Fast spectrally-accurate solution of variable-coefficient elliptic problems

John Strain[1]
Department of Mathematics
University of California
Berkeley, California 94720

April 1993

[1]Research supported by a NSF Young Investigator Award, Air Force Office of Scientific Research Grant FDF49620-93-1-0053, and the Applied Mathematical Sciences Subprogram of the Office of Energy Research, U.S. Department of Energy under Contract DE-AC03-76SF00098.

Abstract

A simple, efficient, spectrally-accurate numerical method for solving variable-coefficient elliptic partial differential equations in periodic geometry is described. Numerical results show that the method is efficient and accurate even for difficult problems including convection-diffusion equations. Generalizations and applications to phase field models of crystal growth are discussed.

1 Introduction

This paper presents a numerical method for solving the variable-coefficient second-order elliptic partial differential equation

$$\mathcal{L}u(x) := \sum_{i,j=1}^{d} a_{ij}(x)\partial_i\partial_j u(x) + \sum_{i=1}^{d} b_i(x)\partial_i u(x) + c(x)u(x) = f(x) \qquad (1.1)$$

in the box $B = [0,1]^d$ in $\mathbf{R}^d$, with periodic boundary conditions imposed on the boundary ∂B of the box. We assume that the coefficients are smooth and periodic, with $c(x) \le 0$, and we assume uniform ellipticity:

$$M|\xi|^2 \ge \sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \ge m|\xi|^2 \qquad (1.2)$$

for all $\xi \in \mathbf{R}^d$ and some constants $M, m > 0$. We do not require self-adjointness.

The method is based on representing u as a volume potential

$$u(x) = \int_B \bar{G}(x,x')\sigma(x')dx' = \bar{\mathcal{L}}^{-1}\sigma(x)$$

formed with the Green function $\bar{G}$ for the constant-coefficient averaged operator

$$\bar{\mathcal{L}} := \sum_{i,j=1}^{d} \bar{a}_{ij}\partial_i\partial_j + \sum_{i=1}^{d} \bar{b}_i\partial_i + \bar{c}. \qquad (1.3)$$

Here $\bar{g} = \int_B g(x)dx$. The operator $\mathcal{A} := \mathcal{L}\bar{\mathcal{L}}^{-1}$ is a bounded invertible operator on $L^2(B)$, and the equation $\mathcal{A}\sigma = f$ is equivalent to (1.1) but easier to solve. The solution u of (1.1) can be recovered from σ by evaluating the volume potential.

This method is spectrally accurate in the sense that the error decreases faster than any power of the grid size h as $h \to 0$, because convolution with the Green function $\bar{G}$ and differentiation can be applied with spectral accuracy. It is efficient because $\mathcal{A}$ is a bounded invertible operator on $L^2(B)$, so reasonable discretizations of $\mathcal{A}$ have bounded condition numbers independent of mesh size, and iterative methods then converge in an asymptotically bounded number of iterations. The method is extremely simple to program and trivial to parallelize, since most of the computational effort is spent performing

the fast Fourier transform. It works well even for convection-diffusion problems where the operator is far from self-adjoint; we note that the coefficients change sign frequently in our numerical examples, but the accuracy obtained depends only on the smoothness of the solution. The solution time grows as the complexity of the problem increases, but for a fixed problem it remains bounded as the mesh size decreases, once the solution is resolved.

We discuss generalizations in §5; the most important is efficient high-order accurate schemes for variable-coefficient problems in arbitrary smooth domains. The method also can be used to solve higher-order elliptic problems and systems; an example of the latter is the application to BDF discretizations of phase field models for crystal growth which we discuss in §4.

