UNCLASSIFIED

ERDA Mathematics and Computing Laboratory

Courant Institute of Mathematical Sciences
New York University

Mathematics and Computers COO-3077-122

HIGH ORDER FAST LAPLACE SOLVERS FOR THE DIRICHLET PROBLEM ON GENERAL REGIONS

Victor Pereyra*, Włodzimierz Proskurowski**, and Olof Widlund***

*Department of Applied Mathematics, California Institute of Technology, Pasadena, California. On leave of absence from Universidad Central de Venezuela, Caracas. The work of this author was supported in part by ERDA, AT(04-3)-326 and W-7405-ENG-48 while visiting Stanford University and the Lawrence Berkeley Laboratory.

**Department of Computer Science, Royal Institute of Technology, Stockholm 70, Sweden.

***Courant Institute of Mathematical Sciences, New York University. The work of this author was supported by ERDA, Contract No. E(11-1)-3077 at New York University.

Also to be issued as a report (LBL 4244) of the Lawrence Berkeley Laboratory.

Contract No. E(11-1)-3077
or E(11-1)-3077; E(04-3)-326

DISCLAIMER

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

Abstract

Highly accurate finite difference schemes are developed for Laplace's equation with the Dirichlet boundary condition on general bounded regions in R^n. A second order accurate scheme is combined with a deferred correction or Richardson extrapolation method to increase the accuracy. The Dirichlet condition is approximated by a method suggested by Heinz-Otto Kreiss. A convergence proof of his, previously not published, is given which shows that, for the interval size h, one of the methods has an accuracy of at least $O(h^{5.5})$ in L_2. The linear systems of algebraic equations are solved by a capacitance matrix method. The results of our numerical experiments show that highly accurate solutions are obtained with only a slight additional use of computer time when compared to the results obtained by second order accurate methods.

AMS (MOS) subject classifications. Primary 65B05, 65N15; secondary 65F05, 65N20.

§1. Introduction

It is the purpose of this paper to develop some highly accurate finite difference methods for the Dirichlet problem for a general bounded region Ω in R^n. The most accurate of these has an L_2 error of order at most $h^{5.5}$, see §4. Our basic schemes use the standard (2n+1)-point formula for the interior mesh points and are therefore only second order accurate. The increased accuracy is achieved by two steps of a deferred correction or Richardson extrapolation procedure. We also discuss the computer implementation of these methods in some detail.

The use of deferred correction and Richardson extrapolation methods is justified by finding asymptotic expansions of the error. Wasow [20] has shown that no useful expansions of this kind exist if the boundary condition is approximated to a low order of accuracy. An obvious remedy for this problem, already mentioned by Wasow, is to use higher order interpolation or extrapolation formulas at any irregular mesh point, i.e. a mesh point in the open set Ω which fails to have all its next neighbors in the closure of Ω. Volkov [19] proposed the use of high order one-dimensional Lagrange polynomials for this purpose. Because of the change of sign of the interpolation coefficients the matrix representing the difference scheme will then, in general, not be of positive type. The standard convergence proof based on a discrete maximum principle, see Forsythe and Wasow [7], will therefore generally not apply. But by allowing the use of values of the mesh functions many mesh lengths away from the boundary, Volkov succeeded in designing schemes with

diagonally dominant matrices. His schemes may however lead to an unacceptably small mesh size even for very simple geometries.

Numerical experiments, see Pereyra [13] and the last section of this paper, clearly demonstrate the need for higher order accuracy at the irregular mesh points if improved solutions through Richardson extrapolation or deferred correction methods are required. In his 1966 paper, Pereyra also reported on successful numerical experiments with methods based on Lagrange interpolation in one variable and employing only mesh points close to the boundary. At that time no convergence proof was known for such methods.

In June of 1968, Kreiss announced an interesting result on the convergence of methods of this type. His result was never published. His schemes are constructed as sums of difference approximations of one-dimensional problems. At the interior mesh points each of these problems is discretized by a three-point formula while at the irregular mesh points this basic formula is combined with high order Lagrange extrapolation. For a detailed description see §2. Kreiss found a method of proof which provides an alternative to the classical technique previously mentioned. His method depends heavily on the special structure just described.

We learned about his results from several conversations and his unpublished notes which were kindly made available to us. Our interest in these methods was recently renewed when we realized that the capacitance matrix, or imbedding, method developed by Proskurowski and Widlund [18] could be adapted for the difference schemes considered by Kreiss.

In this paper, we describe Kreiss' schemes, give detailed proofs of convergence and existence of error expansions and discuss their implementation. We have exclusively used a deferred correction method in our numerical experiments rather than Richardson extrapolation. Our reason is that the deferred correction method, especially for problems in several dimensions, has often proved less costly, see Pereyra [13] and also § 5 of this paper. One advantage is that, in contrast to Richardson extrapolation, deferred correction methods require only one mesh size. The capacitance matrix method allows us to solve the same system of linear equations repeatedly at an expense which decreases considerably once the first problem has been solved.

Our combination of a deferred correction and an imbedding method is quite convenient from a programming point of view. We have also developed a new, practical way of calculating the required difference approximations to the terms of the expansion of the truncation error. This method resolves a long-standing problem in the theoretical justification for the use of more than one deferred correction step for boundary value problems of this type. The imbedding of the region in a rectangle allows us to use certain programs previously developed to perform deferred corrections for problems on rectangular regions.

In the last section, we report on numerical experiments carried out on a CDC 7600 computer at the Lawrence Berkeley Laboratory. They show that very high accuracy is obtained for problems with sufficiently smooth solutions. For problems which fail to have sufficiently many bounded derivatives the corrections

do not spoil the accuracy of the solution. We believe that our method can be developed further into highly efficient and reliable numerical software. We note that fast Laplace solvers are used increasingly to enhance the convergence when solving more general problems, see for example, Bartels and Daniel [1], Concus and Golub [4,5], Jameson [9], O'Leary [10], Martin [11,12] and Widlund [21].

