Iterative solution of elliptic difference equations using fast direct methods*

by PAUL CONCUS

University of California
Berkeley, California

INTRODUCTION

In recent years, fast direct methods have been developed for the numerical solution of the Poisson equation on a rectangle.[1,2] By taking advantage of the special block structure of the approximating discrete equation on a uniform rectangular mesh, these methods obtain the solution with striking efficiency and accuracy. A comparison of fast direct methods with other methods can be found in Reference 3, and the extension to more general separable elliptic equations in Reference 4.

Here, a technique is discussed for using fast direct methods to solve iteratively certain more general formally self-adjoint strongly elliptic equations $\mathfrak{L}u=f$, which are not necessarily separable. Dirichlet conditions on the boundary of the rectangle are considered, although the technique applies with slight alteration to other boundary conditions for which fast methods are suitable. The approach is to utilize a modified form of the iterative procedure

$$-\Delta u_{n+1}=-\Delta u_n-\tau(\mathfrak{L}u_n-f), \quad \Delta\equiv\partial^2/\partial x^2+\partial^2/\partial y^2 \quad (1)$$

proposed for numerical computation in conjunction with alternating-direction methods by D'yakonov[5] and discussed recently by Widlund.[6] This procedure, in addition to being of a form suitable for fast direct methods, has the desirable feature that for well-behaved problems its convergence rate is essentially independent of mesh size.

As it stands, however, iteration (1) may be too slowly convergent to be of practical importance, even when optimal values of the parameter τ are used. The means employed in this paper for improving its convergence rate are: (i) scaling the original problem $\mathfrak{L}u=f$ and iterating instead with the scaled problem $\mathfrak{M}w=q$; (ii) using, instead of (1), the shifted iteration

$$(-\Delta+K)w_{n+1}=(-\Delta+K)w_n-\tau(\mathfrak{M}w_n-q), \quad (2)$$

where K is a suitably chosen constant; (iii) applying Chebyshev acceleration. Algorithms for the fast direct solution of the discrete Poisson equation in a rectangle can handle iteration (2), which requires the repeated solution of a Helmholtz equation, with the same rapidity as they can (1).

Related iterative techniques for elliptic equations are studied in References 7 and 8 in connection with alternating-direction methods and in References 9 and 10 in connection with Stone's sparse factorization method. This latter method is formally similar to the one here; however, the present technique has the desirable property of being based on a more natural splitting of the operator.

ITERATIVE PROCEDURE

Description

In its simplest form, the iterative procedure considered here solves numerically on a uniform rectangular mesh the problem

$$\mathfrak{L}u\equiv-\nabla\cdot[a(x,y)\nabla u]=f(x,y) \quad \text{on } \mathfrak{R} \quad (3)$$

$$u(x,y)=g(x,y) \quad \text{on } \partial\mathfrak{R}, \quad (4)$$

where $\mathfrak{R}$ is the rectangle $0<x<c$, $0<y<d$ and $a(x,y)$ is strictly positive on $\mathfrak{R}$ and its boundary $\partial\mathfrak{R}$. [It is assumed that $a(x,y)$, $f(x,y)$, and $g(x,y)$ are such that the solution $u(x,y)$ is sufficiently well behaved near the corners of $\mathfrak{R}$ so that special numerical methods are not required there.] The positivity of $a(x,y)$ implies that $\mathfrak{L}$ is positive definite.

* This work was supported in part by the U.S. Atomic Energy Commission.

If $a(x, y)$ has bounded second derivatives on the closed rectangle, which is the case of principal interest for use of the procedure, the change of variable is performed.

$$w(x, y) = [a(x, y)]^{1/2} u(x, y). \quad (5)$$

Then, after division by $a^{1/2}$, (3) becomes

$$a^{-1/2}\mathfrak{L}u = \mathfrak{M}w \equiv -\Delta w + p(x, y)w = q(x, y) \quad \text{on } \mathfrak{R}, \quad (6)$$

where $p(x, y) = a^{-1/2}\Delta(a^{1/2})$ and $q(x, y) = a^{-1/2}f$. The effect of this scaling is to transform the operator $\mathfrak{L}$ into one whose differential part is $-\Delta$. Note that the change of variable (5) does not alter the positive definiteness of $\mathfrak{L}$, so that $\mathfrak{M}$ is positive definite as well.

