ed. (Springer-Verlag, Berlin, 1990).

[29] Following the notation of Fernandez and Lu [35] for rectangular guides, mode E^k_{nm} has $n(m)$ antinodes in the $x(y)$ direction, and overall polarization axis $k=x,y$. However as the symmetry line $W=H$ is approached (see Fig. 2), mixing of polarizations occurs for the case $n\neq m$ and the notation becomes more arbitrary.

[30] E. Snitzer, J. Opt. Soc. Am. **51**, 491 (1961).

[31] We made the assumption that the positive and negative detunings caused negligible shifts in the wavelength used for the bound mode calculation. For the largest detunings used in this paper the fractional shift is 1.8%.

[32] P. Mezza, J. Phalippou, and P. Sempere, J. Non-Cryst. Solids **243**, 75 (1999); S. V. Nitta, V. Pisupatti, A. Jain, P. C. Wayner, Jr., W. N. Gill, J. L. Plawsky, J. Vac. Sci. Technol. B **17**, 205 (1999).

[33] G. Rempe, R. J. Thompson, H. J. Kimble, and R. Lalezari, Opt. Lett. **17**, 363 (1992).

[34] J. B. Davies, IEEE Trans. Magn. **29**, 1578 (1993).

[35] Y. Lu and F. A. Fernandez, IEEE Trans. Microwave Theory Tech. **41**, 1215 (1993).

[36] For a comparison of methods, see: S. Selleri and M. Zoboli, J. Opt. Soc. Am. **14**, 1460 (1997).

[37] E. A. Hinds and V. Sandoghdar, Phys. Rev. A **43**, 398 (1991).

[38] L. Spruch and Y. Tikochinsky, Phys. Rev. A **48**, 4213 (1993).

[39] I. E. Dzyaloshinskii, E. M. Lifshitz, and L. P. Pitaevskii, Adv. Phys. **10**, 165 (1961).

[40] Z-C. Yan, A. Dalgarno, and J. F. Babb, Phys. Rev. A **55**, 2882 (1997).

[41] A. Derevianko, W. R. Johnson, M. S. Safronova, and J. F. Babb, Phys. Rev. Lett. **82**, 3589 (1999).

[42] C. Henkel and M. Wilkens, Europhys. Lett. **47**, 414 (1999); e-print quant-ph/9902009.

[43] For some of the earliest bending loss estimation in a rectangular fiber see E. A. J. Marcatili, Bell Syst. Tech. J. **48**, 2103 (1969); and for the case of a cylindrical fiber, D. Marcuse, J. Opt. Soc. Am. **66**, 216 (1976). Most of the work done since then has been on the more complicated situation of fibers with cladding.

[44] P. Desbiolles, M. Arndt, P. Szriftgiser, and J. Dalibard, Phys. Rev. A **54**, 4292 (1996).

[45] Yu. B. Ovchinnikov, J. Söding, and R. Grimm, JETP Lett. **61**, 21 (1995); P. Desbiolles and J. Dalibard, Opt. Comm. **132**, 540 (1996).

AMERICAN MATHEMATICAL SOCIETY
MathSciNet® Mathematical Reviews on the Web

Citations
From References: 5
From Reviews: 0

MR2248896 (2007d:81085) 81Q50 (35J05 35Q40)

Barnett, A. H. (1-DTM)

Asymptotic rate of quantum ergodicity in chaotic Euclidean billiards. (English summary)

Comm. Pure Appl. Math. **59** (2006), *no. 10*, 1457–1488.

This paper contains very precise numerical tests of several conjectures about quantum ergodicity for a hyperbolic Euclidean billiard in two dimensions. Let $\Omega \subset \mathbb{R}^2$ be a compact domain with piecewise smooth boundary and let φ_j and E_j be the eigenfunctions and eigenvalues of the Laplace operator on Ω with Dirichlet boundary conditions. Let A be a function on Ω with average $\overline{A}$; then the autocorrelation function of A is defined as

$$C_A(t) = \int_{S^*\Omega} A \circ \Phi^t A - \overline{A}^{\,2}$$

where Φ^t is the billiard flow, and let $\widehat{C}_A(\omega)$ be its Fourier transform. Then it was conjectured by M. Feingold and A. Peres [Phys. Rev. A (3) **34** (1986), no. 1, 591–595; MR0848319 (87i:81055)] that if the billiard flow is hyperbolic then

$$(1) \quad \frac{1}{N_L(E)} \sum_{E \le E_j \le E+L} |\langle \varphi_j, A\varphi_j\rangle - \overline{A}|^2 = \frac{2\widehat{C}_A(0)}{\mathrm{vol}(\Omega)} E^{-1/2} + o(E^{-1/2})$$

where $N_L(E) = \#\{E_j \colon E \le E_j \le E+L\}$ with $L = O(E^{1/2})$, and furthermore, if one sets $\delta(E) = 2\pi/[E^{1/2}\mathrm{vol}(\Omega)]$, then

$$(2) \quad \frac{\delta(E)}{\varepsilon N_L(E)} \sum_{\substack{E \le E_j \le E+L \\ |E_i^{1/2} - E_j^{1/2} - \omega| \le \varepsilon}} |\langle \varphi_i, A\varphi_j\rangle - \delta_{ij}\overline{A}|^2 = \frac{2\widehat{C}_A(\omega)}{\mathrm{vol}(\Omega)} E^{-1/2} + o(E^{-1/2})$$

where $\varepsilon = O(E^{1/4})$. That the left-hand sides of (1) and (2) tend to zero for sufficiently nice functions A if the system is ergodic or weak mixing, respectively, is the content of quantum ergodicity theorems.

Here these conjectures are tested for a hyperbolic billiard of Sinai type, with A being the characteristic function of some large subset of Ω. Since the approach to the asymptotic regime is fairly slow it is important to be able to go to very large energies, and this is the main achievement of the paper, where energies up to $E \sim 10^7$ are considered. This is done using a numerical method called the scaling method. The results show very close agreement with the conjectures. This is particularly interesting because it has been recently shown that for eigenfunctions of the modular domain conjecture (1) is false, due to the arithmetic properties of the system [W. Z. Luo and P. C.

Sarnak, Ann. Sci. École Norm. Sup. (4) **37** (2004), no. 5, 769–799; MR2103474 (2005k:11101)].

It is furthermore reported that there are no signs of strong scars, which argues in favour of quantum unique ergodicity.

