Computer Physics Communications 224 (2018) 209–221

Contents lists available at ScienceDirect

Computer Physics Communications

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

Spectral algorithms for multiple scale localized eigenfunctions in infinitely long, slightly bent quantum waveguides

John P. Boyd [a,*], Paolo Amore [b], Francisco M. Fernández [c]

[a] Department of Climate and Space Sciences and Engineering, University of Michigan, 2455 Hayward Avenue, Ann Arbor MI 48109, United States
[b] Facultad de Ciencias, CUICBAS, Universidad de Colima, Bernal Díaz del Castillo 340, Colima, Colima, Mexico
[c] NIFTA (UNLP, CCT La Plata-CONICET), Division Quimica Teorica, Blvd. 113 S/N, Sucursal 4, Casilla de Correo 16, 1900 La Plata, Argentina

ARTICLE INFO

Article history:
Received 10 January 2017
Received in revised form 23 October 2017
Accepted 25 October 2017
Available online 15 November 2017

Keywords:
Pseudospectral
Quantum waveguide

ABSTRACT

A "bent waveguide" in the sense used here is a small perturbation of a two-dimensional rectangular strip which is infinitely long in the down-channel direction and has a finite, constant width in the cross-channel coordinate. The goal is to calculate the smallest ("ground state") eigenvalue of the stationary Schrödinger equation which here is a two-dimensional Helmholtz equation, $\psi_{xx} + \psi_{yy} + E\psi = 0$ where E is the eigenvalue and homogeneous Dirichlet boundary conditions are imposed on the walls of the waveguide. Perturbation theory gives a good description when the "bending strength" parameter ϵ is small as described in our previous article (Amore et al., 2017) and other works cited therein. However, such series are asymptotic, and it is often impractical to calculate more than a handful of terms. It is therefore useful to develop numerical methods for the perturbed strip to cover intermediate ϵ where the perturbation series may be inaccurate and also to check the pertubation expansion when ϵ is small. The perturbation-induced change-in-eigenvalue, $\delta \equiv E(\epsilon) - E(0)$, is $O(\epsilon^2)$. We show that the computation becomes very challenging as $\epsilon \to 0$ because (i) the ground state eigenfunction varies on both $O(1)$ and $O(1/\epsilon)$ length scales and (ii) high accuracy is needed to compute several correct digits in δ, which is itself small compared to the eigenvalue E. The multiple length scales are not geographically separate, but rather are inextricably commingled in the neighborhood of the boundary deformation. We show that coordinate mapping and immersed boundary strategies both reduce the computational domain to the uniform strip, allowing application of pseudospectral methods on tensor product grids with tensor product basis functions. We compared different basis sets; Chebyshev polynomials are best in the cross-channel direction. However, sine functions generate rather accurate analytical approximations with just a single basis function.

In the down-channel coordinate, $X \in [-\infty, \infty]$, Fourier domain truncation using the change of coordinate $X = \sinh(Lt)$ is considerably more efficient than rational Chebyshev functions $\mathrm{TB}_n(X; L)$. All the spectral methods, however, yielded the required accuracy on a desktop computer.

Published by Elsevier B.V.

1. Introduction

As reviewed in our previous article [1], there is considerable interest in the localized ground state eigenfunctions that arise when an infinitely long, uniform width quantum waveguide is perturbed by a localized bulge in the wall or by a sharp bend as shown schematically in Fig. 1. Perturbation theory, as developed in our article and by other articles we cite, is a good option when the perturbation parameter is very small. However, it is still desirable to develop numerical methods that can compute the eigenvalues and eigenfunctions with spectral accuracy.

The numerical computations have two large challenges. The eigenfunctions for the uniform, unperturbed waveguide are independent of the down-channel coordinate $x \in [-\infty, \infty]$ and are sinusoids in the cross-channel coordinate y. However, when the perturbation is very small but has a length scale comparable to the width of the waveguide (the usual case), the ground state eigenfunction has two widely disparate length scales. One is the $O(1)$ length scale of the wall perturbation. The other is the $O(1/\epsilon)$ length scale of the slow decay of the eigenfunction in the down-channel direction.

This is one numerical challenge, but verification of perturbation theory is also hard because, to provide any useful information about the accuracy of the perturbative approximation, the numerical method must accurately calculate the tiny *difference* between the perturbed and unperturbed eigenvalues.

[*] Corresponding author.
E-mail addresses: jpboyd@umich.edu (J.P. Boyd), paolo.amore@gmail.com (P. Amore), fernande@quimica.unlp.edu.ar (F.M. Fernández).

https://doi.org/10.1016/j.cpc.2017.10.015
0010-4655/Published by Elsevier B.V.

Table 1
Notation.

(x, y)	Cartesian coordinates for the physical domain
E	Schroedinger equation eigenvalue
$B(x, y)$	Boundary function: its zero isoline is the boundary
D	Total degree of a polynomial (for x^{m+n}, $D = m + n$)
H	Pseudospectral discretization matrix
$\mathcal{H}$	Laplace operator
L	Map parameter for rational Chebyshev functions TB_n
$\mathfrak{L}$	Map parameter for the sinh-Fourier method
M	Number of basis functions in X
N	Number of basis functions in Y, the cross-channel coordinate
N_{total}	Total number of functions in the tensor product basis, MN
P	Number of interpolation points
T	Parameter for parametric specification of the upper boundary curve
$\mathfrak{W}$	Boundary of the region $X \in [\mathfrak{W}, \infty]$ where asymptotic analysis yields an explicit approximation to the eigenfunction
X	Computational coordinate in the "down-channel" direction, $X \in [-\pi, \pi]$
Y	Computational coordinate perpendicular to the walls ("cross-channel coordinate"), $Y \in [0, 1]$
δ	Change in the ground state eigenvalue due to perturbation
ϵ	Perturbation parameter; strength of domain deformation
ν	Mode number for the y-dependent factor (cross-channel factor)
$\psi(x, y)$	Wavefunction [the unknown in the Schroedinger equation]
$\chi_n(X)$	One-dimensional basis in the down-channel computational coordinate X
$\Theta_n(Y)$	One-dimensional basis in the cross-channel computational coordinate Y
$\sigma(x, y)$	Metric factor in the PDE induced by conformal mapping

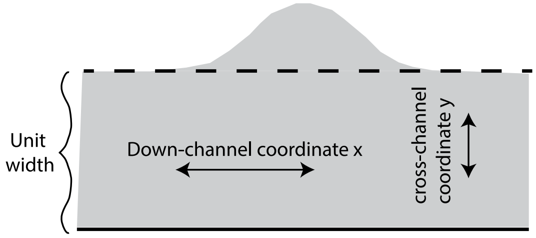

Fig. 1. Schematic of a bent waveguide. The unperturbed waveguide is a strip which is infinitely long in the "down-channel" coordinate x. The waveguides considered here are perturbed by bulges in one wall and are not actually bent. The jargon "bent waveguide" has become a shorthand for the class of "waveguides that are perturbations of a uniform-width, infinitely long rectangle by bend, bulging walls or other deformation that allows a bound state of finite energy".

Thus, a low order method is quite useless. All the algorithms applied here are spectrally accurate.

Spectral methods applied to a phenomenon with a single spatial scale are well understood as cataloged in [2–4]. However, applying spectral methods when there are multiple spatial scales is still an application on the research frontier. SIAM founded its *Journal of Multiscale Modeling and Simulation* not because multiple scales are passé, but because multiple scale methods are the frontier.

The eigenproblem is

$$\psi_{xx} + \psi_{yy} + E\psi = 0, \qquad \psi(x, y) = 0 \ \forall (x, y) \in \partial\Omega \tag{1}$$

where E is the eigenvalue and we impose homogeneous Dirichlet boundary conditions on the walls of the waveguide $\partial\Omega$. Important symbols are listed in Table 1. Note that subscripts with respect to a coordinate denote partial differentiation with respect to that coordinate, a convention employed throughout this article.

2. Strategies for an asymmetric channel

Many strategies have been applied to complicated domains, but we concentrate on approaches that are well-suited to perturbed rectangular domains: conformal mapping and the immersed boundary method. Both transform the waveguide from the "physical coordinates" (x, y) to computational coordinates (X, Y) where the domain is a channel of uniform unit width in the cross-channel coordinate Y, but extending indefinitely in the down-channel X coordinate. Thus, like the unperturbed domain, the computational domain in the coordinates (X, Y) is a rectangle.

We shall now briefly describe each strategy.

In the conformal mapping method, the computational domain is the infinite, uniform width channel in the coordinates (X, Y). This is the image of a non-rectangular domain under a conformal mapping. Because the mapping is conformal, the coordinate transformation merely multiplies the eigenvalue term in the Schrödinger equation by the metric factor. The "crowding" or "Geneva Effect", that is, a highly nonuniform grid, is fatal to most efforts at grid generation by conformal mapping [5]. Here, crowding is not an issue because the map is a *small perturbation* of the *identity transformation*. The conformal mapping used here is given by a simple analytical expression. However, an explicit conformal map may not be available. What then? One option is to calculate conformal maps using perturbation theory as in [6,7]. Another is to apply PDE-solvers that do not require a conformal mapping as elaborated below.

The key idea of an "immersed boundary" method is to embed the physical domain inside a computational domain which, in this case, is an infinite strip of uniform width [8–10]. Boundary conditions are imposed by Krylov's method [6]. That is, if the boundary is specified implicitly as the union of the zero isolines of a function $B(x, y)$, then homogeneous Dirichlet boundary conditions are enforced by writing the approximation as

$$\psi(x, y) = B(x, y)v(x, y) \tag{2}$$

where $v(x, y)$ is an unconstrained sum of tensor product basis functions.

An alternative to these approaches is to map the perturbed waveguide into the channel of uniform unit width using a nonconformal mapping. The bad news is that the metric factors will be numerous, significantly extending the debugging time. However, relaxing conformality opens up a vast spectrum of grid generation techniques for future studies.