Substitution of (6) into (2) then yields as the iteration

$$(-\Delta + K)w_{n+1} = (-\Delta + K)w_n - \tau(-\Delta + p)w_n + \tau q \text{ on } \mathfrak{R}. \quad (7)$$

The boundary condition is

$$w_{n+1} = H(x, y) \quad \text{on } \partial\mathfrak{R}, \quad (8)$$

where $H(x, y) = a^{1/2}g$.

In an attempt to make the operator $-\Delta + K$ on the left of (7) agree closely with $\mathfrak{M}$, the constant K is chosen to approximate $p(x, y)$. The choice of central interest in this study is the minimax value,

$$K = (\beta + B)/2, \quad (9)$$

where β is the minimum and B the maximum value of $p(x, y)$ on the closed rectangle. As will be shown in the next section, this choice leads to an estimate that the optimal value of the single parameter τ to give most rapid convergence in (7) is

$$\tau = 1. \quad (10)$$

For this value of τ, (7) becomes simply

$$(-\Delta + K)w_{n+1} = (K - p)w_n + q \quad \text{on } \mathfrak{R}. \quad (11)$$

The discrete form of the iterative procedure (8, 9, 11) is obtained by placing a uniform rectangular mesh on $\mathfrak{R}$ with spacing h in the x-direction and k in the y-direction and letting W_{ij} correspond to $w(x, y)$ at the mesh points $x = ih$, $y = jk$. Using the standard five point approximation for the operator $-\Delta$ with Dirichlet boundary conditions

$$-\Delta_h W_{ij} \equiv h^{-2}(-W_{i-1,j} + 2W_{ij} - W_{i+1,j}) + k^{-2}(-W_{i,j-1} + 2W_{ij} - W_{i,j+1}), \quad (12)$$

$$i = 1, 2, \ldots, \frac{c}{h} - 1; \quad j = 1, 2, \ldots, \frac{d}{k} - 1,$$

one then obtains for (8, 11)

$$(-\Delta_h + KI)W^{(n+1)} = (KI - P)W^{(n)} + Q, \quad (13)$$

where P is a diagonal matrix with elements $P_{ij} = p(ih, jk)$, Q is a vector with elements $Q_{ij} = q(ih, jk)$, and I is the identity matrix. The solution of (13) is carried out in each iteration by using a fast direct method.

Finally, under the assumption that the eigenvalues of $(-\Delta_h + KI)^{-1}(KI - P)$ lie in the interval $[-\rho, \rho]$, Chebyshev acceleration is applied:[11]

$$\tilde{W}^{(n+1)} = \omega_{n+1}(W^{(n+1)} - \tilde{W}^{(n-1)}) + W^{(n-1)}, \quad (14)$$

where $\omega_0 = 1$, $\omega_1 = 2/(2 - \rho^2)$, $\omega_{n+1} = (1 - \rho^2\omega_n/4)^{-1}$ for $n = 1, 2, \ldots$, and $\tilde{W}^{(n+1)}$ is the improved value of $W^{(n+1)}$, where now $W^{(n+1)}$ satisfies (13) with $W^{(n)}$ replaced by $\tilde{W}^{(n)}$ on the righthand side. This is equivalent to the use in (7) of a sequence $\{\tau_n\}$, rather than a single value of τ, in a manner that is numerically stable and does not require the total number of parameters in the sequence to be specified in advance. If in some cases memory limitations preclude the use of (14), then a fixed sequence $\{\tau_n\}$ could be used instead, ordered in the manner recommended in Reference 12 for numerical stability.