Reviewed by *Roman Schubert*

References

1. Agam, O.; Fishman, S. Semiclassical criterion for scars in wave functions of chaotic systems. *Phys. Rev. Lett.* **73** (1994), no. 6, 806–809. MR1285573 (95c:81028)
2. Aurich, R.; Taglieber, M. On the rate of quantum ergodicity on hyperbolic surfaces and for billiards. *Phys. D* **118** (1998), no. 1–2, 84–102. MR1629251 (99f:81055)
3. Austin, E. J.; Wilkinson, M. Distribution of matrix elements of a classically chaotic system. *Europhysics Lett.* **20** (1992), 589–593.
4. Bäcker, A.; Schubert, R.; Stifter, P. On the number of bouncing ball modes in billiards. *J. Phys. A* **30** (1997), no. 19, 6783–6795. MR1481345
5. Bäcker, A. Numerical aspects of eigenvalue and eigenfunction computations for chaotic quantum systems. *The mathematical aspects of quantum maps*, 91–144. Lecture Notes in Physics, 618. Springer, Berlin, 2003. MR2159324
6. Bäcker, A.; Schubert, R.; Stifter, P. Rate of quantum ergodicity in Euclidean billiards. *Phys. Rev. E (3)* **57** (1998), no. 5, part A, 5425–5447. Erratum: Rate of quantum ergodicity in Euclidean billiards. *Phys. Rev. E (3)* **58** (1998), no. 4, 5192. MR1626188 (99f:81056a)
7. Barnett, A. H. Quasi-orthogonality on the boundary for euclidean laplace eigenfunctions. arXiv:math-ph/0601006, 2006.
8. Barnett, A. H. The scaling method for the Dirichlet eigenproblem. In preparation.
9. Barnett, A. H. *Dissipation in deforming chaotic billiards.* Doctoral dissertation, Harvard University, 2000. Also available at http://www.math.dartmouth.edu/∼ahn/thesis_html.
10. Barnett, A. H.; Cohen, D.; Heller, E. J. Deformations and dilations of chaotic billiards: dissipation rate, and quasi-orthogonality of the boundary wavefunctions. *Phys. Rev. Lett.* **85** (2000), 1412. arXiv:nlin.CD/0003018, 2000.
11. Barnett, A.; Cohen, D.; Heller, E. J. Rate of energy absorption for a driven chaotic cavity. *J. Phys. A* **34** (2001), no. 3, 413–438. arXiv:nlin.CD/0006041, 2000. MR1822784 (2002a:81069)
12. Berry, M. V. Regular and irregular semiclassical wavefunctions. *J. Phys. A* **10** (1977), no. 12, 2083–2091. MR0489542 (58 #8961)
13. Berry, M. V. Quantum scars of classical closed orbits in phase space. *Proc. Roy. Soc. London Ser. A* **423** (1989), no. 1864, 219–231. MR1059234 (92c:81075)
14. Berry, M. Some quantum-to-classical asymptotics. *Chaos et physique quantique (Les Houches, 1989)*, 251–304. North-Holland, Amsterdam, 1991. MR1188420 (94e:81059)
15. Betcke, T.; Trefethen, L. Reviving the method of particular solutions. *SIAM Rev.* **47** (2005), no. 3, 469–491. MR2178637 (2006k:65344)
16. Bogomolny, E. B. Smoothed wave functions of chaotic quantum systems. *Phys. D* **31** (1988), no. 2, 169–189. MR0955627 (89h:81062)
17. Bogomolny, E. B.; Georgeot, B.; Giannoni, M.-J.; Schmit, C. Chaotic billiards generated by arithmetic groups. *Phys. Rev. Lett.* **69** (1992), no. 10, 1477–1480. MR1178884 (93g:81025)

18. Bohigas, O. Random matrix theories and chaotic dynamics. *Chaos et physique quantique (Les Houches, 1989)*, 87–199. North-Holland, Amsterdam, 1991. MR1188418 (94f:82038)
19. Boose, D.; Main, J. Distributions of transition matrix elements in classically mixed quantum systems. *Phys. Rev. E* **60** (1999), 2813–2844.
20. Casati, G.; Prosen, T. The quantum mechanics of chaotic billiards. Classical chaos and its quantum manifestations (Toulouse, 1998). *Phys. D* **131** (1999), no. 1–4, 293–310. MR1726941 (2001m:81082)
21. Cohen, D. Chaos and energy spreading for time-dependent Hamiltonians, and the various regimes in the theory of quantum dissipation. *Ann. Physics* **283** (2000), no. 2, 175–231. MR1769920 (2001g:81081)
22. Colin de Verdière, Y. Ergodicité et fonctions propres du laplacien. *Comm. Math. Phys.* **102** (1985), no. 3, 497–502. MR0818831 (87d:58145)
23. Combescure, M.; Robert, D. Semiclassical sum rules and generalized coherent states. *J. Math. Phys.* **36** (1995), no. 12, 6596–6610. MR1359645 (97d:81069)
24. Donnelly, H. Quantum unique ergodicity. *Proc. Amer. Math. Soc.* **131** (2003), no. 9, 2945–2951. MR1974353 (2005a:58048)
25. Eckhardt, B.; Fishman, S.; Keating, J.; Agam, O.; Main, J.; Müller, K. Approach to ergodicity in quantum wave functions. *Phys. Rev. E* **52** (1995), 5893–5903.
26. Faure, F.; Nonnenmacher, S.; De Bièvre, S. Scarred eigenstates for quantum cat maps of minimal periods. *Comm. Math. Phys.* **239** (2003), no. 3, 449–492. MR2000926 (2005a:81076)
27. Feingold, M.; Peres, A. Distribution of matrix elements of chaotic systems. *Phys. Rev. A (3)* **34** (1986), no. 1, 591–595. MR0848319 (87i:81055)
28. Gardiner, C. W. *Handbook of stochastic methods for physics, chemistry and the natural sciences.* 3rd ed. Springer Series in Synergetics, 13. Springer, Berlin, 2004. MR2053476 (2004m:00008)
29. Gutzwiller, M. C. *Chaos in classical and quantum mechanics.* Interdisciplinary Applied Mathematics, 1. Springer, New York, 1990. MR1077246 (91m:58099)
30. Heller, E. J. Bound-state eigenfunctions of classically chaotic Hamiltonian systems: scars of periodic orbits. *Phys. Rev. Lett.* **53** (1984), no. 16, 1515–1518. MR0762412 (85k:81055)
31. Heller, Eric J. Wavepacket dynamics and quantum chaology. *Chaos et physique quantique (Les Houches, 1989)*, 547–664. North-Holland, Amsterdam, 1991. MR1188425 (94i:81031)
32. Hortikar, S.; Srednicki, M. Random matrix elements and eigenfunctions in chaotic systems. *Phys. Rev. E (3)* **57** (1998), no. 6, 7313–7316. MR1628191 (99a:81045)
33. Kaplan, L.; Heller, E. J. Linear and nonlinear theory of eigenfunction scars. *Ann. Physics* **264** (1998), no. 2, 171–206. MR1619193 (99f:81058)
34. Kaplan, L.; Heller, E. J. Measuring scars of periodic orbits. *Phys. Rev. E (3)* **59** (1999), no. 6. 6609–6628. MR1695455 (2000b:81037)
35. Kuttler, J. R.; Sigillito, V. G. Eigenvalues of the Laplacian in two dimensions. *SIAM Rev.* **26** (1984), no. 2, 163–193. MR0738929 (85k:65086)
36. Li, B.; Hu, B. Statistical analysis of scars in stadium billiard. *J. Phys. A* **31** (1998), no. 2, 483–504. MR1629174 (99b:81061)
37. Lindenstrauss, E. Invariant measures and arithmetic quantum unique ergodicity. *Ann. of Math.*