Acknowledgements

Thanks are due foremost to Heinz-Otto Kreiss for a number of conversations and for copies of his notes. Thanks are also due to Ole Hald and Vidar Thomée who read a draft of this paper and made helpful comments and to Paul Concus and Gene Golub for their interest and their hospitality in Berkeley and Stanford where a main part of our work was done.

§2. Kreiss' Method for Poisson's Equation

We will consider a family of finite difference schemes for the Dirichlet problem for Poisson's equation,

$$
\begin{aligned}
-\Delta u = -\sum_{i=1}^{n} (\partial/\partial x_i)^2 u &= f(x) , \quad x \in \Omega , \\
u(x) &= g(x) , \quad x \in \partial\Omega ,
\end{aligned}
\tag{2.1}
$$

where the region Ω is an open, bounded subset of the n-dimensional, real, Euclidean space R^n with the boundary $\partial\Omega$. We will make no detailed assumptions on the smoothness of $\partial\Omega$ and the data f and g but assume only that they are sufficiently smooth. As is well known, the problem (2.1) then has a unique, sufficiently smooth solution.

A uniform mesh R_h^n is introduced by

$$
R_h^n = \{x \in R^n \mid x_i = n_i h,\ n_i = 0, \pm 1, \pm 2, \ldots\} ,
$$

where $h > 0$ is the mesh size. The position of the origin of our mesh is, of course, arbitrary. We could also have chosen different uniform mesh sizes in the different coordinate directions without affecting the theory or practice of the methods except in some very minor ways.

The set of mesh points of interest to us is

$$
\Omega_h = \Omega \cap R_h^n .
$$

There are no equations for points on $\partial\Omega$. The difference equations are constructed as a sum of approximations of one-dimensional prob-

lems corresponding to the operators $-(\partial/\partial x_j)^2$, $j = 1,\ldots,n$. They are specified by defining a linear equation for each $x \in \Omega_h$. Let the vector e_i be the unit vector in the direction of the positive i-th coordinate axis. A mesh point $x \in \Omega_h$ is called regular if all its closest neighbors $x \pm he_i$, $i = 1,\ldots,n$, belong to Ω_h. For a regular mesh point, we simply use the standard centered difference approximation of each of the second derivatives. This results in the equation

$$2nu^h(x) - \sum_{i=1}^{n} (u^h(x+he_i) + u^h(x-he_i)) = h^2 f(x) .$$

This formula is combined with polynomial extrapolation of a fixed degree k for the remaining, irregular, mesh points of Ω_h. Let us thus suppose that $x \in \Omega_h$ but that $x + he_i \notin \Omega_h$ and that $x - he_i,\ldots,x-(k-1)he_i \in \Omega_h$. This last condition can always be satisfied for a smooth $\partial\Omega$ if h is chosen small enough. Denote by x_i^* the intersection of the boundary $\partial\Omega$ and the segment between x and $x+he_i$ and by $s\cdot h$ the distance between x_i^* and $x+he_i$. Thus $0 \leq s < 1$. A provisional value of $u^h(x+he_i)$ is now defined by the Lagrange interpolation formula,

$$\sum_{j=0}^{k} \alpha_j u^h(x-(j-1)he_i) = u(x_i^*) = g(x_i^*) . \tag{2.2}$$

The coefficients α_j depend only on s and are given by the formula

$$\alpha_j = \prod_{\substack{\ell=0 \\ \ell \neq j}}^{k} (s-\ell)/(j-\ell) .$$

The value of u^h at the point $x+he_i$ is now eliminated by combining

(2.2) with the standard three point formula for the point x. The resulting matrix, which corresponds to the approximation of $-(\partial/\partial x_i)^2$ along a mesh line parallel to e_i, thus typically has the form

(2.3)

$$\begin{pmatrix} (2+\alpha_1'/\alpha_0'), (-1+\alpha_2'/\alpha_0'), \alpha_3'/\alpha_0', \ldots, \alpha_k'/\alpha_0' & & & & & & \\ -1 & 2 & -1 & \ldots & & & \\ 0 & -1 & 2 & \ldots & & & \\ \ldots & & & & \ldots & 2 & -1 & 0 \\ & & & & \ldots & -1 & 2 & -1 \\ & & & & \alpha_k/\alpha_0, \ldots, \alpha_3/\alpha_0, (-1+\alpha_2/\alpha_0), (2+\alpha_1/\alpha_0) \end{pmatrix}$$

Here $\alpha_0', \ldots, \alpha_k'$ are the Lagrange interpolation coefficients related to a second intersection between the boundary and the mesh line. If the mesh line in question intersects $\partial\Omega$ in several points, the matrix representing the difference approximation of $-(\partial/\partial x_i)^2$ along this line will be a direct sum of several matrices of the form (2.3). The matrix A_i which corresponds to the entire approximation of $-(\partial/\partial x_i)^2$ is the direct sum of the matrices introduced for the individual mesh lines parallel to the vector e_i. Finally, the matrix A, which represents the approximation of the entire problem (2.1), is the sum of $P_i^T A_i P_i$ where P_i is a suitable permutation matrix.

We note that if some irregular mesh point x is very close to the boundary, i.e. some s is quite close to one, the ratio α_1/α_0 will become very large. This will give the matrix a very large diagonal element and the coefficient multiplying $g(x_1^*)$ in the right-hand side will be of the same order of magnitude. In practice, we will therefore always scale the rows of the matrix A, making the diagonal elements equal to 2n.