3. A typical asymmetric channel

In the rest of the article, we focus on an example that is representative of a broad class of bent waveguides – more accurately described as "bulging waveguides" – in which the perturbation is a distortion of the shape of the upper boundary, $y = 1$. We shall concentrate on a particular distortion, but the methods applied are general. In our case, the perturbation is generated by the conformal map

$$F(z) = z + \epsilon \tanh(z), \qquad \epsilon \ll 1. \tag{3}$$

The perturbed channel has a lower boundary at $y = 0$, a straight line identical with the unperturbed channel, but the upper boundary is altered from $y = 1$ to a curve described in parametric form by

$$x = T + \epsilon \frac{\sinh(2T)}{\cos(2) + \cosh(2T)} \tag{4}$$

$$y = 1 + \epsilon \frac{\sin(2)}{\cos(2) + \cosh(2T)}, \qquad T \in [-\infty, \infty]. \tag{5}$$

The bulge is symmetric about $x = 0$, so the ground state eigenmode has the same property. We exploit this symmetry by using basis sets restricted to symmetric functions only.

When the bulging channel in (x, y) is conformally mapped to the straight channel in the computational coordinates (X, Y), the transformed problem in the straight channel becomes

$$\begin{aligned} &\psi_{XX} + \psi_{YY} + (1 + \sigma(X, Y))E\psi = 0, \\ &\quad \psi(x, 0) = \psi(x, 1) = 0 \end{aligned} \tag{6}$$

where σ is a known function. (Note that (X, Y) are the computational coordinates on the unbounded rectangle.) To calculate the metric factor $\sigma(X, Y)$, it is necessary to convert the parametric specification of the boundary curve to an implicit form as described in Appendix.

Any conformal mapping yields the simple transformed form $\psi_{XX} + \psi_{YY} + (1 + \sigma(X, Y))E\psi = 0$; for the particular case of the boundary described by the parametric form above,

$$\begin{aligned} \sigma(X, Y) = 4\epsilon &\frac{(\cosh(2X)\cos(2Y) + 1)}{(\cosh(2X) + \cos(2Y))^2} \\ &+ 4\epsilon^2 \frac{1}{(\cosh(2X) + \cos(2Y))^2}. \end{aligned} \tag{7}$$

This function σdecays exponentially fast as $|x| \to \infty$ and therefore the PDE degenerates to a constant coefficient equation for $X > \mathfrak{W}$ for some sufficiently large $\mathfrak{W}$, allowing an asymptotic analysis as $X \to \infty$ as given in the next section.

The conformal mapping is useful for theory and numerical methods. The immersed boundary method allows numerical solutions to alternatively be calculated on the original domain without a change of coordinate.

4. The challenge of multiple spatial scales: asymptotic analysis

The bent waveguide problem is challenging because any successful method must resolve wildly disparate length scales in the down-channel direction. After the perturbed waveguide is conformally mapped to a uniform strip in the computational coordinates (X, Y), an arbitrary solution can always be expanded into a Fourier sine series in Y with X-dependent coefficients. Fig. 2 shows that the lowest sine coefficient, $b_1(X)$, decays very slowly compared to the higher harmonics, $b_n(X)$, $n \geq 2$. The ratio of decay scales only worsens as $\epsilon \to 0$.

To understand why there is this large ratio of scales, we proceed in several steps. First, solve the unperturbed problem

$$\psi_{xx} + \psi_{yy} + E\psi = 0, \qquad \psi(x, 0) = \psi(x, 1) = 0. \tag{8}$$

When the domain is the unperturbed infinite strip, the eigenfunctions and eigenvalues are

$$\psi_n = \sin(n\pi\, y), \qquad n = 1, 2, \ldots \qquad E_n = n^2\pi^2. \tag{9}$$

These eigenfunctions are independent of x. This implies that the modes are unphysical in the sense that the integral of the square of the eigenfunction over the entire infinite strip, which is the usual L_2 norm of the mode, is infinite. The perturbation to the geometry creates an eigenmode which has finite norm and energy.

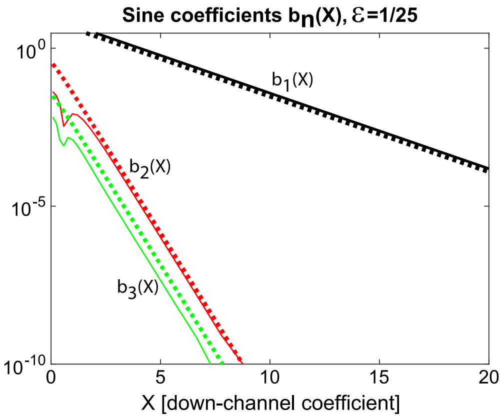

Fig. 2. The solid curves are the lowest three sine coefficients for $\epsilon = 1/25$ in the expansion of the solution to our standard example as $\psi(X, Y) = \sum_{n=1} b_n(X) \sin(n\pi Y)$. The dotted curves are the asymptotic approximations derived in the text as (21).

Next, conformally map the perturbed waveguide to the uniform width infinite strip in the computational coordinates (X, Y). The partial differential equation transforms to

$$\begin{aligned} &\psi_{XX} + \psi_{YY} + (1 + \sigma(X, Y))E\psi = 0, \\ &\quad \psi(x, 0) = \psi(x, 1) = 0 \end{aligned} \tag{10}$$

where σ is uniquely determined by the choice of the waveguide perturbation and thus of the conformal mapping.

For now only qualitative properties are relevant. One such property is that the magnitude of σ is $O(\epsilon)$ where $\epsilon \ll 1$ is the perturbation parameter. Another important property is that $\sigma(X, Y)$ decays exponentially fast as $|X| \to \infty$.

For all X and Y, perturbed eigenmodes can be expanded as a Fourier sine series in the cross-channel coordinate, $u(X, Y) = \sum_{n=1}^{\infty} b_n(X) \sin(n\pi Y)$. Substitution of the sine series into the partial differential equation followed by Galerkin's method gives the set of ordinary differential equations in X

$$b_{n,XX} + (E - n^2\pi^2)b_n + \sum_{m=1}^{\infty} b_m(X) I_{mn}(X) = 0 \tag{11}$$

$$I_{mn} = 2\int_0^1 dY\, \sin(nY)\sigma(X, Y)\sin(mY) \tag{12}$$

$$b_n(X) \to 0 \qquad \text{as}\, X \to \pm\infty \tag{13}$$

$$\begin{aligned} \sigma(X, Y) = 4\epsilon &\frac{(\cosh(2X)\cos(2Y) + 1)}{(\cosh(2X) + \cos(2Y))^2} \\ &+ 4\epsilon^2 \frac{1}{(\cosh(2X) + \cos(2Y))^2} \end{aligned} \tag{14}$$

$$\sim 8\epsilon\, \cos(2Y)\exp(-2X), \qquad |X| \gg 1. \tag{15}$$

For large X, the ODE system simplifies to

$$b_{1,XX} - \hat{\delta}\epsilon^2 b_1 = O(\epsilon) \tag{16}$$

$$b_{2,XX} - 3\pi^2 b_2 = -I_{12} b_1 \tag{17}$$

$$b_{3,XX} - 8\pi^2 b_3 = -I_{13} b_1 \tag{18}$$

where

$$I_{12} \sim -\epsilon\, 8\, \frac{\pi\, (\cos(2) - 1)\, e^{-2X}}{\pi^2 - 1}, \qquad X \gg 1 \tag{19}$$

$$I_{13} \sim \epsilon\, 48\, \frac{\pi\, (\cos(2) + 1)\, e^{-2X}}{9\,\pi^2 - 4}. \tag{20}$$

Then for $X \geq \mathfrak{W}$,

$$b_1(X) \sim b_1(\mathfrak{W}) \exp\left(-\sqrt{\hat{\delta}}\,\epsilon(X-\mathfrak{W})\right) \tag{21}$$

$$b_2 \sim -\frac{8}{3}\,\frac{\pi\,\epsilon\,(\cos(2)-1)}{\pi^2-1} b_1(X)\,\exp(-2X) \tag{22}$$

$$b_3 \sim 6\,\frac{\pi\,\epsilon\,(\cos(2)+1)}{9\pi^2-4} b_1(X)\,\exp(-2X). \tag{23}$$

Fig. 2 shows excellent agreement between the numerical curves and the asymptotic predictions (21).

The ratio of the decay scale for the fundamental b_1 and the higher harmonics $b_n(X)$ is thus

$$r_{scales} \sim \frac{2}{\sqrt{\hat{\delta}}\,\epsilon} \tag{24}$$

$$\sim \frac{1}{\pi^2\,\epsilon} \tag{25}$$

where the last line substitutes the lowest order perturbative approximation for $\hat{\delta}$, $4\pi^4$, from [1]. When $\epsilon \ll 1/10$, the numerical method must resolve two very disparate length scales.

Denoting the number of basis functions in X and Y by M and N, the two-scale dependence demands

$$M \gg N. \tag{26}$$

5. The pseudospectral method

5.1. The general formalism

The eigenfunction is approximated by a series of basis functions $\phi_n(X, Y)$:

$$\psi(X,Y) = \sum_{n=1}^{N_{total}} a_n\,\phi_n(X,Y) \tag{27}$$

where we assume that the basis functions individually satisfy the boundary conditions. Let (X_j, Y_j) denote a set of N_{total} interpolation points. Let the differential eigenproblem be

$$\mathcal{H}\psi + E\mathcal{V}(X,Y)\psi = 0 \tag{28}$$

where $\mathcal{H}$ is an elliptic operator [here, the Laplacian operator, $\mathcal{H} = \partial_{XX} + \partial_{YY}$] and $\mathcal{V}(X,Y)$ is a function smooth on the computational domain but otherwise arbitrary and E is the eigenvalue. The pseudospectral discretization of this is the generalized matrix eigenproblem

$$\mathbf{Ha} = E\mathbf{Va} \tag{29}$$

where the elements of the vector $\mathbf{a}$ are the spectral coefficients a_n and the matrix elements are

$$H_{jn} = \mathcal{H}(\phi_n)|_{X=X_j, Y=Y_j} \tag{30}$$

$$V_{jn} = \mathcal{V}(X_j, Y_j)\phi_n(X_j, Y_j). \tag{31}$$

The pseudospectral algorithm (and Galerkin methods, applied later) is discussed at length in the book [2], especially Chapters 3 to 6.