(2) **163** (2006), no. 1, 165–219. MR2195133 (2007b:11072)
38. Luo, W.; Sarnak, P. Quantum variance for Hecke eigenforms. *Ann. Sci. École Norm. Sup. (4)* **37** (2004), no. 5, 769–799. MR2103474 (2005k:11101)
39. Martinez, A. *An introduction to semiclassical and microlocal analysis.* Universitext. Springer, New York, 2002. MR1872698 (2003b:35010)
40. Prosen, T. Statistical properties of matrix elements in a Hamilton system between integrability and chaos. *Ann. Physics* **235** (1994), no. 1, 115–164. MR1294439 (95g:81040)
41. Prosen, T. Quantization of a generic chaotic 3D billiard with smooth boundary. I. Energy level statistics. *Phys. Lett. A* **233** (1997), no. 4–6, 323–331. MR1474466 (99a:81046)
42. Prosen, T. Quantization of generic chaotic 3D billiard with smooth boundary. II. Structure of high-lying eigenstates. *Phys. Lett. A* **233** (1997), no. 4–6, 332–342. MR1474467 (99a:81047)
43. Rudnick, Z; Sarnak, P. The behaviour of eigenstates of arithmetic hyperbolic manifolds. *Comm. Math. Phys.* **161** (1994), no. 1, 195–213. MR1266075 (95m:11052)
44. Sarnak, P. Arithmetic quantum chaos. *The Schur lectures (1992) (Tel Aviv)*, 183–236. Israel Mathematical Conference Proceedings, 8. Bar-Ilan University, Ramat Gan, Israel, 1995. MR1321639 (96d:11059)
45. Sarnak, P. Spectra and eigenfunctions of Laplacians. *Partial differential equations and their applications (Toronto, ON, 1995)*, 261–276. CRM Proceedings & Lecture Notes, 12. American Mathematical Society, Providence, R.I., 1997. MR1479252 (98i:58240)
46. Sarnak, P. Spectra of hyperbolic surfaces. *Bull. Amer. Math. Soc. (N.S.)* **40** (2003), no. 4, 441–478. MR1997348 (2004f:11107)
47. Schanz, H.; Kottos, T. Scars on quantum networks ignore the lyapunov exponent. *Phys. Rev. Lett.* **90** (2003), 234101.
48. Šnirel'man, A. I. Ergodic properties of eigenfunctions. (Russian) *Uspehi Mat. Nauk* **29** (1974), no. 6 (180), 181–182. MR0402834 (53 #6648)
49. Sinaĭ, Ya. G. Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards. *Uspehi Mat. Nauk* **25** (1970), no. 2 (152), 141–192. MR0274721 (43 #481)
50. Sivia, D. S. *Data analysis. A Bayesian tutorial.* Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1996. MR1420864 (97k:62003)
51. Steiner, F. *Quantum chaos*. arXiv:chao-dyn/9402001, 1994. MR1726564
52. Tanner, G. How chaotic is the stadium billiard? A semiclassical analysis. *J. Phys. A* **30** (1997), no. 8, 2863–2888. MR1450353
53. Tate, T. Some remarks on the off-diagonal asymptotics in quantum ergodicity. *Asymptot. Anal.* **19** (1999), no. 3–4, 289–296. MR1696219 (2000d:58041)
54. Veble, G.; Robnik, M.; Liu, J. Study of regular and irregular states in generic systems. *J. Phys. A* **32** (1999), no. 36, 6423–6444. MR1732841 (2000k:81106)
55. Vergini, E. Estudio cuántico y semiclásico de billares clásicamente caóticos. Doctoral dissertation, Universidad de Buenos Aires, 1995.
56. Vergini, E.; Saraceno, M. Calculation by scaling of highly excited states of billiards. *Phys. Rev. E* **52** (1995), no. 3, 2204–2207.
57. Wilkinson, M. A semiclassical sum rule for matrix elements of classically chaotic systems. *J. Phys. A* **20** (1987), no. 9, 2415–2423. MR0895065 (89f:81071)

58. Zelditch, S. Uniform distribution of eigenfunctions on compact hyperbolic surfaces. *Duke Math. J.* **55** (1987), no. 4, 919–941. MR0916129 (89d:58129)
59. Zelditch, S. Quantum transition amplitudes for ergodic and for completely integrable systems. *J. Funct. Anal.* **94** (1990), no. 2, 415–436. MR1081652 (92b:58181)
60. Zelditch, S. Quantum ergodicity on the sphere. *Comm. Math. Phys.* **146** (1992), no. 1, 61–71. MR1163667 (93e:58191)
61. Zelditch, S. On the rate of quantum ergodicity. I. Upper bounds. *Comm. Math. Phys.* **160** (1994), no. 1, 81–92. MR1262192 (95f:58084)
62. Zelditch, S. On the rate of quantum ergodicity. II. Lower bounds. *Comm. Partial Differential Equations* **19** (1994), no. 9–10, 1565–1579. MR1294471 (96e:58161)
63. Zelditch, S. Quantum mixing. *J. Funct. Anal.* **140** (1996), no. 1, 68–86. MR1404574 (98f:58197)
64. Zelditch, S. A random matrix model for quantum mixing. *Internat. Math. Res. Notices* **1996.** no. 3, 115–137. MR1383753 (97j:81085)
65. Zelditch, S. Quantum ergodicity and mixing of eigenfunctions. *Elsevier Encyclopedia of Mathematical Physics*, to appear. arXiv:math-ph/0503026, 2005. cf. MR 2006b:81086
66. Zelditch, S.; Zworski, M. Ergodicity of eigenfunctions for ergodic billiards. *Comm. Math. Phys.* **175** (1996), no. 3, 673–682. MR1372814 (97a:58193)

Note: This list reflects references listed in the original paper as accurately as possible with no attempt to correct errors.

© *Copyright American Mathematical Society 2007, 2010*

© 2010 European Mathematical Society, FIZ Karlsruhe & Springer-Verlag

---

Zbl 1170.65082

Barnett, A.H.; Betcke, T.