§3. Stability of the Finite Difference Methods

As we saw in §2 the matrix A which corresponds to the full difference approximation of problem (2.1) has the form

$$A = \sum_{i=1}^{n} P_i^T A_i P_i$$

where the P_i are suitable permutation matrices and the A_i are direct sums of matrices of the form (2.3). The original problem (2.1) has a bounded inverse in L_2. The analogous result is that the spectral norm of the inverse of A is bounded by const $\times h^{-2}$. To establish this result we will study the symmetric part of A. In this section we will use the Euclidean vector norm and the spectral matrix norm exclusively.

Lemma 1. Let the symmetric part of a matrix A satisfy

$$(A + A^T)/2 \geq \delta I , \quad \delta > 0 .$$

Then A is nonsingular and $|A^{-1}| \leq 1/\delta$.

Proof: Let $Ax = b$. Then

$$\delta x^T x \leq x^T (A + A^T) x/2 = (x^T b + b^T x)/2 \leq |b| \cdot |x| .$$

Thus $|x| \leq |Ax|/\delta$ which proves the lemma.

Lemma 2. Let $A = \sum_{i=1}^{n} P_i^T A_i P_i$ where the P_i are permutation matrices. If

$$(A_i + A_i^T)/2 \geq \delta I$$

then

$$(A + A^T)/2 \geq n\delta I .$$

The proof follows from an elementary variational argument. The proof of the next lemma is equally easy.

Lemma 3. Let the matrix A_i be the direct sum of certain matrices B_{ij}. If

$$(B_{ij} + B_{ij}^T)/2 \geq \delta I \ , \quad \text{for all } j \ ,$$

then

$$(A_i + A_i^T)/2 \geq \delta I \ .$$

We are now ready to apply these lemmas. Specifically we will study matrices of the form (2.3). For technical reasons we will assume that all these matrices have an order of at least 2k-1. This condition can again be satisfied for any smooth $\partial\Omega$ if the mesh size h is chosen small enough. We will reduce the study of the matrix (2.3) to a simpler case which corresponds to imposing a Neumann condition at one end point.

Lemma 4. Denote by B the matrix defined by formula (2.3). Let B_2 be a matrix of the form

$$B_2 = \begin{pmatrix} 1 & -1 & & & & & & \\ -1 & 2 & -1 & & & & & \\ & \cdot & \cdot & \cdot & & & & \\ & & \cdot & \cdot & \cdot & & & \\ & & & \cdot & \cdot & \cdot & & \\ & & & & -1 & 2 & -1 & \\ & & & & \alpha_k/\alpha_0, \ldots, \alpha_3/\alpha_0, & (-1+\alpha_2/\alpha_0), & (2+\alpha_1/\alpha_0) & \end{pmatrix}$$

and let B_1 be a matrix of the same form generated by $\alpha_0', \ldots, \alpha_k'$. Suppose further that the orders of the matrices B_1, B_2 and B, denoted by n_1, n_2 and m respectively, satisfy the conditions,

$$n_1 \geq k , \quad n_2 \geq k , \quad m = n_1 + n_2 - 1 .$$

If

$$(B_1 + B_1^T)/2 \geq \delta I$$

and

$$(B_2 + B_2^T)/2 \geq \delta I$$

then

$$(B + B^T)/2 \geq \delta I .$$

Proof: Denote by $\tilde{B}_1$ the matrix obtained from B_1 by reversing the order of its rows and columns. The proof follows from the identity,

$$x^T(B + B^T)x/2 = u^T(\tilde{B}_1 + \tilde{B}_1^T)u/2 + v^T(B_2 + B_2^T)v/2 ,$$

where $u^T = (x_1, \ldots, x_{n_1})$ and $v^T = (x_{n_1}, \ldots, x_m)$. This identity can be verified straightforwardly. Hence, by our assumption,

$$x^T(B + B^T)x/2 \geq \delta(u^T u + v^T v) .$$

To conclude the proof we only note that

$$u^T u + v^T v \geq x^T x .$$

We will next use the LDL^T factorization of $S = (B_2 + B_2^T)/2$ to verify that S is positive definite and also give a lower bound for its eigenvalues. We will write S as a block matrix,

$$(3.1) \qquad S = \begin{pmatrix} S_{11} & s_{21}^T \\ s_{21} & s_{22} \end{pmatrix},$$

where $s_{21} = (0,\ldots,0,\ \alpha_k/2\alpha_o,\ldots,\alpha_3/2\alpha_o,\ -1+\alpha_2/2\alpha_o)$ and $s_{22} = 2+\alpha_1/\alpha_o$. Its block factorization takes the form

$$S = \begin{pmatrix} L_{11} & 0 \\ \ell & 1 \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & d \end{pmatrix} \begin{pmatrix} L_{11}^T & \ell^T \\ 0 & 1 \end{pmatrix},$$

where

$$L_{11} = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix},$$

is bidiagonal,

$$\ell = s_{21}L_{11}^{-T} = (0,\ldots,0,\alpha_k/2\alpha_o,(\alpha_k+\alpha_{k-1})/2\alpha_o,\ldots,$$

$$(\alpha_k+\ldots+\alpha_3)/2\alpha_o,\ -1+(\alpha_k+\ldots+\alpha_2)/2\alpha_o)\ ,$$

and

$$d = s_{22} - \ell\ell^T = 2+\alpha_1/\alpha_o - [(\alpha_k/2\alpha_o)^2+\ldots$$

$$+((\alpha_k+\ldots+\alpha_3)/2\alpha_o)^2+(-1+(\alpha_k+\ldots+\alpha_2)/2\alpha_o)^2]\ .$$

By using the fact that $\alpha_o+\ldots+\alpha_k = 1$, we find that

$$(3.2) \qquad d = 1/\alpha_o - [\alpha_k^2+(\alpha_k+\alpha_{k-1})^2+\ldots+(\alpha_k+\ldots+\alpha_2)^2]/4\alpha_o^2\ .$$

Computer results show that the rational function d is strictly

positive for $0 \leq s \leq 1$ and all $1 \leq k \leq 6$. For $k = 7$ and 8 it changes sign in the interval. These results can of course also be verified by a tedious paper and pencil calculation. We note that d goes to positive infinity when s approaches 1 while the components of s_{21} and ℓ remain bounded. We are now ready to establish a lower bound for the eigenvalues of S.