2 The method

Consider the equation (1.1), and average the coefficients over B to produce the constant-coefficient operator $\bar{\mathcal{L}}$ given by (1.3). By uniform ellipticity (1.2) and the linearity and positivity of averaging, $\bar{\mathcal{L}}$ is elliptic with the same m, M as $\mathcal{L}$. If $\bar{c}$ is strictly negative, then $\bar{\mathcal{L}}$ is invertible; otherwise, $\bar{c} = 0$ and we work with the subspace of $L^2(B)$ consisting of functions with mean zero, where $\bar{\mathcal{L}}$ is invertible. For simplicity of exposition, we assume from now on that $\bar{c} < 0$.

Since $\bar{\mathcal{L}}$ has constant coefficients and the boundary conditions are periodic, we use Fourier series. For convenience, we use multiindex notation: Z^d is the space of d-dimensional integer sequences $k = (k_1, k_2, \ldots, k_d)$ where each k_i is a positive or negative integer, and $|k| = \max |k_i|$. We take our Fourier series in the form

$$\sigma(x) = \sum_{k \in \mathbf{Z}^d} e^{2\pi\iota k \cdot x} \hat{\sigma}(k)$$

where $\iota = \sqrt{-1}$ and the Fourier coefficients $\hat{\sigma}(k)$ are defined by

$$\hat{\sigma}(k) = \int_B e^{-2\pi\iota k \cdot x} \sigma(x) dx.$$

Then we can apply $\mathcal{A} = \mathcal{L}\bar{\mathcal{L}}^{-1}$ explicitly:

$$\mathcal{A}\sigma(x) \quad = \quad \sum_{k \in \mathbf{Z}^d} \frac{\sum_{i,j=1}^d a_{ij}(x) 2\pi\iota k_i 2\pi\iota k_j + \sum_{i=1}^d b_i(x) 2\pi\iota k_i + c(x)}{\sum_{i,j=1}^d \bar{a}_{ij} 2\pi\iota k_i 2\pi\iota k_j + \sum_{i=1}^d \bar{b}_i 2\pi\iota k_i + \bar{c}} e^{2\pi\iota k \cdot x} \hat{\sigma}(k)$$

$$= \sum_{k \in \mathbf{Z}^d} \frac{\rho(x,k)}{\bar{\rho}(k)} e^{2\pi\iota k \cdot x} \hat{\sigma}(k)$$

where $\rho(x,k)$ and $\bar{\rho}(k)$ are the symbols of $\mathcal{L}$ and $\bar{\mathcal{L}}$. The ellipticity hypothesis (1.2) and the assumption $\bar{c} < 0$ imply that the denominator $\bar{\rho}(k)$ never vanishes. Thus $\mathcal{A}$ is a bounded invertible operator on $L^2(B)$, since $\bar{\mathcal{L}}^{-1}$ maps L^2 one-to-one and onto the Sobolev space $H^2(B)$, while $\mathcal{L}$ maps $H^2(B)$ back onto $L^2(B)$ (by e.g. Theorem 9.15 of [7], modified for the periodic case). The equation $\mathcal{A}\sigma = f$ therefore has a unique solution $\sigma \in L^2(B)$ if $f \in L^2(B)$, and we can recover u from σ by applying the Green function $u = \bar{\mathcal{L}}^{-1}\sigma$.

The numerical method now has three independent components; first, we need an iteration for solving $\mathcal{A}\sigma = f$, second, we need to approximate $\mathcal{A} = \mathcal{L}\bar{\mathcal{L}}^{-1}$ accurately, and third, we need to apply $\bar{\mathcal{L}}^{-1}$ to σ to get u.

We expect that almost any nonsymmetric iterative method (such as GMRES [13], QMR [6] or BI-CGSTAB [14]) can be used to solve $\mathcal{A}\sigma = f$, because $\mathcal{A}$ is bounded and invertible though neither symmetric nor positive definite in general. The rate of convergence will not depend on the mesh size, since we have no mesh size yet. For GMRES applied to $Ax = f$, for example, the residual $r_m = f - Ax_m$ after m steps satisfies [13]

$$||r_m|| \leq \kappa(X)\epsilon_m ||r_0||$$

where $A = X\Lambda X^{-1}$ is a diagonalizable $N \times N$ matrix, $\kappa(X) = ||X||_2 ||X^{-1}||_2$, and

$$\epsilon_m \leq \left(\frac{D}{d}\right)^{\nu} \left(\frac{R}{C}\right)^{m-\nu}.$$