Convergence properties

The convergence properties of the iterative technique can be examined by standard methods in terms of the eigenvalues of the Laplace operator, which are known explicitly for the rectangle. Consider the discrete form of the iteration (7, 8),

$$(-\Delta_h + KI)W^{(n+1)} = (-\Delta_h + KI)W^{(n)} - \tau[(-\Delta_h + P)W^{(n)} - Q], \quad (15)$$

in which K and τ are not yet specified to be the values (9) and (10). Assume that $K > -\lambda_m$, where λ_m is the smallest eigenvalue of $-\Delta_h$, so that $(-\Delta_h + KI)$ is positive definite. Assume also that the discretization of $\mathfrak{M}$ to $M \equiv -\Delta_h + P$ maintains the positive definiteness. Then one obtains, denoting by ν_m and ν_M the minimum and maximum eigenvalues of the generalized eigenvalue problem

$$M\Phi = \nu(-\Delta_h + KI)\Phi,$$

that the spectral radius ρ for iteration (15) is given by

$$\rho(I - \tau[-\Delta_h + KI]^{-1}M) = \text{Max}(|1 - \tau\nu_m|, |1 - \tau\nu_M|). \quad (16)$$

Since $\nu_m > 0$, there follows the well-known result[13] that *iteration* (15) *converges for any initial approximation* $W^{(0)}$ *if and only if* $0 < \tau < 2/\nu_M$, *and, for a single parameter* τ, *the optimal choice*

$$\tau = \tau_0 \equiv 2/(\nu_m + \nu_M) \tag{17}$$

yields the smallest spectral radius

$$\rho = \rho_0 \equiv (\nu_M - \nu_m)/(\nu_M + \nu_m). \tag{18}$$

The values of ν_m and ν_M can be estimated from the Rayleigh quotient for ν,

$$\frac{\Phi^T M \Phi}{\Phi^T(-\Delta_h + KI)\Phi} = 1 + \frac{\Phi^T(P - KI)\Phi}{\Phi^T(-\Delta_h + KI)\Phi}. \tag{19}$$

One obtains

$$1 + \min\left(\frac{\beta - K}{\lambda_m + K}, \frac{\beta - K}{\lambda_M + K}\right) \leq \nu_m \leq \nu_M \leq 1 + \max\left(\frac{B - K}{\lambda_m + K}, \frac{B - K}{\lambda_M + K}\right), \tag{20}$$

where λ_M is the largest eigenvalue of $-\Delta_h$.

The estimate for ρ obtained from (16) and (20) is least when a choice for K is made such that

$$\beta \leq K \leq B, \tag{21}$$

assuming $\beta > -\lambda_m$ holds. There results that for the corresponding optimal choice

$$\tau = \frac{2(\lambda_m + K)}{2\lambda_m + B + \beta} \tag{22}$$

there holds

$$\rho \leq \rho_u \equiv \frac{B - \beta}{2\lambda_m + B + \beta}. \tag{23}$$

The upper bound (23) on the spectral radius is essentially independent of mesh size, since λ_m is approximately equal to its limiting continuous equivalent of $\pi^2(c^{-2} + d^{-2})$ to order of the square of the mesh length. It is a simple matter to place a rigorous lower bound on λ_m and obtain from (23) the result that $\rho_u < 1$ and hence that convergence is guaranteed, for $\beta > -\lambda_m$.

It is of interest to compare (23) with the analogous spectral radius estimate for the iteration, without scaling and shifting, based on (1). For the latter case one obtains that for the optimal choice $\tau = 2/(\alpha + A)$ there holds

$$\rho \leq (A - \alpha)/(A + \alpha), \tag{24}$$

where $\alpha = \min a(x, y)$ and $A = \operatorname{Max} a(x, y)$ on the closed rectangle. The estimate (24) is independent of the mesh size and is the sharpest such one possible.

The presence of the $2\lambda_m$ term in the denominator of (23) can have the effect of there resulting a considerably smaller bound on ρ for the scaled and shifted iteration than results from (24) for iteration (1). Since (24) is essentially sharp such a smaller bound would imply a faster convergence rate. Thus one concludes that scaling and shifting are most effective when A/α is not especially close to one and a does not vary with excessive rapidity over the rectangle, in which case the resulting improvement in convergence rate could be substantial.