In our applications, the basis is a tensor product basis:

$$\phi_k(X,Y) = \chi_m(X)\Theta_n(Y), \qquad m = 1, 2, \ldots, M; \quad n = 1, 2, \ldots, N \tag{32}$$

where the total number of basis functions is $N_{total} = MN$. The grid is a tensor product grid in the sense that the N_{total} points are all possible combinations, MN in number, of the one-dimensional grids with their indices i and j varying independently. That is, the interpolation points are (X_i, Y_j), $i = 1, 2, \ldots, M$, $j = 1, 2, \ldots, N$ where the X_i are the standard grid associated with the basis functions $\chi_m(X)$ and the Y_j are the canonical one-dimensional grid points appropriate for the basis $\Theta_n(Y)$.

5.2. An overview of our example

We applied a variety of methods to calculate the ground state eigenvalue as displayed in Table 2. Results are expressed in terms of the difference between the perturbed and unperturbed eigenvalues

$$\delta \equiv \pi^2 - E(\epsilon). \tag{33}$$

To recall from previous sections, the channel is an infinite strip of uniform width perturbed so that one wall is deformed to the shape described in parametric form by $x = T + \epsilon\,\sinh(2T)/(\cos(2)+\cosh(2T))$, $y = 1 + \epsilon\,\sin(2)/(\cos(2)+\cosh(2T))$ where T is the parameter. The bulge is symmetric about $x = 0$, so the ground state eigenmode has the same property. We exploit this by using basis functions symmetric about the origin in the down-channel coordinate x.

6. Cross-channel basis functions: Chebyshev polynomials versus sine functions

To enforce homogeneous boundary conditions on the walls, we employ basis functions which *individually* vanish at each wall, either pairs of Chebyshev polynomials or $\sin(\pi n Y)$.

For both the conformal mapping and immersed boundary methods, the computational domain is an infinite, uniform width channel in the coordinates (X, Y). However, the width is different from one for the immersed boundary approach, so we shall derive the Chebyshev polynomial formalism for the interval $Y \in [A, B]$ where A and B are arbitrary.

The Chebyshev polynomials are defined by the recurrence relation $T_{n+1}(z) = 2zT_n(z) - T_{n-1}(z)$ with the starting values $T_1(z) = z$ and $T_0 \equiv 1$. If $z = \cos(t)$, this recurrence becomes the trigonometric identity $\cos([n+1]t) = 2\cos(t)\cos(nt) - \cos([n-1]t)$. Thus $T_n(\cos(t)) = \cos(nt)$ where $T_n(z)$ is the Chebyshev polynomial of degree n. It follows that the Chebyshev polynomials on the interval $Y \in [A, B]$ are the images of a Fourier cosine basis under the change of coordinate

$$T_n(z) = \cos(nt) \tag{34}$$

$$z = (2Y - (B+A))/(B-A) \tag{35}$$

$$t = \arccos(z) = \arccos\left(\frac{2Y-(B+A)}{B-A}\right) \tag{36}$$

where $t \in [0, \pi]$ is the "trigonometric coordinate", $z \in [-1, 1]$ is the Chebyshev polynomial argument and $Y \in [A, B]$ is the cross-channel coordinate.

Derivatives can be calculated in the trigonometric coordinate and then transformed, by applying the chain rule, to the Y derivatives we actually need. The canonical interpolation points in Y are the images of a uniform grid in the trigonometric coordinate. Because we will use basis functions that satisfy the boundary conditions, we construct a Chebyshev–Lobatto grid with $(N+2)$ points and then omit the endpoints

$$Y_k \equiv \frac{B-A}{2}\cos\left(\pi\frac{k}{N+1}\right) + \frac{B+A}{2}, \qquad k = 1, 2, \ldots, N. \tag{37}$$

To satisfy the boundary conditions, we employ basis functions that *individually* satisfy the homogeneous conditions on the walls.

Table 2
Computed eigenvalue changes δ in the ground state for various basis sets in various truncations where $\delta = \pi^2 - E$. M and N are the number of basis functions in X and Y. L and $\mathfrak{L}$ are the TB and SINH "map parameters". "Conformal" indicates that the domain is conformally mapped to the uniform width rectangle. "TB" means that rational Chebyshev functions of even degree were the basis in X; "SINH+COS" implies that the X coordinate was transformed by the hyperbolic sine mapping followed by Fourier cosine domain truncation. "TN" denotes that Chebyshev polynomials were the cross-channel basis; "SINE" denotes $\sin(n\pi Y)$ was used instead. "GAL" means that Galerkin's method replaced the pseudospectral discretization used for all other entries.

Method and basis	L	M	N	$-\delta$	rel. error
$\epsilon = 1/10{,}000$					
Conf TB & TN	$L = 100$	$M = 100$	$N = 20$	−2.240655e−08	−0.42
Conf TB & TN	$L = 100$	$M = 300$	$N = 20$	−3.888793e−08	−0.0016
Conf TB & TN	$L = 100$	$M = 500$	$N = 20$	−3.895123e−08	−1.7e−05
Conf TB & TN	$L = 100$	$M = 700$	$N = 20$	−3.895156e−08	−8.7e−06
Conf TB & TN	$L = 100$	$M = 800$	$N = 20$	−3.895132e−08	−1.5e−05
SINH-COS & TN	$L = 4$	$M = 50$	$N = 24$	−3.89518983831e−08	−5.6e−08
SINH-COS & TN	$L = 4$	$M = 100$	$N = 24$	−3.89519005746e−08	3e−13
SINH-COS & TN	$L = 4$	$M = 200$	$N = 24$	−3.89519005746e−08	3.0e−13
SINH-COS & TN	$L = 6$	$M = 50$	$N = 24$	−3.89511204033e−08	2.0e−5
$\epsilon = 1/1000$ errors computed using −3.884668E−6 as "exact"					
Conf TB & TN	$L = 10$	$M = 400$	$N = 20$	−3.88417526e−06	1.3e−4
Conf TB & TN	$L = 15$	$M = 400$	$N = 20$	−3.88464267e−06	6.4e−6
Conf TB & TN	$L = 20$	$M = 400$	$N = 20$	−3.88464140e−06	7.0e−6
Conf TB & TN	$L = 25$	$M = 200$	$N = 16$	−3.88463041e−06	1.0e−5
Conf TB & TN	$L = 30$	$M = 200$	$N = 16$	−3.88463476e−06	8.5e−6
Conf TB & TN	$L = 30$	$M = 400$	$N = 20$	−3.88462062e−06	1.2e−5
Conf, SINH-COS & TN	$\mathfrak{L}$=5	$M = 200$	$N = 20$	−3.88466781e−06	4.9e−8
Conf SINH-COS & TN	$\mathfrak{L}$=6	$M = 100$	$N = 20$	−3.8846675e−06	1.3e−7
Conf SINH-COS & SINE	$\mathfrak{L}$=6	$M = 100$	$N = 16$	−3.88466916e−06	3.1e−7
Conf, SINH-COS & TN	$\mathfrak{L}$=8	$M = 100$	$N = 20$	−3.88466793e−06	1.8e−8
Conf, SINH-COS & TN	$\mathfrak{L}$=8	$M = 200$	$N = 25$	−3.88466693e−06	2.8e−7
Conf, SINH-COS & SINE	$\mathfrak{L}$=5	$M = 50$	GAL 1	−3.8846500e−06	4.6e−6
Conf, SINH-COS & SINE	$\mathfrak{L}$=5	$M = 100$	GAL 1	−3.8846573e−06	2.8e−6
$\epsilon = 1/100$					
SINH-COS & TN	$L = 6$	$M = 50$	$N = 24$	−0.000378313474209	−2e−05
SINH-COS & TN	$L = 6$	$M = 100$	$N = 24$	−0.000378320991591	−2.3e−10
SINH-COS & TN	$L = 6$	$M = 200$	$N = 24$	−0.000378320991676	0
Conf TB & TN	$L = 10$	$M = 10$	$N = 20$	−0.000216974425828	−0.4265
Conf TB & TN	$L = 10$	$M = 20$	$N = 20$	−0.000370256739341	−0.02132
Conf TB & TN	$L = 10$	$M = 30$	$N = 20$	−0.000377709097205	−0.001617
Conf TB & TN	$L = 10$	$M = 50$	$N = 20$	−0.000378314016604	−1.844e−05
Conf TB & TN	$L = 10$	$M = 60$	$N = 20$	−0.000378321481908	1.296e−06
Conf TB & TN	$L = 5$	$M = 100$	$N = 20$	−0.000378317262797	−9.856e−06
Conf TB & TN	$L = 5$	$M = 200$	$N = 20$	−0.000378316895749	−1.083e−05
Conf TB & TN	$L = 10$	$M = 100$	$N = 20$	−0.00037831654112	−1.176e−05
Conf TB & TN	$L = 20$	$M = 100$	$N = 20$	−0.000378315252239	−1.517e−05
Conf TB & TN	$L = 50$	$M = 100$	$N = 20$	−0.000368892634683	−0.02492
Conf TB & TN	$L = 50$	$M = 200$	$N = 20$	−0.000378281933729	−0.0001032
Conf TB & TN	$L = 50$	$M = 300$	$N = 20$	−0.000378317513349	−9.194e−06
Conf TB & TN	$L = 100$	$M = 100$	$N = 20$	−0.000232233699493	−0.3861
Conf TB & TN	$L = 100$	$M = 300$	$N = 20$	−0.000377766505323	−0.001466
Conf TB & TN	$L = 100$	$M = 500$	$N = 20$	−0.000378314775741	−1.643e−05

Our choices are either $\sin(n\pi Y)$ or the *difference* of two Chebyshev polynomials of the same parity:

$$\begin{aligned} \Phi_n &\equiv T_{n+1}(z(Y)) - T_{n-1}(z(Y)) \qquad \Leftrightarrow \\ \Phi_n &\equiv \cos([n+1]t) - \cos([n-1]t). \end{aligned} \tag{38}$$

Then, with $W = 2/(B - A)$, the chain rule allows us to express derivatives with respect to the cross-channel coordinate Y in terms of the following derivatives of the basis functions with respect to the trigonometric coordinate t:

$$\Phi_{n,t} = -(n+1)\sin([n+1]t) - (n-1)\sin([n-1]t) \tag{39}$$

$$\Phi_{n,tt} = -(n+1)^2\cos([n+1]t) - (n-1)^2\cos([n-1]t) \tag{40}$$

$$\phi_{n,Y} = -\Phi_{n,t}\frac{W}{\sin(t)} \tag{41}$$

$$\phi_{n,YY} = \left\{\sin(t)\Phi_{n,tt} - \cos(t)\Phi_{n,t}\right\}\frac{W^2}{\sin^3(t)}. \tag{42}$$

The alternative basis is

$$\sin(n\pi Y), \qquad n = 1, 2, \ldots \qquad Y \in [0, 1]. \tag{43}$$

Fourier sines are simple, individually satisfy the boundary conditions, and are the cross-channel-dependent factors of the unperturbed eigenmodes. Therefore, Fourier functions have been widely used in eigenvalue computations and boundary value solving even in coordinates where periodicity is lacking, as in the cross-channel coordinate Y here. Theory predicts that asymptotically for large degree, a Fourier basis will converge at a rate proportional to an inverse power of degree rather than exponentially.