Here we assume that A has ν eigenvalues $\lambda_1, \ldots, \lambda_\nu$ in the left half plane and $N - \nu$ in a circle $|C - \lambda| \leq R$ with $C > R > 0$, and we define

$$D = \max_{1 \leq i \leq \nu; \nu+1 \leq j \leq N} |\lambda_i - \lambda_j|$$

and $d = \min_{1 \leq i \leq \nu} |\lambda_i|$. Thus the restarted method GMRES(m) is guaranteed to converge if

$$m > \frac{\log \kappa(X)}{\log C/R} + \nu \left(1 + \frac{\log D/d}{\log C/R}\right),$$

and the convergence rate depends on the problem size *only* through eigenvalue bounds and $\kappa(X)$. For solving $\mathcal{A}\sigma = f$, therefore, the convergence rate of GMRES is independent of mesh size.

Our method can be seen as an analytic preconditioning of the differential operator, rather than a matrix preconditioning of a discretized problem. Matrix preconditioning helps solve a discrete problem even when it is not a good approximation to the continuous problem, but then the value of the computation is unclear. Our method can be seen as an analytic version of the preconditioning presented in [4] and more recently in [5, 8], extended to more general operators. It is also closely related to the approach of [11].

We now approximate $\mathcal{A}\sigma$ in the natural way by evaluating the Fourier coefficients of σ numerically, multiplying by the appropriate factors, and truncating the Fourier series. Thus we lay down a uniform grid of N^d points $x_j = (j_1 h, j_2 h, \ldots, j_d h)$, where $h = 1/N$ and each j_i runs from 1 to N. Since the problem is periodic we identify $x_i = 0$ with $x_i = 1$. We approximate $\hat{\sigma}(k)$ for $|k| \le N/2$ by the trapezoidal rule

$$\hat{\sigma}(k) \approx \hat{\sigma_h}(k) := h^d \sum_{j_i=1}^{N} e^{-2\pi\iota k\cdot x_j}\sigma(x_j)$$

with an error which is spectrally small if σ is smooth;

$$|\hat{\sigma}(k) - \hat{\sigma_h}(k)| \le C||\sigma||_{H^p(B)}h^p \qquad\qquad |k| \le N/2,$$

for every $p \ge d+1$. Here $H^p(B)$ is the pth-order Sobolev space on B [7] and C is a constant depending only on d.

These approximate Fourier coefficients can be evaluated by the FFT, which requires $O(N^d \log N)$ operations, and divided by $\bar{\rho}(k)$ to obtain $\hat{\sigma_h}(k)/\bar{\rho}(k)$. The averages $\bar{a}_{ij}$ are approximated by trapezoidal sums over the mesh points; since this is also an average, it preserves ellipticity just as well as integrating. Then we approximate $\mathcal{A}\sigma$ by

$$\begin{aligned}
\mathcal{A}_h\sigma_h(x) \;:=\; & \sum_{i,j=1}^{d} a_{ij}(x) \sum_{|k|\le N/2} 2\pi\iota k_i 2\pi\iota k_j e^{2\pi\iota k\cdot x}\hat{\sigma_h}(k)/\bar{\rho}(k) \\
+\; & \sum_{i=1}^{d} b_i(x) \sum_{|k|\le N/2} 2\pi\iota k_i e^{2\pi\iota k\cdot x}\hat{\sigma_h}(k)/\bar{\rho}(k) \\
+\; & c(x) \sum_{|k|\le N/2} e^{2\pi\iota k\cdot x}\hat{\sigma_h}(k)/\bar{\rho}(k)
\end{aligned}$$

We have to extract the x-dependence from the sums in order to evaluate them on the mesh with the FFT. If we take advantage of the equality of

mixed partials, we need $1 + (d+1)(d+2)/2$ FFT's and $(d+1)(d+2)N^d/2$ multiplications and additions to evaluate $\mathcal{A}_h\sigma$, assuming that the $(d+1)(d+2)/2$ distinct coefficients of $\mathcal{L}$ have already been evaluated at the mesh points.