Lemma 5. Let d_{min} denote the minimum of the function d(s) defined by formula (3.2). Then there exists a strictly positive constant C, independent of the mesh size h and the region Ω, such that

$$S \geq \delta I$$

where

$$\delta = Cd_{min}h^2/(\text{diameter } (\Omega))^2 .$$

Proof: By using the notations previously introduced in this section, we find

$$x^TSx = x^TLDL^Tx \geq \min (d_{min}, 1)|L^Tx|^2 .$$

Since $d = 1$ for $s = 0$, $x^TSx \geq d_{min}|L^Tx|^2$. To obtain a lower bound for $|L^Tx|$ we will compute an upper bound for $|L^{-T}y|$. Partitioning the vector so that $y^T = (\tilde{y}^T, y_n)$, we find

$$y^TL^{-1} = ((\tilde{y}^T - y_n\ell)L_{11}^{-1}, y_n) .$$

Therefore, if we use the fact that ℓ has a uniformly bounded norm, we find

$$|L^{-T}y|^2 \leq |L_{11}^{-1}|^2(|\tilde{y}| + |y_n|\cdot|\ell|)^2 + y_n^2 \leq C(|L_{11}^{-1}|^2 + 1)|y|^2 .$$

The norm of L_{11}^{-1} equals the square root of the reciprocal of the smallest eigenvalue of $L_{11}L_{11}^T$. Now the matrix

$$L_{11}L_{11}^T = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \cdots & & & \\ & & & -1 & 2 \end{pmatrix}$$

has an order $m \leq \text{diam}\,(\Omega)/h$. As is easily checked, the smallest eigenvalue of $L_{11}L_{11}^T$ equals $4\sin^2(\pi/2(2m+1))$ corresponding to an eigenvector with the components $\cos(\pi(j-1/2)/(2m+1))$, $j = 1,\ldots,m$. This concludes the proof.

By combining our five lemmas and the results from our computation of d_{min}, we obtain, what essentially is Kreiss' result,

Theorem 1. For $k \leq 6$ there exist constants C_k, independent of h, such that,

$$|A^{-1}| \leq C_k(\text{diam}\,\Omega)^2 \times h^{-2} .$$

§4. Convergence and Asymptotic Expansions of the Error

In this section, we will prove the convergence of the schemes introduced in §2 and simultaneously establish asymptotic expansions for the error. We will concentrate on the case k = 6, which is the most accurate of the schemes known to be stable. We will assume throughout that the solution u(x) is sufficiently smooth. We make the Ansatz,

$$u^h(x) = u(x) + h^2 e^{(1)}(x) + h^4 e^{(2)}(x) + r^h(x) \quad . \tag{4.1}$$

The functions $e^{(1)}(x)$ and $e^{(2)}(x)$ will be chosen as solutions of Poisson's equation in a way which will make the remainder $r^h(x)$ a term of higher order.

Asymptotic expansions of this form are basic for the justification of Richardson extrapolation and deferred correction methods. They also easily enable us to give estimates for the rate at which difference quotients of the solution of the discrete problem $u^h(x)$ converge to the corresponding derivatives of the solution u(x).

Let us denote by $h^2 L_h$ the difference operator which has the matrix representation A, see §§2 and 3. The linear system of equations therefore has the form

$$h^2 L_h u^h = F^h \quad . \tag{4.2}$$

A component of the right-hand side F^h, which corresponds to a regular mesh point, has the form $h^2 f(x)$ whereas a component, corresponding to an irregular mesh point, is a sum of $h^2 f(x)$ and terms of the form $g(x_i^*)/\alpha_o(s_i)$. Here $\alpha_o(s_i)$, $0 \leq s_i < 1$, is a Lagrange

polynomial coefficient introduced in §2. To derive equations for the error functions $e^{(1)}(x)$ and $e^{(2)}(x)$, we substitute the expression (4.1) into the equation (4.2) and expand the truncation error in the customary way. We first ignore the contributions from the interpolation formulas used for the irregular mesh points. By setting the fourth and sixth order terms of the resulting expressions equal to zero, we obtain the Poisson equations

$$-\Delta e^{(1)} = (1/12) \sum_{i=1}^{n} (\partial/\partial x_i)^4 u$$

and

$$-\Delta e^{(2)} = (1/360) \sum_{i=1}^{n} (\partial/\partial x_i)^6 u + (1/12) \sum_{i=1}^{n} (\partial/\partial x_i)^4 e^{(1)} .$$

Because of the high order accuracy with which the boundary condition is approximated, it is appropriate to equip these equations with zero Dirichlet boundary conditions

$$e^{(1)}(x) = 0 , \quad e^{(2)}(x) = 0 , \quad x \in \partial\Omega .$$

The functions $e^{(1)}(x)$ and $e^{(2)}(x)$ are, by a standard result on elliptic equations, sufficiently smooth functions.