Fig. 3 shows the spectral coefficients for each basis. The first fifty are the Fourier cosine coefficients $b^{sin}_{p,1}$, $p = 1 \ldots 50$ where

$$\psi(x,y) = \sum_{p=1}\sum_{n=1} b^{sin}_{p,n}\cos(p\,t(X))\sin(\pi n y) \tag{44}$$

for $n = 1$ [black] or of $b^{Cheb}_{p,n}$ in

$$\psi(x,y) = \sum_{p=1}\sum_{n=1} b^{Cheb}_{p,n}\cos(p\,t(X))(T_{n+2}(Y) - T_n(Y)) \tag{45}$$

for $n = 1$ [red]; the two curves are almost indistinguishable in the leftmost quarter of the diagram. The Chebyshev coefficients $b^{Cheb}_{p,n}$ for $n = 2, 3$ are several orders of magnitude larger than the

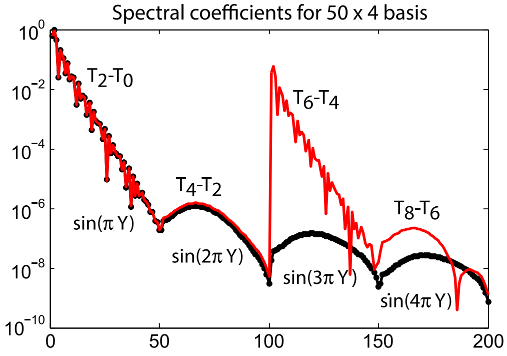

Fig. 3. Comparison of spectral coefficients for two different cross-channel basis sets. The spectral coefficients for all 200 basis functions are shown versus a single index running from 1 to 200 over the entire two-dimensional basis for $\epsilon = 10^{-4}$. The red curve employed four Chebyshev basis functions, each of which is the difference between two Chebyshev polynomials. The black curves with disks connect the coefficients when the cross-channel basis functions are $\sin(\pi\, n\, Y/L)$, which are the cross-channel factors of the unperturbed eigenmode. The Chebyshev polynomial basis predicts $\delta = 0.0023$ which is wrong even as to order of magnitude whereas the sine basis yields $\delta = 0.334620 \times 10^{-6}$ which contains five correct digits. Both curves used fifty Fourier sinh-mapped cosines in the down-channel direction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

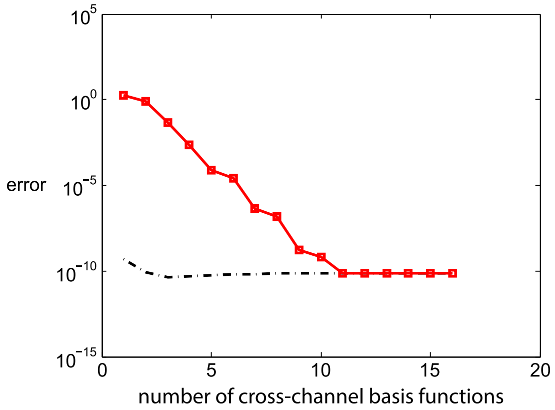

Fig. 4. Errors in δ, the correction to the ground state eigenvalue, for the asymmetric waveguide with $\epsilon = 10^{-4}$ as calculated using $M = 50$ Fourier cosines $\cos(jt(X))$ in the down-channel direction X where $t = (1/6)\text{arcsinh}(X)$ and either a sine basis in Y [black dot–dash curve] or a Chebyshev polynomial basis in Y with N basis functions where in either case N is the horizontal axis. The errors in the pseudospectral Chebyshev polynomial method [red curve with squares] decay exponentially fast to "saturate" when the number of basis functions reaches 11. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

coefficients $b^{sin}_{p,n}$. However, the X Fourier series for $b^{Cheb}_{p,n}$, $n > 1$ decrease much faster with increasing degree p in X than do their counterparts for $b^{sin}_{p,n}(X)$, $n > 1$.

6.1. One-term Galerkin sine approximation

The spectral coefficients show that the expansion in the cross-channel coordinate Y is dominated by the lowest cross-channel mode. (Collocation with many basis functions is applied in the down-channel coordinate X.) This suggests that it is useful to look at approximations that have few cross-channel modes, perhaps even just a single mode.

The pseudospectral method is somewhat easier to program than the Galerkin discretization because functions are *evaluated* instead of *integrated*. The Galerkin method is almost always more accurate, but the accuracy ratio is usually only a factor of two or three. When the basis truncation is large, this accuracy difference is usually insignificant because the pseudospectral method with $(N + k)$ basis functions is as accurate as the Galerkin method with N basis functions where $k \ll N$. When the basis is drastically truncated, however, Galerkin's higher accuracy is often worth the bother.

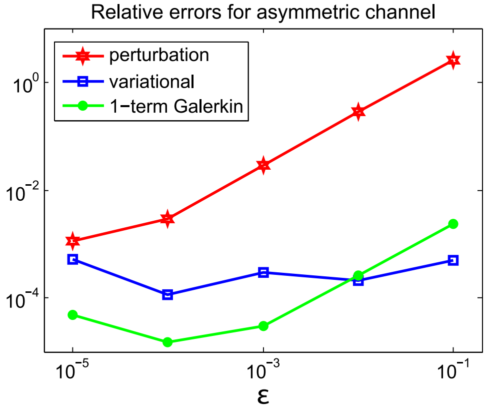

Fig. 5. Relative errors in δ for the asymmetric channel versus the perturbation parameter ϵ. The relative error is defined as $(\delta(method) - \delta_{exact})/\delta_{exact}$ where "method" is the perturbation theory [1], $\delta(perturbation) = 4\pi^4\epsilon^2$, "variational" (from the same source) or "1-term sine-Galerkin". The "exact" answer, used to compute errors, is defined as the best pseudospectral approximation for a given ϵ.

A one-term Galerkin approximation to the partial differential equation with a single sine function as the cross-channel basis, that is, an approximation of the form $\psi(X, Y) = b_1(X)\sin(\pi Y)$, generates the ODE eigenproblem in the unbounded, down-channel coordinate X:

$$b_{1,XX} - \pi^2 b_1 + E\, q(X) b_1 = 0, \qquad b_1(X) \text{ bounded as } |X| \to \infty \tag{46}$$

$$q(X) = 2\int_0^1 dY \sin^2(\pi Y)\{1 + \sigma(X, Y)\}\,. \tag{47}$$

(This ODE is solved by a pseudospectral algorithm in X with M sufficiently large so that all errors in this subsection are dominated by the coarseness of the Y discretization.)

The inferiority of the Chebyshev basis to sines for ten or fewer basis functions in the cross-channel coordinate Y is clearly illustrated by Fig. 4. Table 3 reiterates this by comparing the accuracy of various Chebyshev truncations with the single sine-Galerkin approximation. Even ten Chebyshev basis functions are less accurate than a single sine function.

Fig. 5 shows that even with just a single cross-channel basis function, the Galerkin approximation is better than or equal to the best variational estimates from [1] for all ϵ except $\epsilon = 1/10$.

For larger N, however, the Chebyshev polynomial basis is always better because of its (much) faster rate of convergence. (Note the tiny errors for $N = 11$ and $N = 12$ in the table.) When moderate accuracy is acceptable, a one-term sine Galerkin is more accurate and simpler than any Chebyshev approximation of comparable simplicity and number of basis functions, but the Chebyshev basis is always *asymptotically superior* in the limit $N \to \infty$.

7. Rational Chebyshev functions vs. cosines-with-sinh-mapping as the down-channel basis

7.1. Rational Chebyshev functions

We quote some facts and definitions collected in [2]. The rational Chebyshev functions $\text{TB}_n(X; L)$ are a Fourier cosine series under

Table 3
Relative error in computing δ, the correction to the ground state eigenvalue, for $\epsilon = 1/10,000$ for various numbers of Chebyshev polynomial basis functions. All calculations used 100 basis functions in X with the sinh-Fourier scheme with $\mathfrak{L}$. The relative error with $\sin(\pi Y)$ as the single cross-channel basis function and a Galerkin discretization is also shown.

N	relative error in δ	$\delta(N; \epsilon)$
1	4.8E05	−1.9
2	2.05E05	0.80
3	1.2E04	−0.046
4	596.6	−0.00232
5	20.3	7.5E−05
6	6.9	2.30E−05
7	0.116302648871	−4.336467E−06
8	0.036	−4.026818E−06
9	0.000397	−3.88312E−06
10	1.40E−04	−3.88412E−06
11	5.56E−07	−3.884672E−06
12	1.40E−08	−3.88467005455E−06
1-sine Galerkin	1.47E−05	−3.884657E−06

Table 4
Rational Chebyshev functions for the infinite interval: $TB_n(X)$. (For map parameter $L = 1$.)

n	$TB_{2n}(X)$ [Symmetric about $X = 0$]
0	1
2	$(X^2 - 1)/(X^2 + 1)$
4	$(X^4 - 6X^2 + 1)/(X^2 + 1)^2$
6	$(X^6 - 15X^4 + 15X^2 - 1)/(X^2 + 1)^3$
8	$(X^8 - 28X^6 + 70X^4 - 28X^2 + 1)/(X^2 + 1)^4$
10	$(X^{10} - 45X^8 + 210X^6 - 210X^4 + 45X^2 + 1)/(X^2 + 1)^5$

the change of coordinate

$$\mathrm{TB}_n(X; L) = \cos(nt[X; L]) \tag{48}$$

$$t = \operatorname{arccot}(X/L) \tag{49}$$

$$X = L\cot(t) \tag{50}$$

where $t \in [0, \pi]$ is the "trigonometric coordinate" and $X \in [-\infty, \infty]$ is the computational coordinate. It is typical for the rational Chebyshev spectral coefficients to decrease at a "subgeometric" rate, that is, proportional to $\exp(-pn^r)$ for some constant p and an exponent $r < 1$. "Root-exponential" convergence is the most common case, errors falling as $\exp(-p\sqrt{n})$.