Thus each application of $\mathcal{A}_h$ costs $O(N^d \log N)$ operations, even though the matrix has N^{2d} elements. If we can solve $\mathcal{A}_h\sigma_h = f$ in a number of iterations independent of the mesh size, then the total cost will be $O(N^d \log N)$, only a constant factor times the cost of solving a constant-coefficient problem and much smaller than the cost of solving $\mathcal{L}u = f$ by standard iterative or direct methods or even the cost of a standard multiplication by $\mathcal{A}_h$. Our experiments with GMRES show that in fact, once the solution is resolved, the number of iterations does not increase as the mesh is refined.

Once we solve $\mathcal{A}_h\sigma_h = f$, we have σ_h, so we compute an approximate solution u_h in the natural way. We divide the Fourier coefficients $\hat{\sigma_h}$ by $\bar{\rho}(k)$ and evaluate the resulting truncated Fourier series u_h on the mesh. Usually u_h is even more accurate than σ_h, since the higher modes are damped by $\bar{\mathcal{L}}^{-1}$.

It may be worthwhile to compare our method with some of the many other techniques available for this problem. The advantage of our method over multigrid methods [1] (which are equally efficient for a given grid size but less accurate) is its spectral accuracy, while the advantage over standard spectral methods [3] (which are equally accurate but less efficient for a given grid size) is its efficiency.

3 Numerical results

Our numerical results use $d = 2$ dimensions and a solution u given by

$$u(x) = \exp(\cos(2\pi k_1 x_1)\cos(2\pi k_2 x_2)).$$

We calculated σ and f from u by applying $\bar{\mathcal{L}}$ and $\mathcal{L}$ exactly, then solved the problem numerically and calculated the error in σ and u.

The variable coefficients of $\mathcal{L}$ were constructed from six $(M+1)^2$-term Fourier cosine series

$$F_s(x) = \sum_{k_1,k_2=0}^{M} F_k \cos(2\pi k_1 x_1)\cos(2\pi k_2 x_2)$$

with coefficients F_k generated randomly on [-1,1] for each $s = 1$ through 6. Since we want $\mathcal{L}$ elliptic, we generated a 2 by 2 upper triangular matrix F with entries F_1, F_2 and F_3, and set $(a_{ij}) = I + F^T F$ where I is the 2 by 2 identity matrix. Thus $a_{11} = 1 + F_1^2$, $a_{12} = 2F_1F_2$, $a_{21} = 0$, and $a_{22} = 1 + F_2^2 + F_3^2$. The hypothesis of uniform ellipticity is satisfied with $m = 1$. The first-order coefficients b_i were given by random Fourier series F_4 and F_5, multiplied by a convection coefficient β which was varied to increase the convective terms. The zero-order coefficient c was formed by setting $c = -F_6^2$, to ensure $c(x) \leq 0$. Note that the second-order and zero-order coefficients can vary on scales twice as small as the first-order terms, since they are quadratic functions of the F_i's.

The choice of starting values is important in iterative methods; we experimented with four starting strategies of increasing accuracy. First $\sigma = 0$, second, σ randomly generated, third, $\sigma = f$; and fourth, σ constructed recursively by solving the problem on a coarser grid and using trigonometric interpolation. The first three methods required more time than the last, so we present results only for the last strategy, with the solution initialized on the coarsest grid by setting $\sigma = f$. We display results in Figure 1 in the form of log-log plots of maximum error in u versus total CPU time T on a Cray-2 in seconds; the time plotted is the cumulative time required for all the solves on smaller grids as well as the current grid. GMRES(10) was used, with a stopping tolerance of 10^{-11} for the norm of the residual. We present results for solution wavenumbers $k_1 = k_2 = 1$, 5 and 9, with coefficient wavenumbers $M = 1$, 5 and 9 and $\beta = 10$. More detailed information is presented in Table 1. The number of iterations required to solve these problems depended strongly on the regularity of the solution, weakly on the variation in the coefficients, and not at all on the mesh size. This is extremely encouraging since one of the main applications of this type of solver is to nonlinear problems, where the coefficients are no smoother than the solution.