We now derive a difference equation for the remainder term $r^h(x)$. This equation has the form

$$h^2 L_h r^h = G^h .$$

It is easy to show that a component of G^h, corresponding to a regular mesh point, is of the form,

$$h^8((1/20160)\sum_{i=1}^{n}(\partial/\partial x_i)^8 u + (1/360)\sum_{i=1}^{n}(\partial/\partial x_i)^6 e^{(1)}$$

$$+ (1/12)\sum_{i=1}^{n}(\partial/\partial x_i)^4 e^{(2)}) + O(h^{10}) \ .$$

A component of G^h, corresponding to an irregular mesh point is the sum of a term of this form and of one or more seventh order error terms of the Lagrange interpolation formulas (2.2). The latter terms are multiplied by factors $1/\alpha_o(s_i)$ which grow as const$/(1-s_i)$ when $s_i \to 1$. It can, however, be shown, by a straightforward calculation, that this increasing factor $1/\alpha_o(s_i)$ is fully compensated by a decreasing factor in the error bound for Lagrange interpolation, see Isaacson and Keller [8 , p. 190]. These components of G^h are therefore uniformly $O(h^7)$ for all sufficiently smooth solutions $u(x)$.

We are now ready to use Theorem 1 to obtain a bound for $r^h(x)$. It is natural to work with the norm,

$$\|r^h\|_2 = \Big(\sum_{x\in\Omega_h} h^n |r^h(x)|^2\Big)^{1/2} \ ,$$

for which the estimate of Theorem 1 still holds. We first estimate $\|G^h\|_2$. The components of G^h are $O(h^8)$ for the regular mesh points and $O(h^7)$ for the irregular mesh points. Since there are only of the order $h^{-(n-1)}$ irregular points, $\|G^h\|_2 = O(h^{7.5})$. We use Theorem 1 to establish,

Theorem 2. Let $u^h(x)$ be the solution of the finite difference scheme with $k = 6$ and let $u(x)$ be the sufficiently smooth solution of the differential equation (2.1). Then there exist two

sufficiently smooth functions $e^{(1)}(x)$ and $e^{(2)}(x)$ such that

$$u^h(x) = u(x) + h^2 e^{(1)}(x) + h^4 e^{(2)}(x) + r^h(x) , \quad x \in \Omega_h ,$$

where the L_2 norm of $r^h(x)$ is $O(h^{5.5})$.

Similar results hold for smaller values of k. We expect that Theorem 2 is not sharp. We conjecture that the remainder term should be of the form

$$r^h(x) = h^6 e^{(3)}(x) + O(h^7)$$

in the maximum norm. We are led to this conjecture by results, previously established by Bramble and Hubbard [2], for the operators of strictly positive type which result when k = 1 and 2. If the estimate of Theorem 2 can be sharpened in this way, we would be justified in applying Richardson extrapolation three times to obtain a seventh order accurate method.

§5. Methods of Increasing the Accuracy

Richardson extrapolation and deferred correction methods are available to improve the second order accuracy of the basic solution $u^h(x)$. We will again concentrate on the case $k = 6$. We will first discuss the Richardson extrapolation method which is simpler both conceptually and in terms of its implementation.

The solution is first found on a basic mesh Ω_{h_o} and then for a sequence of refined meshes Ω_{h_i}, where $h_i = h_o/r_i$, $1 < r_1 < r_2 < \ldots$. It is very important that the sequence $\{r_i\}$ grows slowly for multidimensional problems since the number of variables grows rapidly. The improved solution is obtained only on the intersection of the meshes Ω_{h_i}. If we require the improved solution at all points of Ω_{h_o} and use two extrapolation steps, the number of mesh points on the finest mesh will be at least about nine (twenty seven) times larger in two (three) dimensions. Core storage can therefore easily be exhausted and less advantage can also be taken of the savings which often can be realized when direct methods are used to solve linear systems repeatedly.

If enough terms of an asymptotic error expansion, in even powers of h, exist, we obtain improved solutions $\tilde{u}_i^j$ by the recursion formula,

$$\tilde{u}_i^j = (\tilde{u}_i^{j-1} - (r_{i+j}/r_i)^2\, \tilde{u}_{i+1}^{j-1})/(1 - (r_{i+j}/r_i)^2)$$

with $\tilde{u}_i^o$ the restriction of u^{h_i} to the intersection of the meshes Ω_{h_i}. The error $\tilde{u}_i^j - u$ will be of the order h_o^{2j+2}. A useful a posteriori error bound,

$$\tilde{u}_i^j - u \approx (\tilde{u}_i^j - \tilde{u}_{i-1}^j)/(1 - (r_{i+j}/r_i)^2)\ ,$$

can also be computed, for details see Bulirsch and Stoer [3].

By using Theorem 2, we can easily show that two steps of Richardson extrapolation will give an accuracy of the order $h^{5.5}$, if we use the scheme with $k = 6$.

The deferred correction method requires only one mesh. The method has been discussed in detail in a number of papers, see for example Pereyra [13-17]. Here, approximations of the leading terms of the local truncation error of the discrete operator h^2L_h are computed and a corrected solution is then found by solving an additional system of linear equations with the same matrix A as before. Further corrections may be obtained in s similar way.

We will describe the variant of the method which we have used in our experiments. In the first step we take into account only the first truncation error terms, resulting from the approximation of $(\partial/\partial x_i)^2$, $i = 1,\ldots,n$, by the three point approximations. We know from §4 that these leading terms are

$$h^4(1/12)(\partial/\partial x_i)^4 u\ , \qquad i = 1,\ldots,n\ . \tag{5.1}$$

We attempt to approximate them to within $O(h^6)$ by using centered five point differences of the second order accurate solution u^h. For a periodic problem this procedure is very simple, but for a region with a boundary special one-sided differentiation formulas must be used for the mesh points which are within 2h of the boundary along a mesh line. One-sided formulas can introduce additional error terms for the corrected solution through the special contributions to the truncation error at the points where these formulas are employed. An additional correction step may be justified by an

asymptotic error expansion of the corrected solution, but note that an unfortunate choice of one-sided differentation formulas would lead to difficulties very similar to those already discussed by Wasow [20].