Because the eigenfunctions of the perturbed domain are symmetric with respect to the down-channel computational X, the basis can be restricted to rational Chebyshev functions of *even degree*. The lowest six symmetric rational Chebyshev functions are given in Table 4.

Derivatives can be evaluated by the chain rule. A Matlab function to evaluate the nth rational Chebyshev function and its two derivatives is

```
function [TB,TBX,TBXX]= TBbasis(X,LL,n)
% input: X \in [-oo,oo] is the coordinate
%           LL  > 0 is the user-choosble "map parameter".
%           n is the degree of the basis function.
% output:     TB is TB(X,LL)_n
%             TBX is the X-derivative of TB(X,LL)_n
%             TBXX is the second derivative
% step one:  convert to  ``trigonometric coordinate"  t
      t= acot(X/LL)
      TB = cos(n*t)
% step two:   apply the chain rule
C=cos(t);  S=sin(t);
 TBt= n*cos(n*t);   TBtt=-n*n*TB;  %  evaluate t-derivatives
TBX = - S * S * TBt / LL;   % convert to X-derivatives
TBXX = S*S*S* (S*TBtt + 2*C*TBt) / (LL*LL);
```

The interpolation points are the images of a uniform Fourier grid, as is optimum for a basis of sines and cosines,

$$y_j = L\cot\left(\pi\,\frac{2j-1}{2N}\right), \quad j = 1, 2, \ldots, N. \tag{51}$$

The remaining steps are as described in the general, abstract treatment of the pseudospectral method given earlier.

7.2. Fourier domain truncation with sinh change of coordinate

The alternative discussed here is domain truncation combined with a Fourier basis in the coordinate $t(X; \mathfrak{L})$ where t results from the change of coordinate

$$X = \sinh(\mathfrak{L}\,t) \tag{52}$$

where $\mathfrak{L}$ is a positive constant, the "sinh-scaling parameter". The domain is truncated to the images of $t = \pm\pi$ under the mapping

$$X \in [-X_{max}, X_{max}] \tag{53}$$

$$X_{max} = \sinh(\mathfrak{L}\pi) \approx (1/2)\exp(-\mathfrak{L}\pi). \tag{54}$$

This will be dubbed the "sinh-Fourier" method or, when the solution is symmetric about $X = 0$ and the basis can be restricted to cosines only, the 'sinh-cosine" method. The waveguides below are symmetric with respect to the origin, and so, too, is the ground state eigenfunction. This can exploited by (i) restricting the grid to $X \geq 0$ and (ii) using only even rational Chebyshev functions, $\mathrm{TB}_{2n}(X; L)$, or a Fourier cosine basis $\cos(n\,\operatorname{arcsinh}(X/\mathfrak{L}))$.

Theory asserts that as $n \to \infty$, the sinh-Fourier/sinh-cosine method should be superior to the rational basis expansion [11,12]. However, this is an *asymptotic* prediction, and sometimes true only for impractically huge n [13]; experimentation in the context of a specific application is the final arbiter for that application.

To analyze errors in the sinh-Fourier scheme, we need a slight generalization of the Fourier domain truncation theory of [14], which does not incorporate a change of coordinate as used here. The error theorem proved there can be generalized as follows.

Theorem 1 (*Fourier Truncation Error*). *Let the mapping function be*

$$X = f(t), \qquad t \in [-\pi, \pi]. \tag{55}$$

Let $X_{min} = \min_{t\in[-\pi,\pi]}(X(t))$ *and* $X_{max} = \max_{t\in[-\pi,\pi]}(X(t))$. *Approximate a function* $u(X)$ *by a Fourier series in* t *within the truncated domain and approximate the function by zero outside the truncated domain:*

$$u(X) = \begin{cases} a_0 + \sum_{n=1}^{\infty} a_n\,\cos(nf^{-1}(X)) + \sum_{n=1}^{\infty} b_n\,\sin(nf^{-1}(X)), \\ \qquad X \in [-X_{max}, X_{max}] \\ 0, \qquad |X| > X_{max} \end{cases}$$

where f^{-1} *is the inverse of* $f(t)$. *If* u *is not periodic in* t, *then as* $n \to \infty$,

$$a_n \sim \frac{(-1)^n\,\{f_t(\pi)u_X\,(f(\pi)) - f_t(-\pi)u_X\,(f(-\pi))\}}{\pi\,N^2} \tag{56}$$

$$b_n \sim \frac{(-1)^{n+1}\,\{u(f(\pi)) - u(f(-\pi))\}}{\pi\,N}. \tag{57}$$

Proof. Theorem 2 of [14] followed by application of the chain rule. ■

For the sinh-mapping, $X = \sinh(\mathfrak{L}\,t)$, specializing to problems symmetric about $X = 0$, the theorem gives

$$a_n \sim \frac{2}{\pi}\,(-1)^n\,\mathfrak{L}\,\cosh(\mathfrak{L}\pi)\,\frac{u_X\,(\sinh(\mathfrak{L}\pi))}{N^2}. \tag{58}$$

We can now return to the questions: Why is the sinh-Fourier method superior for sufficiently large n, and why is the rate of convergence of the rational Chebyshev series only "root-exponential" instead of the usual geometric convergence (proportional to $\exp(-pn)$)? A brief answer is that even when the domain is not explicitly truncated, any scheme with grid points is implicitly truncated to the "span" of the grid points, which we define as the interval between the largest and smallest grid points. It is absurd to suppose that any interpolation scheme can be accurate beyond the interval where interpolation constraints are imposed. This truncation, either explicit or implicit, to $|X| < X_{max}$, implies an error of $O(\max_{|X|>X_{max}}(|u(X)|))$.

To approximate a function like, say, $\mathrm{sech}(X)$, a numerical scheme must bow to two masters. One is the need for increasing the *density* of grid points to resolve finer and finer features of the target function $u(X)$. If h denotes the grid spacing near the origin where the target function is largest, the spectral series truncation error will fall like $\exp(-q/h)$ for some constant q, regardless of the basis. On a finite interval, h is inversely proportional to N, the truncation of the spectral series. However, on an unbounded interval, decreasing the grid spacing as $1/N$ does not converge the error to zero, but only to the domain truncation error, $O(u(X_{max}))$. The other master that the numerical scheme must honor is that to approximate a function over the entire real axis in the limit $N \rightarrow \infty$, the *span* of the grid points, that is, the interval spanned by the line segment connecting the smallest interpolation point to the largest, must increase steadily with finite but increasing N. If $u(X)$ is decreasing with X, then increasing the largest grid point X_{max} will decrease the domain truncation error, which is $O(u(X_{max}))$, or simply $|u(X_{max})|$ if $U(X)$ decays monotonically with increasing $|X|$. To converge to zero error, one must somehow increase both the span of the grid and the density of the grid points *simultaneously*. For $\mathrm{sech}(X)$, the best compromise is to weight these factors equally so that

$$h \sim constant\ /\sqrt{N} \qquad \& \qquad X_{max} \sim constant\ \sqrt{N}. \tag{59}$$

Then both $\exp(-q/h)$ and $\mathrm{sech}(X_{max}) \sim (1/2)\exp(-X_{max})$ decay as $N \rightarrow \infty$ with "root-exponential" convergence, that is, proportional to $\exp(-\mathrm{constant}\sqrt{N})$.

The sinh-Fourier method is better because for sufficiently high degree, X_{max} increases exponentially with $\mathfrak{L}$, and only a small increase of $\mathfrak{L}$ will produce a huge decrease in the domain truncation error. For a function like $u(X) = \mathrm{sech}(X)$, $u(X_{max}) \sim \exp(-\exp(L\pi))$. This exponential-of-exponential growth of the domain truncation implies that a slow logarithmic increase of $\mathfrak{L}$ with N such as $\mathfrak{L} = \log(q'N)/\pi$, where q' is a positive constant, will yield a domain truncation error that is exponential in N. The grid spacing in the neighborhood of $t = X = 0$ is proportional to $\mathfrak{L}/N$, so a logarithmically increasing $\mathfrak{L}$ implies that $\exp(-q/h) \sim \exp(-q\pi N)/\log(q'N)$, which is not quite geometric convergence, but falls short only by the logarithm. This rate of convergence is "quasi-geometric" [12,15].

If the desired error tolerance is an accuracy of 10^{-d}, that is, d decimal digits, then the best practical strategy is to choose the smallest map parameter $\mathcal{L}$ for the sinh-Fourier method such that the domain truncation error is less than 10^{-d}. Here, we simply experimented with various combinations of $\mathfrak{L}$, M and N, pursuing approximations whose accuracy was limited not by the choice of $\mathfrak{L}$ but rather by floating point round off error.

In the cross-channel direction, we found that increasing the number of Chebyshev polynomials in Y reached a plateau when $N = 12$ for the sinh-Fourier method. A similar flattening of the error-versus degree curve occurred at slightly larger truncations, $N = 14$ to $N = 16$, when the X basis is a set of rational Chebyshev functions. To assess the relative merits of the two contending infinite interval basis sets, we fixed N at these limiting values and then varied M and $\mathcal{L}$.