The numerical results clearly display the spectral accuracy and efficiency of the method over a wide range of solution and coefficient parameters, and reveal another interesting feature of the method; it informs the user when the solution is sufficiently resolved on the current grid by requiring zero iterations to satisfy the stopping criterion. If the solution on the previous grid is already accurate to the desired tolerance, then the iteration is satisfied on the current grid as well and only one matrix-vector multiplication is required, to compute the residual. Thus if one computes the solution on a sequence of grids,

$k_1 = k_2 = 5, M = 5, \epsilon = 10^{-11}$			
N	I	T	E
16	84	0.80	0.84E+00
32	112	3.44	0.35E+00
48	111	6.72	0.14E-01
64	105	10.37	0.13E-03
80	82	12.49	0.33E-06
96	64	14.31	0.12E-08
112	39	17.09	0.39E-10
128	11	4.48	0.79E-11
144	3	1.87	0.73E-11
160	1	1.09	0.50E-11
176	0	1.03	0.47E-11
192	0	0.79	0.49E-11
208	0	1.48	0.42E-11
224	0	1.57	0.45E-11
240	0	1.21	0.42E-11
256	0	1.41	0.44E-11

$k_1 = k_2 = 9, M = 9, \epsilon = 10^{-11}$			
N	I	T	E
16	152	1.43	0.10E+01
32	171	5.33	0.10E+01
48	242	14.81	0.72E+00
64	276	27.48	0.19E+00
80	298	44.92	0.12E-01
96	297	65.54	0.19E-02
112	276	117.72	0.61E-04
128	239	89.58	0.64E-05
144	199	96.35	0.12E-06
160	169	93.74	0.50E-07
176	134	142.05	0.56E-08
192	97	81.28	0.71E-09
208	53	80.72	0.11E-09
224	18	30.34	0.46E-10
240	5	7.34	0.36E-10
256	1	2.86	0.42E-10

Table 1: Maximum error E in u, divided by the maximum of u, versus the mesh size N, the number of GMRES iterations required I and the CPU time required T per mesh.

the method will become extremely inexpensive as the desired resolution is approached.

4 Application to crystal growth

One of our motivations in developing the elliptic solvers presented above is the phase field model of crystal growth, a continuum problem requiring the solution of 2 by 2 second-order parabolic systems in two or three space dimensions [2]. The boundary conditions are simple, since the interest is in fundamental physics rather than engineering, and periodic boundary conditions are thus appropriate. The phase field equations can be put in the