This problem can be avoided in a systematic way. Let x be the irregular mesh point introduced in our discussion in §2. We will use high order Lagrange extrapolation, employing only values of the mesh function u^h at x, $x-he_i,\ldots,$ to obtain provisional values of u^h at $x+he_i$ and $x+2he_i$. The same centered five point differentiation formula can then be used for all points in Ω_h, see further discussion in §6. The expansion of the truncation error which is due to the use of the five point approximation of the fourth derivative $(\partial/\partial x_i)^4$ will have the same leading terms and differ only in a higher order term. The order of this higher order term will of course depend on the degree of the Lagrange extrapolation polynomial. The Lagrange polynomial coefficients are the same at every point since we use values at mesh points only. The approximation of the expressions (5.1) are added to the original data F^h and the linear system of equations is solved a second time.

In a second correction step

$$(5.2)\qquad h^4(1/12)(\partial/\partial x_i)^4 u + h^6(1/360)(\partial/\partial x_i)^6 u\ , \quad i = 1,\ldots,n\ ,$$

is approximated by a centered seven point formula with an error which is $O(h^8)$ for a sufficiently smooth function. We thus use the once corrected solution and high order extrapolated values thereof in a way very similar to the previous step to obtain a new

right-hand side and a second corrected solution, see further discussion in §6.

Our error bounds for the deferred correction method are rather weak. When we estimate the truncation error due to the discretization of the expression (5.1), we find that the three first terms of the expansion given in Theorem 2 give a contribution of the order h^6. Since the operator h^2L_h has an inverse bounded by const h^{-2} they contribute a term of the order h^4 to the error of the corrected solution. In contrast the undivided differences of the remainder term of r^h create difficulties. Since undivided difference operators are bounded, independently of h, the contributions of r^h to the truncation error and the error of the corrected solution are bounded by $h^{5.5}$ and $h^{3.5}$, respectively. In order to prove a result as strong as that for the Richardson extrapolation method this loss of two powers of h must be eliminated. This would be achieved if we were able to give as sharp a bound for the norm of the second order divided differences of the solution as for the norm of the solution itself. The analogue of this desired estimate holds for second order elliptic equations on regions with sufficiently smooth boundaries. We have not been able to obtain this result in the discrete case. A modification of the argument of §3 leads to an improved bound for divided differences of the first order. This proves that at most one power of h can be lost in each connection step. For numerical evidence see §7.

§6. The Capacitance Matrix Method

All our experiments have been carried out for regions in the plane and we will therefore confine our discussion to that case. We have used a modification of the capacitance, or imbedding, method which was developed by Proskurowski and Widlund [18] to solve our linear systems of equations. We refer to that paper for a detailed discussion of the method. Here we will confine ourselves to a few brief remarks on the method concentrating on the changes required by the deferred correction algorithm.

A main part of any capacitance matrix program is a fast Poisson solver on a region for which separation of the variables can be applied. Our subroutine, SOLVE, implements a Fourier-Toeplitz method on an infinite parallel strip with periodic boundary conditions in one direction, see Proskurowski and Widlund [18, Section 6] and Fischer, Golub, Hald, Leiva and Widlund [6]. Our region Ω is imbedded in a rectangular subset of this strip. The fast solver requires of the order $mn \log_2 n$ operations for the exact solution of the five point discrete Poisson equation. Here n, the number of mesh points across the strip, is preferably a power of two and m is the number of mesh points used along the strip. We will see below that it is convenient to place the region Ω inside a centered subset, of size $(m-6)\times(n-6)$, of the set of $m \times n$ mesh points which is used by SOLVE.

An extended system of linear equations with a matrix

$$\tilde{A} = B + UZ^T$$

is solved. The matrix B corresponds to the five point formula on the

strip while $\tilde{A}$ contains our matrix A, see §§2 and 3, as a principal minor. The matrices U and Z are sparse and have p columns where p is the number of irregular mesh points. The matrix U is chosen so that Uv, v any p-vector, is an extension by zero of the corresponding mesh function v defined only on the set of irregular mesh points. The matrix Z^T is thus a compact representation of the matrix $\tilde{A}$-B from which zero rows have been eliminated. A change of the approximation of the boundary condition involves a change of the matrix Z. The right-hand side F^h of our original system of equations is extended, in an arbitrary way, to the complement of Ω_h. The matrix $\tilde{A}$ is constructed in such a way that the restriction of the solution of the extended system to the set Ω_h is the solution of our original system of equations.

There are two main parts of our capacitance matrix program. We execute the first one only once for a particular choice of h (a mesh size), k (a member of our family of difference schemes) and a region Ω. In this first part a $p \times p$ nonsymmetric dense capacitance matrix C is computed at an expense of one call of the subroutine SOLVE and of the order p^2 additional operations. A solution for a specific set of data, which is accomplished in the second part, requires essentially only two calls of the subroutine SOLVE and the solution of a capacitance matrix system of equations of the form $C\mu = \tilde{g}$. In our implementation the capacitance matrix C is very well conditioned and this equation can therefore be solved accurately by a conjugate gradient method at an expense of the order p^2 operations. We have however chosen to use Gaussian elimination. The matrix C is factored only once, at an

expense of the order p^3 operations, and the factors are then stored and used for any additional set of data. Any subsequent problem therefore requires only of the order $(mn \log_2 n + p^2)$ operations. The method is numerically very stable and the linear system of equations is solved very accurately.