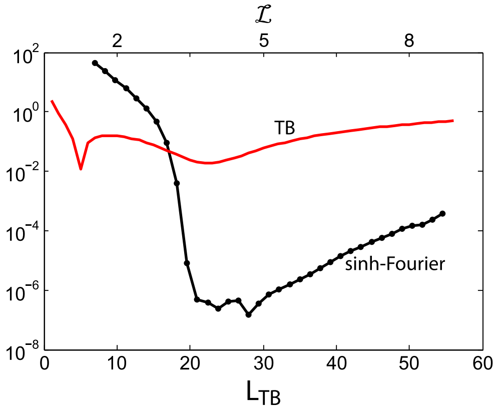

Fig. 6. Errors in the ground state eigenvalue as computed by the TB basis [red] and the sinh-domain-truncation-Fourier-cosine basis [black curve with disks]. $M = 50$ down-channel basis functions were combined with $N = 16$ Chebyshev polynomials as the cross-channel basis. Note that the map parameter values for the sinh-Fourier method have been scaled by multiplication by a factor of 7 to facilitate comparisons; the unscaled values are the labels on the upper axis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Compared to the sinh-Fourier computations, the rational TB calculations used double the down-channel truncation M, but the minimum error is still considerably larger for the rational basis as illustrated in Fig. 6. Boyd showed in [16] that the optimum map parameter L varies strongly with the truncation M; this is confirmed in Fig. 6 and also by the heavy dashed line which connects the optima for different M in Figs. 7 and 8.

The error contours for the rational Chebyshev basis are very smooth. Again as explained in [16], the error is a smooth analytic function of L for a given truncation M, so the error contours are smooth and flat around the best value of L where the derivative of the error with respect to L is zero. Choosing L to be double or half the best value for a given M increases the error only slightly from the error of the best choice of L. A plot of TB error versus L is U-shaped as seen in Fig. 6.

In contrast, errors in the sinh-Fourier scheme are the sum of two independent analytic functions. One is the domain truncation error E_D, which is the maximum of the solution $u(X)$ outside the truncated domain $|X| \leq \sinh(\mathfrak{L}\pi)$. The other is the series truncation error $E_S(M)$ which is the difference between the truncated Fourier series and $u(X)$. These two errors are independent analytic functions, one increasing exponentially with $\mathfrak{L}$ while the other decreases exponentially fast with the parameter. The sum of the two errors is V-shaped (slightly rounded by roundoff error in Fig. 6).

In the right contour plots of Figs. 7 and 8, the error is dominated by domain truncation error for $\mathfrak{L} < 3$, the error contours are vertical, and therefore independent of M because the series truncation errors (tiny) have little to do with the total error. For $\mathfrak{L} > 3$, however, the series truncation error dominates. The optimum choice of map parameter is slightly larger than 4 with little dependence on M.

Table 2 shows that it is possible to achieve high accuracy with either basis.

7.3. Hermite functions

For variety, we also applied a Hermite pseudospectral method with basis functions $\psi_m(\alpha y)$ [2,4] where α is a user-choosable scaling constant analogous to L or $\mathcal{L}$. Fig. 9 shows that a relative error of 10^{-6} is possible for relatively large ϵ.

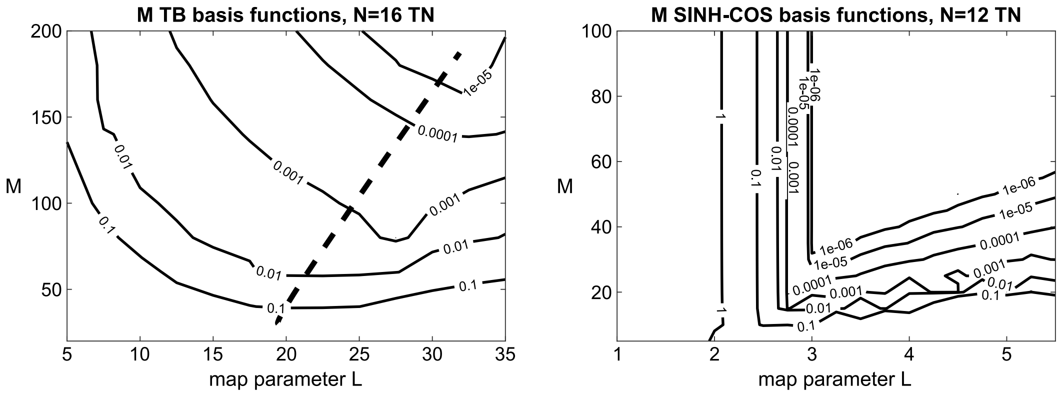

Fig. 7. Isolines of the relative error, $(\delta^{benchmark} - \delta^{numerical})/\delta^{benchmark}$, for the change in the ground state eigenvalue induced by the perturbation when $\epsilon = 1/10{,}000$. All calculations used Chebyshev polynomial basis functions in Y. The left contour plot shows the error of a basis of rational Chebyshev functions $\mathrm{TB}_m(X; L)$ where $m = 0, 2, 4, \ldots (2M - 2)$. The thick dashed line connects the values of the map parameter L which are optimal for various truncations M. Right panel: the same except for a down-channel basis of Fourier cosines with domain truncation after application of the sinh map with various values of the map parameter $\mathcal{L}$.

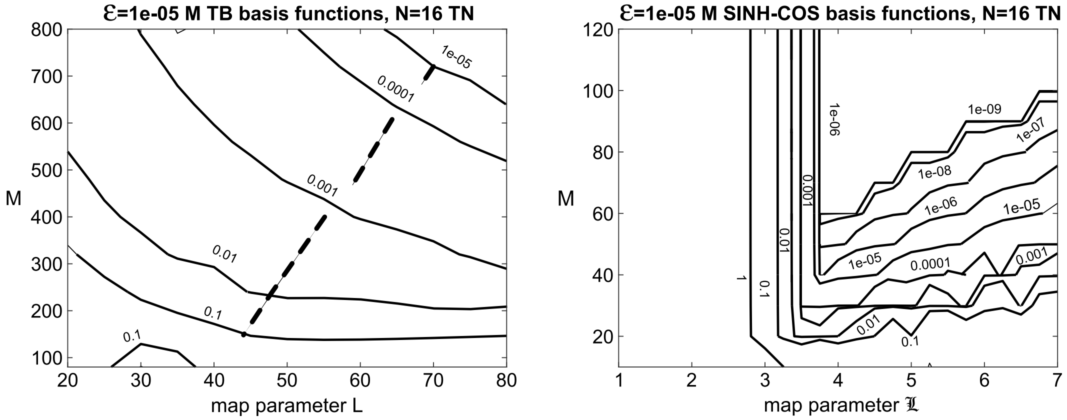

Fig. 8. Same as previous figure but with ϵ decreased to $\epsilon = 1/100{,}000$. Isolines of the relative error, $(\delta^{benchmark} - \delta^{numerical})/\delta^{benchmark}$. The left contour plot shows the error of a basis of rational Chebyshev functions $\mathrm{TB}_m(X; L)$ where $m = 0, 2, 4, \ldots (2M - 2)$. The thick dashed line connects the values of the map parameter L which are optimal for various truncations M. Right panel: the same except for a down-channel basis of Fourier cosines with domain truncation after application of the sinh map with various values of the map parameter $\mathcal{L}$.

Fig. 9 is similar to previous contour plots except that the basis in the unbounded, down-channel coordinate has been changed to Hermite functions. The figure shows an accuracy of 1 part in a million for the *correction* to the eigenvalue. However, no obvious improvement or advantages were seen in comparisons to the sinh-Fourier and rational Chebyshev methods. Indeed, a rerun of this case with all the same except for ϵ reduced to 1/10,000 yielded relative errors greater than one everywhere. We shall be content with this single figure.

7.4. Condition numbers

In recent years, as Chebyshev spectral methods are applied to more and more challenging problems, there have been concerns about the mild ill-conditioning of Chebyshev discretizations. This has inspired a revival of Petrov–Galerkin methods that, among other virtues, greatly reduce condition numbers [17,18].

However, the concentration of grid points near the boundaries, associated with very rapid oscillations near the endpoints, is a vice only of Chebyshev *polynomials*. The grids associated with the rational Chebyshev functions and the Fourier domain truncation with sinh change of coordinate, are roughly uniform with a grid spacing $O(1/N)$ near the center of the domain, and then become sparser and sparser as $|X| \to \infty$.

Table 5 shows that the condition number of the discretization matrix for the Laplace operator is a function almost entirely of N, the number of Chebyshev *polynomials* in the cross-channel direction. Both the rational Chebyshev basis and sinh-Fourier basis are well-conditioned. Condition number cannot be used to prefer one infinite interval discretization over another.

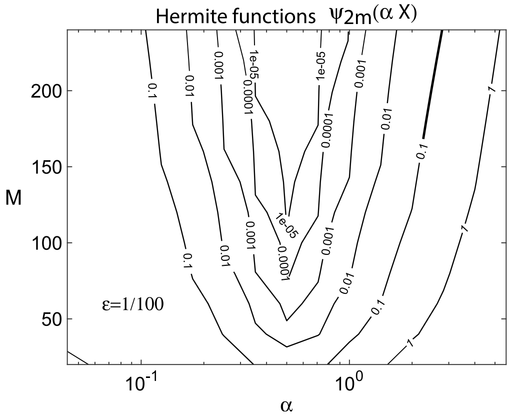

Fig. 9. Isolines of the relative error, $(\delta^{benchmark} - \delta^{numerical})/(\delta^{benchmark})$, for various numbers of functions when $\epsilon = 1/100$. All calculations used Hermite functions $\psi_m(\alpha X)$ where α is the scaling parameter, a user-choosable constant, and where $m = 0, 2, 4, \ldots (2M-2)$. The mth Hermite function is the product of $\exp(-[1/2]Y^2)$ with the Hermite polynomial of degree M.

Table 5
Condition Number of the Discretization of the Laplace Operator $\mathcal{H} = \partial_{XX} + \partial_{YY}$ Chebyshev polynomial basis functions were the cross-channel basis. M and N are the number of basis functions in the unbounded, down-channel coordinates X and in Y, respectively.