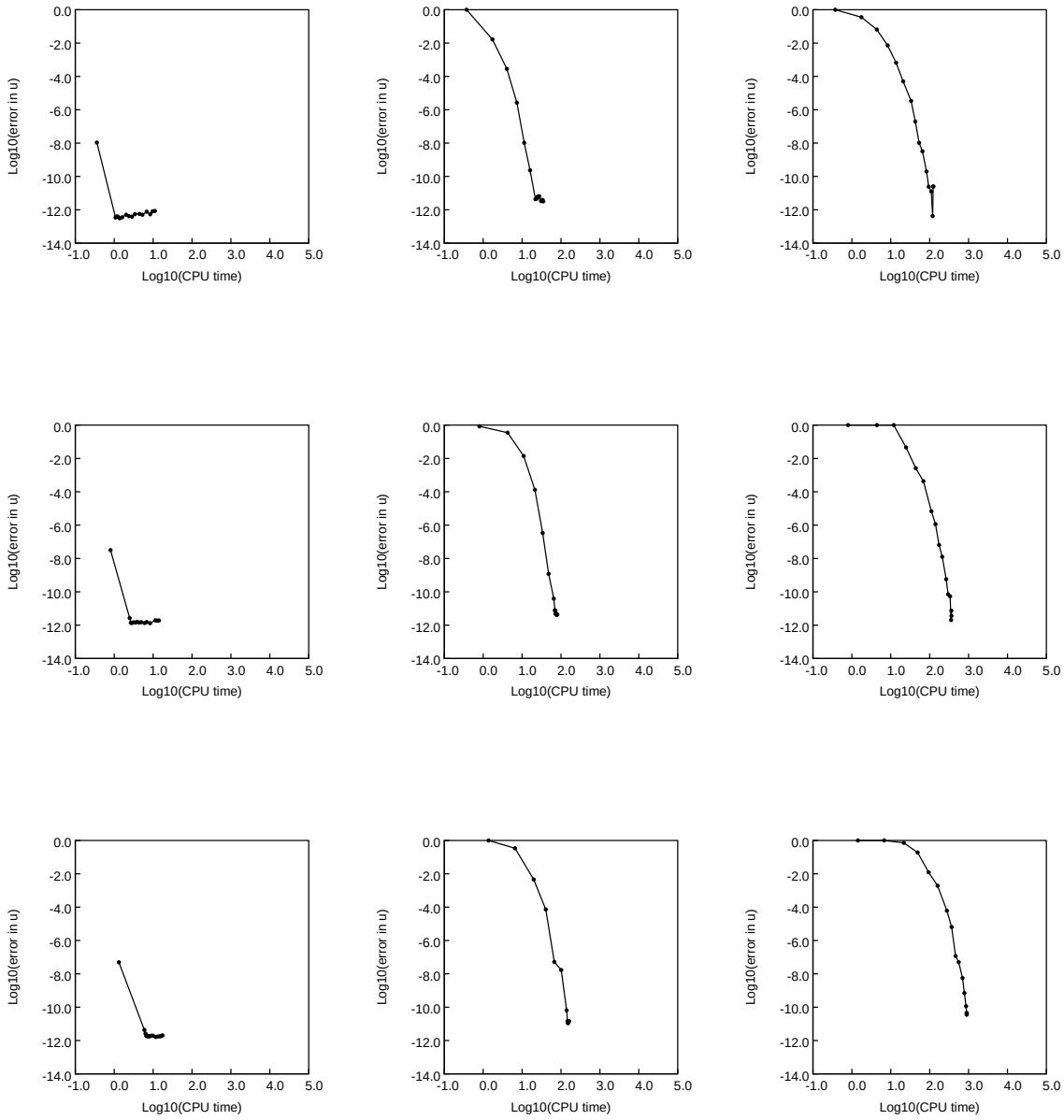

Figure 1: Graphs of maximum error in u versus Cray-2 CPU time for our method, with $K = 1$, 5 and 9 (left to right) and $M = 1$, 5 and 9(top to bottom), with $\beta = 10$. For each plot, $N = 16$ through 256 in steps of 16. The starting value was constructed by trigonometric interpolation of the solution on the previous mesh, except for $N = 16$ where we set $\sigma = f$.

form

$$U_t = \Delta A U + F(U)$$

where $U = (u_1, u_2)^t$, Δ is the Laplacian, A is a 2 by 2 matrix of contants, and $F(U)$ is a cubically nonlinear function. This system is stiff, and therefore should be discretized in time by implicit backward difference formulae [9]. At each time step, one needs to solve a nonlinear elliptic system with a good initial guess available from the previous time step. The elliptic system can be linearized with a damped Newton method, giving a sequence of linear variable-coefficient elliptic systems which are ideal applications for the technique developed in this paper. The extension of our method to solve these systems is straightforward; since the principal part is constant-coefficient already, it need not be averaged.