Two rectangular arrays of the dimension $m \times n$ are used to store the data and the solution. The first array initially contains the original data F^h, arbitrarily extended to the complement of Ω_h. The second order accurate solution u^h is computed and stored in the second array. This solution is then extended to certain exterior points by extrapolation in the x_1-direction, see §5. A first contribution to the modified right-hand side of the equation is then computed by using a five point numerical differentiation formula on all mesh lines parallel with the x_1-axis. The resulting mesh function is added to F^h, the content of the first array. This process is now repeated in the other direction. We thus extrapolate $u^h(x)$ in the x_2-direction to the appropriate exterior mesh points and use a differentiation formula in the x_2-direction to obtain the final contribution to the new right-hand side. We note that we can simplify the programming by using the numerical differentiation formula over the entire rectangular region since the restriction of the solution on the strip to the set Ω_h is independent of the values of the data outside Ω_h. The second part of the capacitance matrix solver is now used, with the new right-hand side, to produce a fourth order accurate solution. It is stored in the second array which also serves as work space during this part of the calculation. The final corrected solution is

computed similarly. The original data F^h is read into the first array and approximations to the expressions in formula (5.2) are added. In this step seven point differentiation formulas are used. We note that since we placed Ω_h inside a rectangle, leaving three extra mesh lines on all sides, we can carry out all the necessary extrapolations while using only the storage locations provided for in the second $m \times n$ array. This admittedly introduces an additional constraint on the choice of mesh size for certain nonconvex regions but this aspect of the implementation of our method can of course easily be changed. The extrapolation and numerical differentiation steps are very straightforward and require very little computer time, see §7.

§7. Numerical Experiments

A FORTRAN program incorporating the ideas of this paper was prepared and run in single precision (between 14 and 15 decimal digits) on a CDC 7600 computer at the Lawrence Berkeley Laboratory using a RUN 76 compiler. We report on experiments using second and sixth order Lagrange interpolation formulas, $k = 2$ and 6, for the irregular mesh points, see §2. In all our experiments the region was a circle of radius one centered at the origin and the mesh size was $h = 1/23$. There were 1653 mesh points of which 128 were irregular and the region was imbedded in a 64×64 mesh.

By ε_∞ and ε_2 we denote the maximum and L_2 norms of the error, i.e.,

$$\varepsilon_\infty = \max_{x \in \Omega_h} |u^h(x) - u(x)| ,$$

and

$$\varepsilon_2 = \{(1/N) \sum_{x \in \Omega_h} |u^h(x) - u(x)|^2\}^{1/2} ,$$

where N is the number of points in Ω_h.

In Table 1, we report on the solution of

$$-\Delta u(x) = 2 \sin (x_1 + x_2)$$

with boundary values and exact solution equal to $u(x) = \sin (x_1 + x_2)$. This is a problem with a very smooth solution and served basically as a test that the program and algorithm really worked. We note that we obtain close to full word accuracy.

The next problem, see Table 2, was

$$-\Delta u(x) = 53 \sin (2x_1 - 7x_2)$$

with the boundary values and exact solution equal to $u(x) = \sin(2x_1 - 7x_2)$. This problem is more difficult than the first since the solution is more oscillatory. We tried sixth and second order interpolation at the irregular mesh point. According to results of Bramble and Hubbard [2] there is an expansion of the form

$$u^h(x) = u(x) + h^2 e^{(1)}(x) + O(h^3)$$

when second order interpolation, $k = 2$, is used. We note that the first correction step gives a smaller improvement in the case $k = 2$ than when $k = 6$ and that the second correction step gives no improvement for $k = 2$. This experiment thus confirms the observations of Wasow [20], Pereyra [13] and others on the importance of the existence of asymptotic error expansions. We also note that the two second order methods, obtained before the correction steps, perform equally well.

A final series of experiments were carried out to study the effects of lack of smoothness of the solutions. The problems had the form

$$-\Delta u = \begin{cases} -2\ell(\ell-1)(x_1 + x_2)^{\ell-2}, & \text{if } x_1 + x_2 \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

with the boundary values and exact solution equal to

$$u(x) = \begin{cases} (x_1 + x_2)^{\ell}, & \text{if } (x_1 + x_2) \geq 0 \\ 0, & \text{otherwise .} \end{cases}$$

We tried $\ell = 2$, 4 and 6. The solution then has a jump discontinuity

in derivatives of order ℓ. The results are given in Table 3. The performance of the method with $k = 2$, $\ell = 6$, is consistent with our previous observations. For $k = 6$ and with $\ell = 2, 4$ it appears as if a ℓ-th order accurate method is obtained for these solutions which have a jump in the ℓ-th derivatives. Care must of course be exercised when trying to draw such conclusions from our very limited experimental evidence. We feel however that our results are encouraging. We note that when the solutions fail to be smooth enough the corrections do not destroy the accuracy obtained in the previous steps.

The total CPU-time for a problem with $k = 6$ was 10.28 seconds. The first part of the capacitance matrix program, see §6, computed the second order accurate solution $u^h(x)$ in 8.77 seconds. The first correction required an additional 0.66 seconds and the second correction took an additional 0.85 seconds. In the correction steps the extrapolation to exterior mesh points and the differentiation steps required less than 10% of the time. The execution time could be reduced by optimizing our program and by changing to a faster compiler.

Correction	0	1	2
ε_∞, $k = 6$	1.9×10^{-5}	1.0×10^{-9}	5.6×10^{-12}
ε_2, $k = 6$	1.0×10^{-5}	5.4×10^{-10}	3.4×10^{-12}

Table 1

L_2- and maximum-norm errors for a problem with the solution $u(x) = \sin(x_1 + x_2)$. Sixth order interpolation used at the boundary points.

Correction	0	1	2
ε_∞, k = 2	8.8×10^{-3}	1.3×10^{-3}	1.4×10^{-3}
ε_2 , k = 2	4.7×10^{-3}	3.3×10^{-4}	3.4×10^{-4}
ε_∞, k = 6	9.2×10^{-3}	5.3×10^{-5}	1.3×10^{-5}
ε_2 , k = 6	4.8×10^{-3}	2.8×10^{-5}	3.4×10^{-6}

Table 2

L_2- and maximum-norm errors for a problem with the solution $u(x) = \sin(2x_1 - 7x_2)$. Second and sixth order interpolation are used.