Condition number	M	N
Fourier cosine basis in X with sinh map with $\mathfrak{L} = 6$		
97.0	50	8
97.0	100	8
122.1	200	8
626.3	50	16
626.3	100	16
626.3	200	16
4532.1	50	32
Rational Chebyshev functions TB(X; L) with $L = 20$		
18.9	50	4
96.9	50	8
96.9	100	8
625.6	50	16
4527.2	50	32

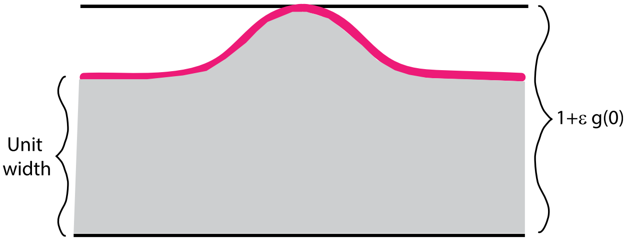

Fig. 10. The computational domain for the immersed boundary method is a strip of uniform width of $[-\infty, \infty] \otimes [0, 1 + \epsilon g(0)]$ where $g(0)$ is the maximum of the function $g(t)$ which describes the bulge in the top boundary of the waveguide. The physical domain is shaded, bounded by $y = 1 + \epsilon g(t)$ [thick pink curve]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

8. The immersed boundary method

Fig. 10 illustrates the physical domain [shaded] and computational domain [infinite rectangle of uniform width] for the immersed boundary method. No change of coordinate is necessary, and thus $X = x$ and $Y = y$. The computational domain completely encloses ("immerses") the physical domain.

By using the implicitization procedure described in the Appendix, the variation of the upper boundary of the waveguide with the down-channel coordinate X is obtained as

$$y_{top} = 1 + \epsilon \frac{\sin(2)}{d} + \epsilon^2 \, 2 \, \frac{\sin(2)\,(\sinh(2X))^2}{d^3} + O(\epsilon^3) \qquad (60)$$

$$d = \cos(2) + \cosh(2X) \, . \qquad (61)$$

The maximum deviation of the upper wall is

$$y_{max} = \max_x (y_{top}(x)) = \epsilon \frac{\sin(2)}{\cos(2) + 1} \approx 1.557\epsilon \, . \qquad (62)$$

The physical domain is then embedded into a computational domain which is $[0, \infty] \otimes [0, y_{max}]$. (Because the wall perturbation is symmetric about $x = 0$, all eigenmodes have definite parity with respect to $x = 0$; by employing a basis of functions symmetric in x, we restrict the numerical domain to $x \geq 0$.) No coordinate change is necessary; spectral methods are happy with rectangular domains. But how can we impose Dirichlet boundary conditions on a curve that does not coincide with the boundaries of the computational domain?

The answer is that we assume a solution in the form

$$\psi(x, y) = B(x, y) v(x, y) \qquad (63)$$

where $v(x, y)$ is an unconstrained sum of tensor product basis functions and where $B(x, y) = 0$ everywhere on the boundaries of the asymmetric waveguide. To mimic the sinusoidal structure of the unperturbed ground state, we used

$$B(x, y) = \sin\left(\pi \frac{y}{y_{top}(x)}\right) . \qquad (64)$$

The partial differential equation for v is

$$B\left\{v_{xx} + v_{yy}\right\} + 2B_x v_x + 2B_y v_y + \left\{B_{xx} + B_{yy}\right\} v + EBv = 0. \qquad (65)$$

Construction of the immersed boundary pseudospectral matrix is more expensive than for the analogous matrix of the conformal mapping approach because it is necessary at each point of the tensor product grid to evaluate v and its two first derivatives and also B and four of its derivatives. However, the resulting numerical eigenmodes will vanish at all points on the boundaries and a coordinate change was not needed.

Table 6 shows that, with the bulging wall parameterized as a function $y(x)$ using the non-perturbative series method of Appendix, the immersed boundary method gives exactly the same numbers to all digits shown as does conformal mapping. The choice between the two strategies must be based on convenience and ease of programming, not accuracy.

9. Timings and extensions to very large *N*

Although only approximations to the lowest eigenvalue are reported, we had no difficulty in computing many eigenvalues using the QZ method to calculate all eigenvalues of the discretization matrix. The techniques of discriminating between matrix eigenvalues which are accurate and inaccurate approximations to eigenvalues of the differential equation are explained in Chapter 7 of [2] and in [19]. Unfortunately, the cost is about $O(10M^3N^3)$ operations for a matrix of dimension $N_{total} = M \times N$.

In most "bent waveguide" problems, however, only the ground-state eigenvalue is of interest. One can enormously reduce execution time by using one-or-a-few eigenvalue solvers such as those based on the Lanczos or Arnoldi algorithms. In Matlab, it is merely necessary to replace **eig(AA,BB)** by **eigs(AA, BB, 1, lambda0)**. where λ_0 is an approximation to the targeted eigenvalue. Here, the perturbatively-predicted eigenvalue is always a very good initialization.

High-order perturbation theory may require accuracy beyond the limits of Matlab, or pseudospectral matrices too big for a workstation's memory or both. It is important to note that it is never necessary, with Arnoldi/Krylov space methods, to explicitly compute the pseudospectral matrix. All that is needed is the *vector* which is the matrix–vector product of the pseudospectral matrix with the vector of spectral coefficients. This matrix–vector product is just the *spectral residual* of the differential equation and can be computed without matrices using the Fast Fourier Transform (Chap. 15 of [2].) Navarra expressed this happy reality nearly thirty years ago [20]:

> one feature that makes this method very suitable ... is that the matrix A need not be stored ... only the vectors Ax are needed. ... this operation may be performed numerically by using a linearized version of a [mixed spectral-finite difference] General Circulation Model. (pg. 144 of [20]).

Unfortunately the Arnoldi iteration is not enough. Xue and Elman write in the introduction to [21]:

> Many scientific and engineering applications require a small group of eigenvalues closest to a specified shift or those with largest or smallest real parts. The shift-invert and Cayley transformations are the two most commonly used spectral

Table 6
Comparison between the conformal mapping strategy and the immersed boundary strategy: Computed eigenvalue changes δ in the ground state for various basis sets in various truncations where $\delta = \pi^2 - E$ for $\epsilon = 1/1000$. "Conformal" and "Immersed" indicate which strategy was applied. M is the number of Fourier cosine basis functions in the down-channel direction after the X coordinate was transformed by the hyperbolic sine mapping ; N is the number of Chebyshev polynomials in the cross-channel basis; $\mathcal{L}$ is the scaling factor used in the hyperbolic sine mapping.

Method and basis	L	M	N	$\lvert\delta\rvert$	rel. error
Conformal map	$\mathcal{L} = 6$	$M = 50$	$N = 24$	0.000378313474209	−2e−05
Immersed	$\mathcal{L} = 6$	$M = 50$	$N = 24$	0.000378313474209	−1.9871e−05
Conformal map	$\mathcal{L} = 6$	$M = 100$	$N = 24$	0.000378320991591	−2.3e−10
Immersed	$\mathcal{L} = 6$	$M = 100$	$N = 24$	0.000378320991591	−2.2591e−10
Conformal map	$\mathcal{L} = 6$	$M = 150$	$N = 24$	0.000378320991676	0
Immersed	$\mathcal{L} = 6$	$M = 150$	$N = 24$	0.000378320991676	0
Conformal map	$\mathcal{L} = 6$	$M = 200$	$N = 24$	0.000378320991676	0
Immersed	$\mathcal{L} = 6$	$M = 200$	$N = 24$	0.000378320991676	0

> transformations to map these eigenvalues to the dominant ones of the transformed operator, so that they can be readily computed by eigenvalue algorithms. The major challenge of this approach is that a linear system of equations involving a shifted matrix needs to be solved in each step (outer iteration) of the eigenvalue algorithm.

What they mean is that the Arnoldi and Lanczos iterations are, in their simplest form, variants of the Power Eigenvalue Method in that they compute the eigenvalue of *largest* magnitude. For bent quantum waveguides, the relevant eigenvalue is the *smallest*. The "shift-invert" strategy is based on the fact that the groundstate eigenvalue is the largest eigenvalue of the *inverse* of the matrix after the eigenvalue has been shifted. Thus, to compute the smallest eigenvalue of $Av = \lambda Bv$, the Arnoldi algorithm is applied to

$$(A - \lambda_0 B)^{-1} Bv = \frac{1}{\lambda - \lambda_0} v \tag{66}$$

where the shift λ_0 is an approximation to the desired eigenvalue λ. The property that large matrices such as A and B and so on need never be explicitly computed or stored can be preserved by performing all matrix inversions, etc., through preconditioned iterations.

Our paper would be incomplete without this discussion of the fact that FFT and Arnoldi methods can extend our spectral method to much larger numbers of basis functions than displayed in our tables. However, such fast but complicated algorithms were not needed here.

Fig. 11 shows that the QZ algorithm is indeed cubic in the matrix size. The discretization matrices are dense matrices of dimension $M \times N$ with $2M^2N^2$ matrix elements combined. The cost of assembling the matrix directly as done here also scales as $O(N_{total}^2)$; as noted just above, this cost can be reduced by an order of magnitude by Arnoldi/FFT methods which do not explicitly construct the matrix. We preferred overnight runs to the additional programming effort and used the slower set-up, element-by-element, in our program.

The single-eigenvalue computation has the same scaling with N and M as the non-FFT assembly of the eigenvalue matrices, but with a proportionality constant an order of magnitude smaller.

Table 7 gives a few numerical values. With 160 basis functions in X and 25 in Y, there are a total of 4000 basis functions. Computing all the eigenvalues of the 4000 × 4000 generalized matrix eigenproblem takes about 2 min. Extrapolating to larger matrix sizes using the cubic scaling, a 10,000 × 10,000 matrix would require about fifteen times as many floating point operations and can be completed in a little less than eight hours, an overnight run. However, to find just *one* eigenvalue with 4000 unknowns takes [adding setup time to the cost of Arnoldi eigensolver] only 6.5 s, and about six times as much (forty seconds) for $N_{total} = 10{,}000$.