5 Generalizations

The method employed in this paper admits generalizations to elliptic systems, to other boundary conditions, and to arbitrary domains. Elliptic systems appear trivial once single equations can be solved.

When we have Dirichlet or Neumann boundary conditions on ∂B, spectral accuracy requires the use of orthogonal polynomial basis functions rather than trigonometric functions [3]. These bases do not diagonalize constant-coefficient operators, so other operators should be used to form more appropriate potentials. In each case, the averaged operator should be constructed with a weighted average and a structure which is diagonalized by the basis used. Thus the method generalizes to any domain which admits spectrally accurate methods for classes of operators produced by averaging.

The method also generalizes to arbitrary domains, using fast Helmholtz solvers. The basic idea is the same: given a linear variable-coefficient elliptic equation $\mathcal{L}u = f$ with homogeneous boundary conditions, we convert it to an integral equation $\mathcal{A}\sigma = f$ with the averaged constant-coefficient operator $\bar{\mathcal{L}}$ with the same boundary conditions. Iteration of $\mathcal{A}$ (using GMRES, QMR or BI-CGSTAB) converges in a number of steps independent of the mesh size since $\mathcal{A}$ is invertible and bounded on L^2. The operator $\bar{\mathcal{L}}^{-1}$ can no longer be approximated with the FFT, since the problem is not periodic; but $\bar{\mathcal{L}}$ can be transformed to the Helmholtz operator $\Delta + K$ by change of variable. The

Helmholtz operator can be inverted efficiently with a fast Helmholtz solver [12, 10].

References

[1] W. L. Briggs. *A multigrid tutorial.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987.

[2] G. Caginalp. Stefan and Hele-Shaw type models as asymptotic limits of the phase-field equations. *Phys. Rev. A*, 39:5887–5896, 1989.

[3] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral Methods in Fluid Dynamics.* Springer-Verlag, New York, 1987.

[4] P. Concus and G. H. Golub. Use of fast direct methods for the efficient numerical solution of nonseparable elliptic equations. *SIAM J. Numer. Analysis*, 10:1103–1120, 1973.

[5] M. Deville and E. Mund. Chebyshev pseudo-spectral solution of second-order elliptic equations with finite element preconditioning. *Jour. Comput. Phys.*, 60:517–533, 1985.

[6] R. Freund and N. M. Nachtigal. QMR: a quasi-minimal residual method for non-Hermitian linear systems. *Numerische Mathematik*, 60:315–339, 1991.

[7] D. Gilbarg and N. S. Trudinger, editors. *Elliptic partial differential equations of second order.* Springer-Verlag, 1983.

[8] H. Guillard and J. A. Désidéri. Iterative methods with spectral preconditioning for elliptic equations. In C. Canuto and A. Quarteroni, editors, *Spectral and high order methods for partial differential equations: proceedings of the ICOSAHOM '89 Conference, Villa Olmo, Como, Italy, 26-29 June, 1989.* Elsevier Science Pub. Co., 1990.

[9] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential equations II : Stiff problems.* Springer-Verlag, 1991.

[10] W. Proskurowski and O. Widlund. On the numerical solution of Helmholtz's equation by the capacitance matrix method. *Math. Comp.*, 32:103–120, 1978.

[11] V. Rokhlin. Application of volume integrals to the solution of partial differential equations. *Comp. and Maths. with Appls.*, 11:667–679, 1985.

[12] V. Rokhlin. Rapid solution of integral equations of scattering theory in two dimensions. *J. Comp. Phys.*, 86:414–439, 1990.

[13] Y. Saad and M. R. Schultz. GMRES: A generalized minimum residual method for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 7:856–869, 1986.

[14] H. A. van der Vorst. BI-CGSTAB: a fast and smoothly converging variant of BI-CG for the solution of nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.*, 13:631–644, 1992.

1991 Subject Classifications: 65N35, 65T20, 65N22.

Key words and phrases: elliptic solvers, preconditioning, spectral methods.

E-mail address: strain@math.berkeley.edu.