Stability and convergence of the method of fundamental solutions for Helmholtz problems on analytic domains. (English)

J. Comput. Phys. 227, No. 14, 7003-7026 (2008). ISSN 0021-9991
http://dx.doi.org/10.1016/j.jcp.2008.04.008
http://www.sciencedirect.com/science/journal/00219991

This paper is concerned with the numerical approximation of solutions to the Helmholtz equation $\Delta u + k^2 u = 0$ in Ω, subject to Dirichlet boundary condition $u = v$ on $\partial\Omega$. Here $\Omega \subset \mathbb{C}$ is a simply connected domain with analytic boundary. The idea of the proof is to use the Method of Fundamental Solutions (MFS), that is, to approximate the solution u as a linear combination of fundamental solutions, $\sum_{j=1}^{N} \alpha_j H_0^{(1)}(k|x-y_j|)$ with $y_j \in \mathbb{C} \setminus \overline{\Omega}$, where $H_0^{(1)}$ is the Hankel function of the first kind with order zero. The main goal of the paper is to investigate conditions on the charges y_j that lead to an accurate and stable numerical scheme in the framework of the MFS. The authors emphasize that this fact depends on how far into the complex plane the solution u can be analytically continued. More precisely, it is obtained in the paper that for a general analytic domain Ω, the position of the charge points relative to the singularities of the analytic continuation of u is crucial for the accuracy and numerical stability of the MFS. Various numerical experiments on many domain shapes are presented that support the conclusion.

Marius Ghergu (Dublin)

Keywords : Helmholtz equation; method of fundamental solutions; analytic continuation; charge curve; stability; convergence; high frequency waves; numerical experiments

Classification :

- ∗65N12 Stability and convergence of numerical methods (BVP of PDE)
- 65N35 Collocation methods (BVP of PDE)
- 35J05 Laplace equation, etc.
- 78M25 Numerical methods in optics
- 78A25 General electromagnetic theory

AMERICAN MATHEMATICAL SOCIETY
MathSciNet® Mathematical Reviews on the Web

Citations
From References: 0
From Reviews: 0

MR2519590 (2010i:65250) 65N25 (35J05 35P15 47A55 47A75 65N15)

Barnett, A. H. (1-DTM)

Perturbative analysis of the method of particular solutions for improved inclusion of high-lying Dirichlet eigenvalues. (English summary)

SIAM J. Numer. Anal. **47** (2009), *no. 3,* 1952–1970.

The author derives an inclusion bound on the Dirichlet eigenvalues of the Laplace operator. The bound reflects the true slopes of the curves appearing in the method of particular solutions and, for higher values of the eigenvalue λ, it improves the classical bound of C. B. Moler and L. E. Payne [SIAM J. Numer. Anal. **5** (1968), 64–70; MR0226833 (37 #2420)] by a factor of $O(\lambda^{1/2})$. The paper also includes a list of interesting open questions.

Reviewed by *Alan L. Andrew*

References

1. A. B"acker, *Numerical aspects of eigenvalue and eigenfunction computations for chaotic quantum systems*, in The Mathematical Aspects of Quantum Maps, Lect. Notes Phys. 618, Springer, Berlin, 2003, pp. 91–144. MR2159324
2. A. H. Barnett and T. Betcke, *Stability and convergence of the method of fundamental solutions for Helmholtz problems on analytic domains*, J. Comput. Phys., 227 (2008), pp. 7003–7026. MR2435440 (2009g:65178)
3. A. H. Barnett, *Dissipation in Deforming Chaotic Billiards*, Ph.D. thesis, Harvard University, Cambridge, MA, 2000, http://www.math.dartmouth.edu/~ahb/thesishtml/.
4. A. H. Barnett, *Asymptotic rate of quantum ergodicity in chaotic Euclidean billiards*, Comm. Pure Appl. Math., 59 (2006), pp. 1457–88. MR2248896 (2007d:81085)
5. T. Betcke and L. N. Trefethen, *Reviving the method of particular solutions*, SIAM Rev., 47 (2005), pp. 469–491. MR2178637 (2006k:65344)
6. T. Betcke, *A GSVD formulation of a domain decomposition method for planar eigenvalue problems*, IMA J. Numer. Anal., 27 (2007), pp. 451–478. MR2337576 (2008f:65071)
7. T. Betcke, *The generalized singular value decomposition and the method of particular solutions*, SIAM J. Sci. Comput., 30 (2008), pp. 1278–1295. MR2398865 (2009f:65082)
8. J. Descloux and M. Tolley, *An accurate algorithm for computing the eigenvalues of a polygonal membrane*, Comput. Methods Appl. Mech. Engrg., 39 (1983), pp. 37–53. MR0712205 (84i:73056)
9. T. A. Driscoll, *Eigenmodes of isospectral drums*, SIAM Rev., 39 (1997), pp. 1–17. MR1439483 (97m:65198)
10. S. C. Eisenstat, *On the rate of convergence of the Bergman–Vekua method for the numerical solution of elliptic boundary value problems*, SIAM J. Numer. Anal., 11 (1974), pp. 654– 680. MR0375805 (51 #11995)
11. S. C. Eisenstat, *Bounds for Eigenvalues and Eigenvectors of Self-adjoint Operators*, preprint,

2006.
12. R. Ennenbach and H. Niemeyer, *The inclusion of Dirichlet eigenvalues with singularity functions*, Z. Angew. Math. Mech., 76 (1996), pp. 377–383. MR1397260 (97b:35030)
13. L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 1998. MR1625845 (99e:35001)
14. G. Fairweather and A. Karageorghis, *The method of fundamental solutions for elliptic boundary value problems*, Adv. Comput. Math., 9 (1998), pp. 69–95. MR1662760 (2000a:65165)
15. G. Fix and R. Heiberger, *An algorithm for the ill-conditioned generalized eigenvalue problem*, SIAM J. Numer. Anal., 9 (1972), pp. 78–88.1970 A. H. BARNETT MR0305569 (46 #4699)
16. L. Fox, P. Henrici, and C. Moler, *Approximations and bounds for eigenvalues of elliptic operators*, SIAM J. Numer. Anal., 4 (1967), pp. 89–102. MR0215542 (35 #6382)
17. L. Friedlander, *Some inequalities between Dirichlet and Neumann eigenvalues*, Arch. Ration. Mech. Anal., 116 (1991), pp. 153–160. MR1143438 (93h:35146)
18. P. R. Garabedian, *Partial Differential Equations*, John Wiley & Sons, New York, 1964. MR0162045 (28 #5247)
19. P. G'erard and E. Leichtnam, *Ergodic properties of eigenfunctions for the Dirichlet problem*, Duke Math. J., 71 (1993), pp. 559–607. MR1233448 (94i:35146)
20. G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd ed., Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 1996. MR1417720 (97g:65006)
21. C. Gordon, D. Webb, and S. Wolpert, *Isospectral plane domains and surfaces via Rieman- nian orbifolds*, Invent. Math., 110 (1992), pp. 1–22. MR1181812 (93h:58172)
22. M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics*, Interdiscip. Appl. Math., Springer, New York, 1990. MR1077246 (91m:58099)
23. A. Hassell, *personal communication*, 2005.
24. A. Hassell and T. Tao, *Upper and lower bounds for normal derivatives of Dirichlet eigenfunctions*, Math. Res. Lett., 9 (2002), pp. 289–305. MR1909646 (2003k:58047)
25. E. J. Heller, *Bound-state eigenfunctions of classically chaotic Hamiltonian systems: Scars of periodic orbits*, Phys. Rev. Lett., 53 (1984), pp. 1515–1518. MR0762412 (85k:81055)
26. E. J. Heller, *Wavepacket dynamics and quantum chaology*, in Chaos et Physique Quantique (Les Houches, 1989), North-Holland, Amsterdam, 1991, pp. 547–664. MR1188425 (94i:81031)
27. A. Karageorghis, *The method of fundamental solutions for the calculation of the eigenvalues of the Helmholtz equation*, Appl. Math. Lett., 14 (2001), pp. 837–842. MR1849238 (2002e:65167)
28. J. R. Kuttler and V. G. Sigillito, *Inequalities for membrane and Stekloff eigenvalues*, J. Math. Anal. Appl., 23 (1968), pp. 148–160. MR0226226 (37 #1816)
29. J. R. Kuttler and V. G. Sigillito, *Bounding eigenvalues of elliptic operators*, SIAM J. Math. Anal., 9 (1978), pp. 768–778. MR0492948 (58 #11996)
30. J. R. Kuttler and V. G. Sigillito, *Eigenvalues of the Laplacian in two dimensions*, SIAM Rev., 26 (1984), pp. 163–193. MR0738929 (85k:65086)
31. J.-L. Lions and E. Magenes, *Probl'emes aux limites non homog'enes et applications, Vol.* 3, Dunod, Paris, 1970. Travaux et Recherches Math'ematiques, 20. MR0291887 (45 #975)
32. W. C. H. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge

University Press, London, 2000. MR1742312 (2001a:35051)
33. C. B. Moler and L. E. Payne, *Bounds for eigenvalues and eigenvectors of symmetric operators*, SIAM J. Numer. Anal., 5 (1968), pp. 64–70. MR0226833 (37 #2420)
34. A. Nachman, J. Sylvester, and G. Uhlmann, *An n-dimensional Borg-Levinson theorem*, Comm. Math. Phys., 115 (1988), pp. 595–605. MR0933457 (89g:35082)
35. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed., Cambridge University Press, Cambridge, 2002. MR1880993 (2003a:65005)
36. M. Reed and B. Simon, *Methods of Modern Mathematical Physics*. IV. *Analysis of Operators*, Academic Press, New York, 1978. MR0493421 (58 #12429c)
37. M. Reed and B. Simon, *Methods of Modern Mathematical Physics*. I. *Functional Analysis*, 2nd ed., Academic Press, New York, 1980. MR0751959 (85e:46002)
38. F. Rellich, *Darstellung der Eigenwerte von* $\Delta u + \lambda u = 0$ *durch ein Randintegral*, Math. Z., 46 (1940), pp. 635–636. MR0002456 (2,56d)
39. F. Rellich, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, 1969. MR0240668 (39 #2014)
40. N. Saito, *Data analysis and representation on a general domain using eigenfunctions of Laplacian*, Appl. Comput. Harmon. Anal., 25 (2008), pp. 68–97. MR2419705 (2009d:35038)
41. N. L. Schryer, *Constructive approximation of solutions to linear elliptic boundary value problems*, SIAM J. Numer. Anal., 9 (1972), pp. 546–572. MR0321327 (47 #9860)
42. G. Still, *Computable bounds for eigenvalues and eigenfunctions of elliptic differential operators*, Numer. Math., 54 (1988), pp. 201–223. MR0965922 (89k:65130)
43. L. N. Trefethen and T. Betcke, *Computed eigenmodes of planar regions*, Contemp. Math., 412 (2006), pp. 297–314. MR2259116 (2008a:35042)
44. L. N. Trefethen, *Spectral Methods in MATLAB*, Vol. 10, Software, Environments, and Tools, SIAM, Philadelphia, PA, 2000. MR1776072 (2001c:65001)
45. H. E. Tureci, H. G. L. Schwefel, P. Jacquod, and A. D. Stone, *Modes of wave-chaotic dielectric resonators*, Prog. Optics, 47 (2005), pp. 75–137.
46. E. Vergini and M. Saraceno, *Calculation by scaling of highly excited states of billiards*, Phys. Rev. E, 52 (1995), pp. 2204–2207.
47. S. Zelditch, *Quantum ergodicity and mixing of eigenfunctions*, in Elsevier Encyclopedia of Mathematical Physics, Vol. 1, Academic Press/Elsevier Science, Oxford, 2006, pp. 183–196, arXiv:math-ph/0503026.

Note: This list, extracted from the PDF form of the original paper, may contain data conversion errors, almost all limited to the mathematical expressions.

© *Copyright American Mathematical Society 2010*

My favorites ▼ | Sign in

mpspack

A MATLAB toolbox to solve Helmholtz PDE, wave scattering, and eigenvalue problems using particular solutions and integral equations

Search projects

Project Home Downloads Wiki Issues Source

Summary | Updates | People

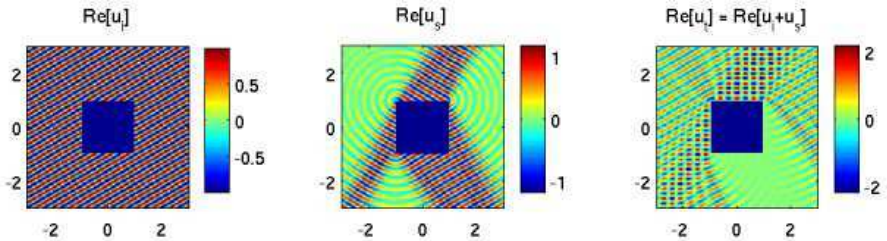
MPSpack is a user-friendly and fully object-oriented MATLAB toolbox that implements the method of particular solutions, nonpolynomial FEM, and related boundary methods (e.g. fundamental solutions, layer potentials) for efficient and highly accurate solution of Laplace eigenvalue problems, interior/exterior Helmholtz boundary-value problems (e.g. wave scattering), and related PDE problems, on piecewise-homogeneous 2D domains.

We released Version 1.0 in September 2009. As of June 2010, the current Version 1.1 beta includes new methods for quasi-periodic (grating) scattering.