Remarks

Lower bound on β

It is required above that β, the minimum of $p(x, y)$ on the rectangle, satisfy $\beta > -\lambda_m$. In the case for which $\beta \leq -\lambda_m$ (the positive definiteness of M does not preclude P dipping below $-\lambda_m$ over a portion of the rectangle) the estimate (20) no longer yields an upper bound on ρ that is less than one, hence it does not guarantee convergence. In the numerical experiments performed on such cases, iteration (15) usually converged, but at a relatively slower rate. In general, the best candidates for the iterative procedure are those cases for which $\beta > -\lambda_m$.

Shift parameter

The choice of the particular value (9) for K out of the possible ones (21) yielding the best convergence rate estimate (23), corresponding to (22), is made for two reasons. One is that for the corresponding value $\tau = 1$, which is obtained from (22) for the shift (9), the resulting discrete Picard iteration (13) requires fewer computer operations than does the one for general τ (15). The other is that for this shift the actual convergence rate observed in numerical experiments is somewhat more rapid than it is for shifts near the end points of the interval $[\beta, B]$, at least for those problems for which $p(x, y)$ varies smoothly without rapid changes (see NUMERICAL EXAMPLES).

Calculation of P

In practice, an alternative to the analytic calculation of $p(x, y) = a^{-1/2}\Delta(a^{1/2})$ and its subsequent numerical evaluation to obtain the elements of P in (13) may be desirable. One could, instead, difference $a^{1/2}(x, y)$

TABLE I—Results after 5 iterations

	a(x,y)	K	τ	Chebyshev Acceleration	ρ_e	Maximum Error
		0	0.868	none	0.13	3.7(−5)
		0	0.868	using ρ_u	—	2.4(−6)
(a)	$[1+\frac{1}{2}(x^4+y^4)]^2$	3	1	none	0.039	3.9(−8)
		3	1	usingρ_u	—	1.1(−6)
		3	1	using ρ_e	—	4.3(−9)
		0	16/15	none	0.066	1.2(−6)
		0	16/15	using ρ_u	—	7.2(−8)
(b)	$[1+\sin\frac{1}{2}\pi(x+y)]^2$	$-\pi^2/8$	1	none	0.061	2.3(−7)
		$-\pi^2/8$	1	using ρ_u	—	3.2(−8)
		$-\pi^2/8$	1	using ρ_e	—	2.3(−8)
		0	0.829	none	0.31	3.4(−4)
		0	0.829	using ρ_u	—	5.9(−3)
(c)	$[2+\tanh 4(x+y-1)]^2$	4.07	1	none	0.26	2.6(−4)
		4.07	1	using ρ_u	—	1.5(−3)
		4.07	1	using ρ_e	—	3.4(−5)

directly to obtain approximate elements p_{ij}^h of P, $p_{ij}^h=\Delta_h a_{ij}^{1/2}/a_{ij}^{1/2}$, where $a_{ij}^{1/2}=[a(ih,jk)]^{1/2}$. The discretization error introduced by using p_{ij}^h instead of $p(ih,jk)$ would be of the same order as that already introduced by (12).

Spectral radius estimate

In applying Chebyshev acceleration (14) to iteration (13), one can either use the estimate (23) for the spectral radius or else obtain an estimate by observing the convergence rate when solving the problem first on a coarse grid. This latter procedure is often worth the small extra expenditure of computing effort, because the estimate (23) may be pessimistic and, since ρ is essentially independent of mesh size, the observed value usually is more accurate.

NUMERICAL EXAMPLES

Well-suited cases

The ideal case for the basic technique (13, 9) is one in which $p=a^{-1/2}\Delta(a^{1/2})$ is constant on the rectangle [e.g., $a=\cos^2(x+y)$, $a=J_0^2([x^2+y^2]^{1/2})$, etc.]. Then from (23) one obtains that the optimal spectral radius is $\rho=0$, hence the problem is solved completely (to round-off accuracy) in only one iteration. This result corresponds to the fact that in each iteration a Helmholtz equation (13) is solved directly.