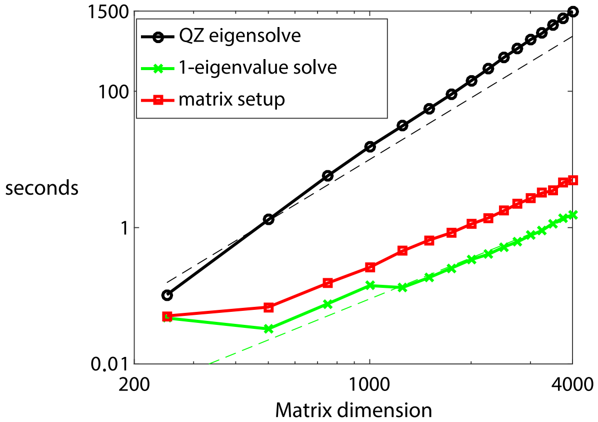

Fig. 11. The dashed lines are power-law fits; the upper dashed line (blue) is N_{total}^3 while the lower line is N_{total}^2. The cost of the QZ method, applied to find all eigenvalues and eigenfunctions, is cubic in matrix dimension while the other two procedures are only quadratic. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

10. Summary

We have developed strategies and multiscale spectral algorithms to solve the stationary Schrödinger equation (Helmholtz equation) in a waveguide which is a small perturbation of an infinitely long strip of uniform width. A major numerical challenge is that in the down-channel coordinate, there are two distinct spatial scales that differ by $O(1/\epsilon)$. These spatial scales do not apply in different and distinct regions, as true for boundary layer flows, but rather are commingled in the neighborhood of the wall perturbation. We show that classical spectral methods are quite adequate for this problem if the tensor product basis is highly anisotropic and wisely chosen.

We compare two strategies for coping with the perturbed boundary. The first strategy is to use a conformal mapping to transform the physical domain into a computational domain which is identical in shape (a rectangle) to the unperturbed waveguide. For strongly deformed domains, conformal mapping is not usually a good option for grid generation because the user has little control over the density of grid points. However, when the conformal mapping is only a small perturbation of the identity transformation, this difficulty is missing. It is likely that one can generate the conformal maps by perturbation theory as in [7], though we have not pursued this. For a general coordinate mapping, the transformed equation is awash with messy, coding-error-producing metric factors. With a conformal mapping, however, the Schrödinger equation is modified only in that the eigenvalue is multiplied by a map-dependent function $1 + \sigma(X, Y)$.

Table 7
Timings (in seconds) on an HP420 workstation running Windows 7 and Matlab 2012.

M	N	N_{total}	Time: QZ	Time: 1 eigenvalue	Time: matrix creation
40	25	1000	14.8	0.089	0.07
80	25	2000	146.5	0.33	1.2
120	25	3000	570.3	0.77	2. 7
160	25	4000	1494.0	1.5	5.0

The second strategy is to embed the deformed domain inside the uniform strip (no longer of unit cross-channel width) and solve the problem without mapping. Boundary conditions are imposed by writing the unknown as $\psi(x, y) \equiv B(x, y)v(x, y)$ where the boundary is the curve $B(x, y) = 0$ and $v(X, Y)$ is computed by the pseudospectral method unconstrained by boundary conditions.

Both strategies allow a tensor product basis and grid. This is good both for efficiency and ease-of-programming. We compared a variety of basis functions to arrive at the following.

1. When a conformal map from this physical domain to a uniform waveguide is available, there is only a single metric factor and the pseudospectral code is very simple and reliable.
2. Fourier domain truncation combined with the change of coordinate $X = \sinh(\mathfrak{L}\, t)$ gives much faster convergence than the rational Chebyshev basis TB_n although the latter also converges exponentially fast.
3. Chebyshev polynomials in the cross-channel coordinate Y converge much faster than sine functions even though the latter are the unperturbed normal modes.
4. Using a single sine function $\sin(\pi Y)$ in the cross-channel coordinate with a Galerkin discretization in this coordinate gives approximations that are remarkably good, more accurate than variational eigenvalue estimates for all ϵ and improving upon lowest order perturbation theory until ϵ is very, very tiny. More cross-channel sine functions produce little improvement while the Chebyshev polynomials decrease exponentially until the error curves for the two different Y basis sets merge, being constrained for $N > 10$ by the down-channel (X coordinate) resolution.
5. For sufficiently small perturbation parameter ϵ, the perturbation theory is always better because the spectral methods are limited by roundoff error, though not severely; ten digits of accuracy for the ground state eigenvalue in sixteen decimal place floating point arithmetic was the normal roundoff limit for our example.
6. The immersed boundary method worked well, but the Krylov strategy of implementing boundary conditions requires evaluating additional derivatives of the unknown and all derivatives up to second order for $B(x, y)$, the function whose zero contours implicitly determine the boundaries. Conformal mapping is simpler whenever a conformal map is available.

Checking perturbation problems is a demanding application because perturbation theory is accurate only when the parameter ϵ that measures the strength of the perturbation is small. Recall the definition that $\delta \equiv E(\epsilon) - E(0)$ is the pertubation-induced change in the eigenvalue. Perturbation theory shows that $\delta \sim O(\epsilon^2)$. To verify the perturbation theory, numerically-computed eigenvalues must be accurate to within an error small compared to δ. It is very difficult to achieve such high accuracy without a spectral method.

All the calculations presented here used Matlab in the usual sixteen decimal place floating point arithmetic. Although only approximations to the lowest eigenvalue are reported, we had no difficulty in computing many eigenvalues as discussed in the previous section.

In most "bent waveguide" problems, however, only the ground-state eigenvalue is of interest. One can enormously reduce execution time by using one-or-a-few eigenvalue solvers. In Matlab, it is merely necessary to replace **eig(AA,BB)** by **eigs(AA, BB, 1, lambda0)**. where λ_0 is an approximation to the targeted eigenvalue, which need not be very accurate. The perturbatively-predicted eigenvalue was always a very good initialization.

High-order perturbation theory may require accuracy beyond the limits of Matlab, or pseudospectral matrices too big for a workstation's memory or both. It is important to note that it is never necessary, with Arnoldi/Krylov space methods, to explicitly compute the pseudospectral matrix. All that is needed is the *vector* which is the matrix–vector product of the pseudospectral matrix with the vector of spectral coefficients. This matrix–vector product is just the spectral residual of the differential equation and can be computed without matrices using the Fast Fourier Transform (Chap. 15 of [2]) as discussed in the previous section.

It is annoying that both an inner and an outer iteration are needed for huge N_{total}, but with such an approach, the spectral methods described here can be extended on a desktop workstation to very large basis size N_{total} and extremely high accuracy.

In three dimensions, conformal mapping is not available except for special cases. Non-conformal maps can be applied; the only drawback is that there are many more metric factors instead of just one in the transformed PDE.

The immersed boundary method also applies in three dimensions. It is necessary to compute many more derivatives of the boundary function $B(x, y, z)$, but no change in underlying principles.

Acknowledgment

This work was supported by the National Science Foundation of the U.S. under DMS-1521158.

Appendix. Reparameterization by interpolation

In computer graphics, parametric and implicit descriptions of a curve are both useful for different purposes, and a common task is to convert from one form to the other. One conversion strategy is to parameterize y in terms of x by using an infinite series, preferably the same type of series used as the down-channel basis. Let x_k denote the usual interpolation points associated with the basis. If we set the parameter t equal to each of these points in turn, we obtain the set of points

$$\tilde{x}_k = x_k + \epsilon f(x_k) \tag{A.1}$$

$$\tilde{y}_k = 1 + \epsilon g(x_k), \qquad k = 1, 2, \ldots, M. \tag{A.2}$$

The $\tilde{x}_k$ differs from the x_k only by $O(\epsilon)$, which implies that interpolation at the $\tilde{x}_k$ inherits almost the same spectral accuracy as interpolation at the canonical interpolation points. The upper boundary has the series representation

$$y(x) = \sum_{m=1}^{M} \aleph_m \, \chi_m(x) \tag{A.3}$$

where the $\aleph_m$ are the elements of a vector that solves a standard matrix problem whose right-hand side is the vector whose elements are $\tilde{y}_k$ and the elements of the Vandermonde matrix are $\chi_{col}(\tilde{x}_{row})$ where "row" and "col" are the row and column indices of the matrix and the $\chi_m(X)$ are the basis functions in X, the same set that was used to approximate $\psi(X, Y)$.

The interpolation series can be evaluated at the canonical interpolation points to compute $y(x)$ with high accuracy at the points where its values are needed for the pseudospectral discretization.

References

[1] P. Amore, J.P. Boyd, F.M. Fernandez, M. Jacobo, P. Zhevandrov, Anziam J. Math. (2016) (in press).
[2] J.P. Boyd, Chebyshev and Fourier Spectral Methods, second ed., Dover, Mineola, New York, 2001, p. 665.
[3] L.N. Trefethen, Spectral Methods in Matlab, Society for Industrial and Applied Mathematics, Philadelphia, 2000, p. 220.
[4] J. Shen, T. Tang, L.-L. Wang, Spectral Methods: Algorithms, Analysis and Applications, Springer Series in Computational Mathematics, Springer, Heidelberg, 2011, p. 500.
[5] T.A. Driscoll, L.N. Trefethen, Schwarz-Christoffel Mapping, Vol. 8, Cambridge University Press, 2002, p. 150.
[6] L.V. Kantorovich, V.I. Krylov, Approximate Methods of Higher Analysis, Interscience, New York, 1958, p. 681. Trans. by Curtis D. Benster.
[7] J.P. Boyd, Amer. Math. Monthly 123 (3) (2015) 241–257.
[8] J.P. Boyd, Appl. Math. Comput. 161 (2) (2005) 591–597.
[9] L. Badea, P. Daripa, Numer. Algorithm 30 (2002) 199–239.
[10] M. Lyon, O.P. Bruno, J. Comput. Phys. 229 (9) (2010) 3358–3381.
[11] L.N. Trefethen, J.A.C. Weideman, SIAM Rev. 56 (3) (2014) 385–458.
[12] J.P. Boyd, J. Comput. Phys. 110 (1994) 360–372.
[13] J.P. Boyd, Appl. Math. Comput. 301 (2016) 214–223.
[14] J.P. Boyd, J. Sci. Comput. 3 (1988) 109–120.
[15] J. Andre C. Weideman, A. Cloot, Comput. Methods Appl. Mech. Engrg. 80 (1990) 467–481.
[16] J.P. Boyd, J. Comput. Phys. 69 (1987) 112–142.
[17] S. Olver, A. Townsend, SIAM Rev. 55 (3) (2013) 462–489.
[18] Z. Huang, J.P. Boyd, J. Comput. Phys. 300 (1) (2015) 1–4.
[19] J.P. Boyd, J. Comput. Phys. 126 (1996) 11–20 Corrigendum, 136, no. 1, 227-228 (1997).
[20] A. Navarra, J. Comput. Phys. 69 (1) (1987) 143–162.
[21] F. Xue, H.C. Elman, SIAM J. Matrix Anal. 33 (2) (2012) 433–459.