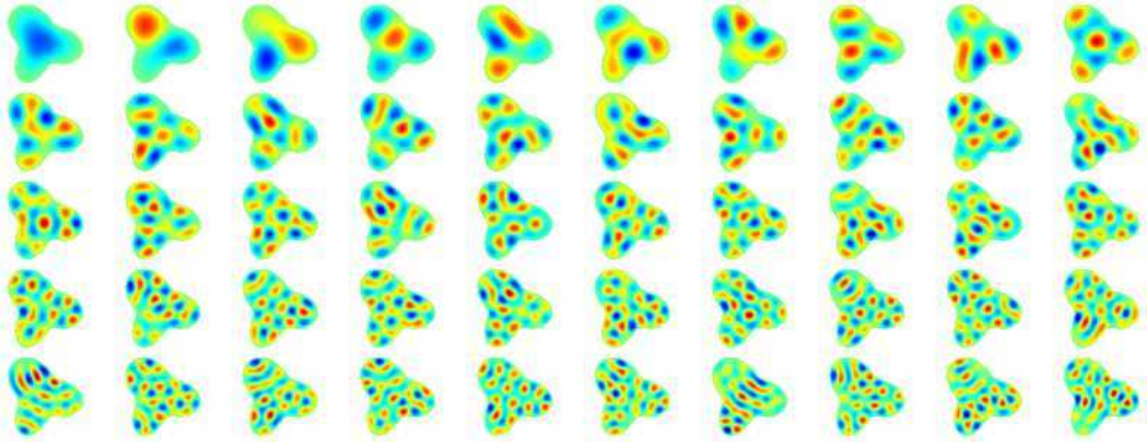
Please see the **Downloads** page for a gzipped tar archive of the package, the manual which has installation instructions, and the all-important tutorial. See the **Source** page for how to download via `svn` (subversion); this gives access to the current development version, including e.g. a Dirichlet eigenmode solver.

This material is based upon work supported by the National Science Foundation under grant DMS-0811005, and Engineering and Physical Sciences Research Council Grant EP/H00409/1.

First is an example showing scattering from a square, accurate to 10 digits, computed in a few seconds on a laptop. Spectral convergence is achieved using the following ingredients: decomposition into subdomains (nonpolynomial FEM), fractional-order Fourier-Bessel expansions around corner singularities, and an exterior fundamental solutions representation. With MPSpack this needs no more than 20 lines of code.



Next is an example showing the first 45 Dirichlet eigenmodes of a smooth planar domain, computed to 12 digit accuracy and evaluated on a grid of 3600 points, in around 1 second per mode. Convergence is again spectral, using layer potential representations, and analytic root-finding on a Fredholm determinant. With MPSpack the code is 9 lines long.



Activity: Medium

Code license:
GNU General Public License v3

Labels:
MATLAB, Helmholtz, ParticularSolutions, FundamentalSolutions, BoundaryIntegral, scattering, Trefftz, PDE, wave, numerical, nonpolynomial, quasi-periodic, eigenvalue, acoustic, eigenfunction

External links:
Alex Barnett website
Timo Betcke website

Feeds:
Project feeds

Owners:
alex.h.barnett, timo.betcke

People details »

©2010 Google - Terms - Privacy - Project Hosting Help

Powered by Google Project Hosting

My favorites ▼ | Sign in

mpspack

A MATLAB toolbox to solve Helmholtz PDE, wave scattering, and eigenvalue problems using particular solutions and integral equations

Search projects

Project Home Downloads Wiki Issues Source

Search Current downloads for Search

1 - 8 of 8

	Filename ▼	Summary + Labels ▼	Uploaded ▼	Size ▼	DownloadCount ▼	...
	README	README for version 1.1 beta	Aug 02	380 bytes	23	
	mpspack-1.1beta.tar.gz	MPSpack version 1.1beta (tarball)	Jun 22	2.9 MB	12	
	mpspack-1.1beta.zip	MPSpack version 1.1beta (zip archive)	Jun 22	3.0 MB	50	
	tutorial.pdf	Tutorial (version 1.1beta), building up from simple through complex examples (35 pages)	Jun 21	1.3 MB	238	
	mpspack-1.0.zip	MPSpack version 1.0 (zip archive)	Aug 2009	1.6 MB	131	
	mpspack-1.0.tar.gz	MPSpack version 1.0 (tarball)	Aug 2009	1.6 MB	44	
	manual.pdf	User manual (version 1.1beta) (37 pages)	Aug 2009	313 KB	335	
	p.pdf	Relevant preprint by Barnett-Betcke, appeared, SIAM J. Sci. Comp. 32 (3), 1417-1441 (2010)	Aug 2009	623 KB	210	