Other highly suitable cases for the technique are those not departing strongly from the ideal one. The experimental results for two such cases are summarized in Table Ia, b. Both cases were solved numerically by using (15) on the unit square $0<x<1$, $0<y<1$ with uniform mesh spacing $h=k=2^{-l}$, for the values $l=4$, 5, and 6. (The number of rows of interior mesh points should be 2^l-1, l an integer, in at least one direction for fast direct methods to apply efficiently.)

The entries in Table I are the rounded values for a mesh with 64×64 interior points; for the other mesh sizes the values differed from these only slightly, if at all. A value of K equal to 0 or to $(\beta+B)/2$ was used, along with the corresponding value (22) for τ. When Chebyshev acceleration was included, either the estimate ρ_u from (23) or the experimentally observed estimate ρ_e was used to approximate the spectral radius ρ in (14) of $(I-\tau[-\Delta_h+KI]^{-1}[-\Delta_h+P])$. The entries for the value of ρ_e are the observed approximate limiting values of the ratio

$$||W^{(n)}-W^{(n-1)}||_\Delta/||W^{(n-1)}-W^{(n-2)}||_\Delta,$$

where $||W||_\Delta=[W^T(-\Delta_h+KI)W]^{1/2}$. The maximum error, which is listed in the last column, is the maximum of the differences at the mesh points between $W^{(5)}$ and the solution. The initial maximum error had the value of approximately 1.

For the example in Table Ia, $\beta=0$ and $B=6$. Thus the estimate (23) for the optimal spectral radius is $\rho\leq\rho_u\approx0.132$ (using $2\pi^2$ for λ_m), and the shift (9) is $K=3$. For the example in Table Ib, one has $\beta=-\pi^2/4$,

$B=0$, and $\rho_u \approx 1/15$. In this case, the improvement obtained by using the shift $K=(\beta+B)/2$, instead of $K=0$, is not so great as it is for example Ia.

The effect of scaling and shifting can be found by comparing the results for these two examples with the estimate (24). For both there holds $\alpha=1$ and $A=4$, so that the spectral radius estimate without scaling and shifting in each case is 0.6.

Less well-suited cases

For the example summarized in Table Ic, $p(x, y)$ deviates more strongly from the ideal case. The task of calculating the actual extremal values of $p(x, y)$ on $\mathcal{R}$ was not carried out for this example; instead, the discrete equivalents $\beta=\beta_h=\min P_{ij}$, $B=B_h=\max P_{ij}$ were used. For the 64×64 mesh, $\beta_h \approx -9.62$ and $B_h \approx 17.77$, for which $\rho_u \approx 0.575$. Note that here $K=0$ does not correspond to an end point of the interval $[\beta, B]$.

An investigation of the possible non-sharpness of estimate (20) and non-optimality of (9) and (10), which are more important here than in a nearly ideal case, was carried out by fixing τ at the value one and observing the change in ρ_e as K was varied. A local minimum was found at approximately $K=3.0$, for which ρ_e is approximately 0.23.

For the more extreme case

$$a(x, y)=[2+\tanh 10(x+y-1)]^2,$$

in which the change in the value of a in crossing the line $x+y=1$ is very abrupt, β_h and B_h becomes approximately -60 and 111, respectively. In this case $\beta<-\lambda_m$; hence, the estimate (23) yields merely that $\rho \leq \rho_u > 1$. The iteration did converge, however, with the observed spectral radius $\rho_e \approx 0.63$ and a maximum error of 2.5×10^{-2} after five iterations for the usual test problem, with $K=(\beta_h+B_h)/2$ and $\tau=1$. With the inclusion of Chebyshev acceleration based on this value of ρ_e, the maximum error after five iterations was reduced to 6.3×10^{-3}. The value of ρ_e can be decreased in this case, with τ fixed at 1, to a locally minimum value of approximately 0.54 at approximately $K=14$.

Computational requirements

All the above experiments were carried out using the subroutine BUXYDY, written by B. L. Buzbee at Los Alamos Scientific Laboratory, which solves the Helmholtz equation on a rectangle using Buneman's algorithm for odd-even reduction.[4] The subroutine requires approximately 0.06 seconds on the CDC 7600 computer to solve a problem on a 64×64 mesh.