Correction	0	1	2
ε_∞, ℓ = 2, k = 6	9.9×10^{-3}	9.9×10^{-3}	9.9×10^{-3}
ε_∞, ℓ = 4, k = 6	1.2×10^{-3}	4.8×10^{-6}	4.8×10^{-6}
ε_∞, ℓ = 6, k = 6	7.4×10^{-3}	9.4×10^{-7}	7.7×10^{-8}
ε_∞, ℓ = 6, k = 2	6.7×10^{-3}	1.5×10^{-3}	1.5×10^{-3}

Table 3

Maximum-norm error for a problem with the solution $u(x) = (x_1 + x_2)^\ell$, if $(x_1 + x_2) \geq 0$, $u(x) = 0$ otherwise. Sixth and second order interpolation are used.

References

[1] Bartels, R. and Daniel, J. W., "A Conjugate Gradient Approach to Nonlinear Elliptic Boundary Value Problems in Irregular Regions," Conference on the Numerical Solution of Differential Equations, Dundee, Scotland, July 1973, Lecture Notes of Mathematics, Springer, vol. 363, pp. 1-11.

[2] Bramble, J. H. and Hubbard, B. E., "Approximation of Derivatives by Finite Difference Methods in Elliptic Boundary Value Problems," Contr. Diff. Eqns., vol. 3, 1964, pp. 399-410.

[3] Bulirsch, R. and Stoer, J., "Fehlerabschätzungen und Extrapolation mit rationalen Funktionen bei Verfahren vom Richardson-Typus," Numer. Math., vol. 6, 1964, pp. 413-427.

[4] Concus, P. and Golub, G. H., "The Use of Fast Direct Methods for Efficient Numerical Solution of Nonseparable Elliptic Equations," SIAM J. Numer. Anal., vol. 10, 1973, pp. 1103-1120.

[5] Concus, P. and Golub, G. H., "A Generalized Conjugate Gradient Method for Nonsymmetric Systems of Linear Equations," Proc. 2nd Int. Symp. on Computing Methods in Applied Sciences and Engineering, IRIA, Paris, Dec. 1975 (to appear). Computer Science Dept. Stanford Report, 1976, CS-76-535.

[6] Fischer, D., Golub, G., Hald, O., Leiva, C. and Widlund, O., "On Fourier-Toeplitz Methods for Separable Elliptic Problems," Math. Comp., vol. 28, 1974, pp. 349-368.

[7] Forsythe, G. E. and Wasow, W. R., "Finite-Difference Methods for Partial Differential Equations," Wiley, 1960.

[8] Isaacson, E. and Keller, H. B., "Analysis of Numerical Methods," Wiley, 1966.

[9] Jameson, A., "Accelerated Iteration Schemes for Transonic Flow Calculations Using Fast Poisson Solvers," N.Y.U. ERDA Report C00-3077-82, 1975.

[10] O'Leary, D. P., "Hybrid Conjugate Gradient Algorithms for Elliptic Systems," Computer Science Dept. Stanford Report, to appear.

[11] Martin, E. D., "Progress in Application of Direct Elliptic Solvers for Transonic Flow Computations," to appear in Aerodynamics Analyses Requiring Advanced Computers, NASA SP-347, 1975.

[12] Martin, E. D., "A Fast Semidirect Method for Computing Transonic Aerodynamic Flows," to appear in Proceedings of the AIAA 2nd Computational Fluid Dynamics Conference, June 1975.

[13] Pereyra, V., "Accelerating the Convergence of Discretization Algorithms," MRC. Tech. Rep. No. 687, Univ. of Wisconsin, 1966; SIAM J. Numer. Anal., vol. 4, 1967, pp. 508-533.

[14] Pereyra, V., "Iterated Deferred Corrections for Nonlinear Operator Equations," Numer. Math., vol. 10, 1967, pp. 316-323.

[15] Pereyra, V., "Iterated Deferred Corrections for Nonlinear Boundary Value Problems," Numer. Math., vol. 11, 1968, pp. 111-125.

[16] Pereyra, V., "Highly Accurate Numerical Solution of Casilinear Elliptic Boundary Value Problems in n Dimensions," Math. Comp., vol. 24, 1970, pp. 771-783.

[17] Pereyra, V., "High Order Finite Difference Solution of Differential Equations," Computer Science Dept. Stanford Report, 1973, CS-73-348.

[18] Proskurowski, W. and Widlund, O., "On the Numerical Solution of Helmholtz's Equation by the Capacitance Matrix Method," ERDA-NYU Report, 1975. Math. Comp., 1976, to appear.

[19] Volkov, E. A., "An Analysis of One Algorithm of Heightened Precision of the Method of Nets for the Solution of Poisson's Equation," (Russian) Vych. Matem., vol. 1, 1957, pp. 62-80. AMS Trans. Series, vol. 2, 1964, pp. 117-136.

[20] Wasow, W., "Discrete Approximations to Elliptic Differential Equation," Z. Angew. Math. Phys., vol. 6, 1955, pp. 81-97.

[21] Widlund, O., "A Lanczos Method for a Class of Non-Symmetric Sysrems of Linear Equations," to appear.

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Administration, nor any person acting on behalf of the Administration:

A. Makes any warranty or representation, express or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Administration" includes any employee or contractor of the Administration, or employee of such contractor, to the extent that such employee or contractor of the Administration, or employee of such contractor prepares, disseminates, or provides access to, any information pursuant to his employment or contract with the Administration, or his employment with such contractor.