1 - 8 of 8

©2010 Google - Terms - Privacy - Project Hosting Help

Powered by Google Project Hosting

ISSN 0002-9920

Notices

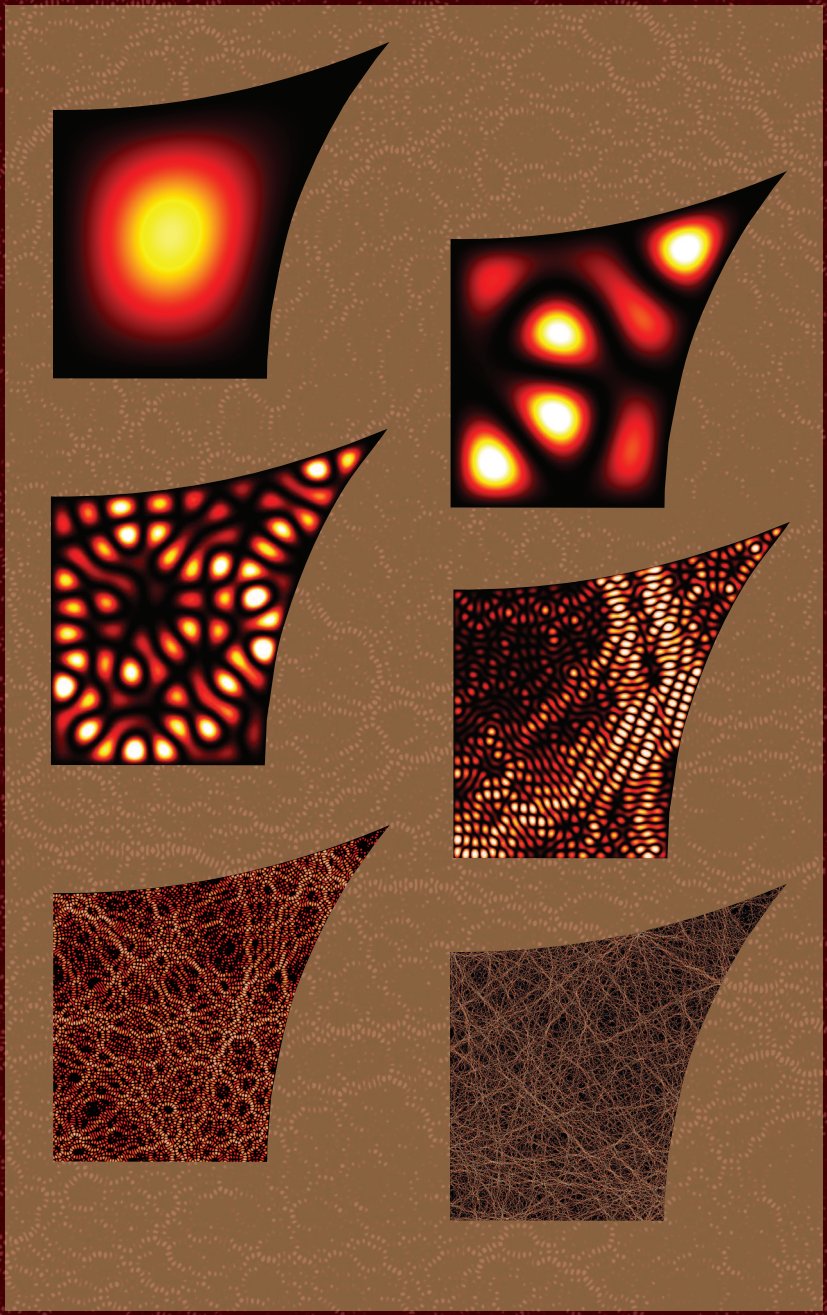
of the American Mathematical Society

January 2008 Volume 55, Number 1

An Evaluation of Mathematics Competitions Using Item Response Theory
page 8

Your Hit Parade: The Top Ten Most Fascinating Formulas in Ramanujan's Lost Notebook
page 18

New York Meeting
page 98

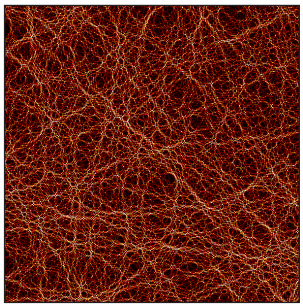


Quantum chaos (see page 41)

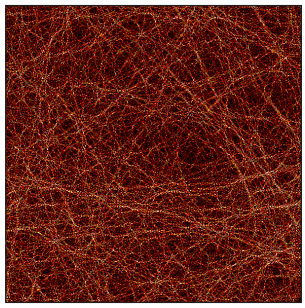
About the Cover

Quantum Chaos

This month's cover accompanies Ze'ev Rudnick's article on quantom chaos. The smaller images illustrate plots of eigenfunctions at progressively higher frequencies for the Dirichlet problem on the curvilinear quadrilateral shown in Rudnick's article. They start from the first and go through the 100,000-th eigenvalue in steps of powers of 10. The background of the cover is a random combination of plane waves, and the point is that in a region giving rise to quantum chaos eigenfunctions of high frequencies are conjectured to be locally similar to such random combinations.



Eigenfunctions of high frequency (cropped from an image on the cover).



Random combination of plane waves. The statistical meaning of the stringy structures in these images does not seem to be well understood (cropped from an image on the cover).

All of the images were produced by Alex Barnett of Dartmouth College, who also provided the data for the figures in Rudnick's article. They were computed using a variant of the Method of Particular Solutions (MPS) called the scaling method, which he has had a hand in developing. It is faster by an order of 10^3 than competing techniques, and is explained in recent papers of his. A good introduction is "Asymptotic rate of quantum ergodicity in chaotic Euclidean billiards", *Comm. Pure Appl. Math.*, 2006.

—Bill Casselman, Graphics Editor
(notices-covers@ams.org)

convolution: *son et lumière*

Alex Barnett

October 5, 2009

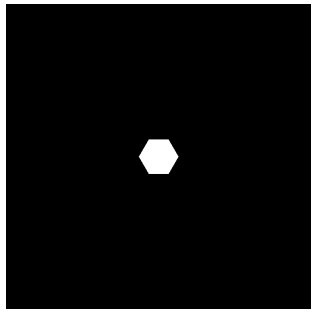
Exterior night, cityscape. Thousands of distant points of light lie still in the summer air. Below them a neon diner sign flickers silently across a nearby empty street. Suddenly the close up face of the main character moves into the frame, the focus pulls, bringing their silent profile into crisp silhouette, while blurring the urban backdrop into a splash of color. Each point of light blooms into a hexagon, the neon sign now an unreadable splotch of red. The main character pauses for a second, then moves on into the night. Fade to black.

Something special just happened in this imagined movie scene, when viewed through a scientific lens. The cityscape image, in its transition from crisp to out-of-focus, underwent a *convolution*, a process with a precise mathematical meaning that will become *our* main character in this short piece. I will try to explain as accurately as possible the concept of convolution, using familiar examples from our everyday lives. Convolution is in fact all around us—we just have to attune ourselves to it, to shift to a viewpoint in which lights and sounds and motions become patterns of numbers and signals. It could be argued that this type of shift, with the logical power to model and predict that it brings, is an idea that has opened the doors to centuries of tremendous human scientific and technological creation. I hope to leave you with a glimpse of how an engineer or scientist, or at least this applied mathematician, thinks about the world.[1]

The classic cinematic trope above, toying with depth of field, mirrors what our eyes and brain do daily, mostly unconsciously, as we shift attention in a complex three-dimensional visual world. The myopic among us may recreate the same blurring effect for free, simply by taking off our glasses and staring into the distance (here the hexagons are instead discs). Let's examine more carefully what took place. We may treat the cityscape image as a pattern of light intensity existing across a two-dimensional plane, as in a rectangular photographic still, or a grid of pixels on your computer screen. Just like Manhattan's grid of avenues and streets, each point in this plane has a location given by its horizontal distance from the left side (x coordinate) and vertical distance up from the bottom (y coordinate). At each such point the image has a brightness, which we may treat as a number, zero for black, one for bright, two for very bright, ... and so forth. So we have a number at each point in the plane: mathematically

[1] In footnotes I will add a flavor of the mathematical formulae, but these may be skipped without missing much of the story.

scene image f



aperture image g

convolution image of f and g

this is called a function.[2]

Where did the hexagons come from? They are copies of the shape of the camera aperture or iris, the other ingredient in the mix. Return to the opening scene. If the cityscape had instead consisted of only a *single* bright point of light, you can picture what would happen upon pulling focus: the single point blurs into a single, uniformly-bright hexagon, surrounded by blackness. The key is to realise that this bright-hexagon-on-black-background is *also* itself an image (a brightness pattern existing over a two-dimensional plane)—we might call it the 'aperture image'. We now have two images: the cityscape 'scene image' and the aperture image (see figure). The act of blurring the scene image (a function, let's call it f) using the aperture image (another function, let's call it g), is precisely to *convolve* one by the other, making a new image written mathematically as,[3]

$$f * g$$

The rule is that each point of light in the scene is replaced by an identical copy of the hexagonal aperture image, and these individual copies are then layered on top of each other, or added up, to give the final blurry picture.[4] If you like, we are painting the original scene using a hexagon-shaped paintbrush. The act was performed by turning a focus ring on a camera, sending a flurry of photons into different places on the film. However, it may also be performed mathematically or computationally, and, once it has been defined precisely, abstracted to any images, functions or data, real or imagined: therein lies its power.