Qualitative comparison of the computational requirements of the technique with those of other methods can be made using the operation-count table given in Reference 3. One finds, for example, that for a 64×64 mesh the operations required for one iteration of (13) are equivalent to those required for about 4 or 4½ SOR iterations, and that about 85 SOR iterations are required to reduce the initial error by a factor $N^{-2} \approx 2.5\times10^{-4}$ (discretization error order) in the numerical solution of the Poisson equation when optimal parameters are used. The solution of (3) or (6) by SOR would generally require even more iterations.

The memory requirements of (13, 14) exceed those of SOR by about $3N^2$ locations if both $P-KI$ and $\tilde{W}^{(n-1)}$ are stored. This value can be reduced to N^2, however, in exchange for recomputing $P-KI$ at each iteration and using a form of Chebyshev acceleration that requires, instead of $\tilde{W}^{(n-1)}$, a sequence of parameters $\{\tau_n\}$.

One concludes that for well-suited cases, such as those in Table Ia, b, the basic technique is an extremely efficient one and compares very favorably with standard iterative and elimination methods. Its advantages are especially striking for problems with a large number of mesh points. For less well-suited problems, the technique may be very satisfactory in some cases, but further study would be helpful to clarify the best means for estimating the parameters.

EXTENSIONS

The iterative technique can be modified to handle more general equations and boundary conditions than those discussed here and to solve problems that are discretized on a mesh with non-uniform spacing.[14]

ACKNOWLEDGMENTS

This study was performed in collaboration with Gene H. Golub and is reported under our joint authorship in a larger paper, which includes further elaboration of many of the points discussed here.[14]

REFERENCES

1 O BUNEMAN
A compact non-iterative Poisson solver
Report 294 Stanford University Institute for Plasma Research Stanford California 1969

2 R W HOCKNEY
The potential calculation and some applications
Methods in Computational Physics vol 9 B Adler S Fernbach and M Rotenberg eds Academic Press New York and London 1969 pp 136–211

3 F W DORR
The direct solution of the discrete Poisson equation on a rectangle
SIAM Rev 12 1970 pp 248–263

4 B L BUZBEE G H GOLUB C W NIELSON
On direct methods for solving Poisson's equations
SIAM J Numer Anal 7 1970 pp 627–656

5 E G D'YAKONOV
On an iterative method for the solution of finite difference equations
Dokl Akad Nauk SSSR 138 1961 pp 522–525

6 O B WIDLUND
On the use of fast methods for separable finite difference equations for the solution of general elliptic problems
Sparse Matrices and Applications D J Rose and R A Willoughby eds Plenum Press New York 1972 pp 121–134

7 J E GUNN
The numerical solution of $\nabla a \nabla u = f$ by a semiexplicit alternating direction iterative method
Numer Math 6 1964 pp 181–184

8 J E GUNN
The solution of elliptic difference equations by semiexplicit iterative techniques
SIAM J Numer Anal 1 1965 pp 24–25

9 H L STONE
Iterative solution of implicit approximations of multi-dimensional partial differential equations
SIAM J Numer Anal 5 1968 pp 530–558

10 T DUPONT R P KENDALL H H RACHFORD JR
An approximate factorization procedure for solving self-adjoint elliptic difference equations
SIAM J Numeral Anal 5 1968 pp 559–573

11 R S VARGA
Matrix iterative analysis
Prentice-Hall Englewood Cliffs New Jersey 1962 p 141 prob 8

12 V I LEBEDEV S A FINOGENOV
On the order of choice of the iteration parameters in the Chebyshev cyclic iteration method
Zhur Vych Mat i Mat Fiz 11 1971 pp 425–438 English translation in Report CS72-304 Computer Science Department Stanford University Stanford California 1972

13 E L STIEFEL
Über einige methoden der relaxationsrechnung
Z angew Math Phys 3 1952 pp 1–33

14 P CONCUS G H GOLUB
Use of fast direct methods for the efficient numerical solution of nonseparable elliptic equations
To appear. Also available as Report 72-278
Computer Science Dept Stanford University
Stanford California 1972