It is a beautiful and not obvious fact that if we swapped the roles of our two images, the picture produced would be exactly the same as before. In other words convolution has a *symmetry*,[5] and that is something that delights the mathematical glands, for both aesthetic and practical reasons. How may

[2] We could name this function f, in which case the brighness at a point (x, y) is $f(x, y)$.

[3] Note that the asterisk means convolution, not mere multiplication.

[4] The adding or summing up operation needs to be done for every single point in the scene image: there are a continuum of points, which mathematically turns the summation into an *integral*, and since there are two directions to sum over, a double integral. Hence the formula for a two-dimensional convolution, $(f * g)(x, y) = \int\int f(w, z)g(x - w, y - z)dw\,dz$.

[5] We may state $g * f = f * g$ as functions, for all f and g. We won't prove it rigorously here.

we convince ourselves of this fact by mental visualization? In the swapped situation we take each point of light within the aperture hexagon and replace it by a copy of the whole cityscape scene, suitably jiggled to be centered at the point on the hexagon from which it came. Again we add up all these copies. The brightness at any point is now a sum of intensities over all the nearby points in the scene image lying in a hexagon centered at that point. But, thinking back to the original situation, we realise this describes that too. Painting a scene using an aperture-shaped brush is equivalent to painting an aperture using a scene-shaped brush!

Let's switch senses from light to sound; we will find convolution lurking almost everywhere we listen. Speak aloud in a large echoey room and your usual speech is changed by the acoustics of the space into a boomy, muddied version: this is *convolution of two sounds*. What do our actors f and g signify here? A sound is air pressure or vibrational motion changing very rapidly in time, which we may describe as a function.[6] Speak aloud outdoors in still air and your voice sounds crisp and dry—this is, give or take, your voice's source function f. On the other hand, clap in the silent large room, and you will hear an echoey reverberation dying away—this is the room's 'response function' g. Now combine the two: emit your voice into the room and the resulting[7] audio signal is precisely $f * g$. To help visualize this the reader might enjoy messing around with a wonderful online interactive demonstration of convolving two hand-drawn functions [1].

It is worth drawing some connections between this and our original case of images. Muddying of your voice by a large echoey room is analogous to blurring of the scene by a wide out-of-focus camera aperture. In order to produce a final result that was simply g, we sent in for f a 'clap' (very short spike-like signal) in the audio case, or a 'bright point' (very narrow spike-like image) in the image case.[8] The power of applied mathematics is that these seemingly unrelated phenomena are described by (essentially) the same formulae.

The sonic example allows us to explain a final property of convolution—one which has many applications in signal processing—again using everyday experiences. I mentioned that sounds are pressure variations as a function of time; however, there is a complementary way to describe them by the strength (amplitude) with which each component *frequency* is present.[9] Many home stereos (and audio player applications) show with a set of illuminated bars how much signal strength is present in the ranges 50-250Hz, 250-1000Hz, etc, from the bottom to the top of our hearing range. Moreover, by adjusting the 'graphic equalizer' on your stereo (or pressing buttons such as 'bass boost') you are able to selectively amplify or reduce the strengths in each of these frequency

[6] Now the function depends on only one variable, the time t. It is often called a signal.

[7] The relevant formula for convolution in one variable is $(f * g)(t) = \int f(s)g(t-s)ds$.

[8] Mathematically these special short signals are known as *delta distributions*.

[9] Recall that frequency is how many times per second something vibrates, and, crudely speaking, corresponds to musical pitch. In isolation, a single frequency sounds like a pure tone such as that made by a tuning fork. Combinations of frequencies create all the wonderful variations in musical timbre. The frequency description is called the *Fourier transform*.

bands, 'shaping' or coloring the sound according to a curve. Mathematically this multiplies the amplitudes by a function which depends on frequency: for instance if you make this function larger at high frequencies than at low ones you will get a bright, tinny, or sibilant effect. If instead you choose a 'flat' or constant function you don't color the sound at all, just like the humble volume knob which merely multiplies the amplitude by the *same* number at all frequencies.

We now have a picture of coloring a sound by enhancing or subduing certain frequencies: it turns out that this is in fact *precisely* what an echoey room does to the sound of your voice in the previous example, when viewed through a 'frequency lens'. Probably the most convincing demonstration of this is by singing in the shower: a shower cabinet is a small echoey room so we know it has a response function (clap and you'll hear it); on the other hand you know that when you sing or hum you find that certain notes of the scale stand out as surprisingly loud and resonant compared to others, indicating heavy coloration of your voice. The remarkable property we have thus shown is that this process of coloring (or multiplying) in the frequency picture is identical to convolving the sound with some response function. All of this can be proven rigorously,[10] and used to manipulate audio with computers to great effect [2].

So, how does this mathematical toolbox help humanity? One real-world problem that repeatedly crops up is that of *deconvolution*: say you took an out-of-focus image, or your camera moved by accident during the exposure, turning your image into a riot of useless squiggly lines—how could you recover the original unsullied image? Applications vary from forensic investigation of blurred photographs of license plates, to astronomy [3], seismic imaging, and improving the MRI images used to diagnose cancer [4]. Maybe the discussion above has given you some hints: transform into the 'frequency picture', then undo the 'graphic equalizer' effect (by dividing rather than multiplying), finally transform back to the signal or image space. Barring certain computational details and tricks, this is in fact what is done: the results can be amazingly good. Problems and failures can occur, most notably when the aperture function is very smooth, or is unknown. Worse yet, there are situations where a convolution framework is simply not general enough, such as when blurriness or aperture size varies across a single image. Something beyond deconvolution would be required; mathematicians and computer scientists are hard at work on this and related signal recovery and imaging problems.

Convolution is in the sun's dappled shadows on the grass, in the vocal reverb you use to add pizazz to your demo CD, in the way heat spreads out on your dining table after you have placed your hot coffee mug on it, and in the multiplication of binary numbers happening every nanosecond in every computer on the planet.[11] Mathematics is woven through the world around us . . . if only we choose to bring it into focus.

[10] If the hat ˆ symbol signifies the *Fourier transform* (frequency picture) of a signal, then what I have just explained is known as the *convolution theorem*: $\widehat{f * g} = \hat{f}\hat{g}$.

[11] The first two of these examples are similar to the cases I described earlier; the latter two are more sophisticated, yet convolutions nonetheless.

The author thanks Liz Canner and Wendy Hui Kyong Chun for fruitful discussions.

References

[1] S. Crutchfield, *Joy of Convolution*, web applet (accessed 10/5/09), `http://www.jhu.edu/signals/convolve/`

[2] S. J. Orfanidis, *Introduction to Signal Processing* (Prentice Hall, Englewood Cliffs, NJ, 1996)

[3] J. L. Starck and F. Murtagh, *Astronomical Image and Data Analysis* (Springer, 2006)

[4] M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging* (Taylor & Francis, 1998)