Title

FAST POISSON SOLVERS ON GENERAL TWO DIMENSIONAL REGIONS FOR THE DIRICHLET PROBLEM

Permalink

https://escholarship.org/uc/item/7ck6d810

Author

Shieh, A.S.L.

Publication Date

1977-12-01

UC-32

To be Submitted for Publication

LBL-4669 c.1
Preprint

FAST POISSON SOLVERS ON GENERAL TWO DIMENSIONAL REGIONS FOR THE DIRICHLET PROBLEM

A. S. L. Shieh

December 1977

Prepared for the U. S. Department of Energy
under Contract W-7405-ENG-48
also Issued as Mathematics Research Center,
University of Wisconsin--Madison,
Technical Summary Report No. 1820.
U. S. Army under Contract No. DAAG29-75 D-0024

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBL-4669

FAST POISSON SOLVERS ON GENERAL TWO DIMENSIONAL REGIONS FOR THE DIRICHLET PROBLEM

A. S. L. Shieh

December 1977

ABSTRACT

It is shown that by using the simplest construction of discrete dipoles, the operation count for solving the Dirichlet problem of Poisson's equation by the capacitance matrix method does not exceed constant times $n^2 \log n$, $n = 1/h$ for certain first and second order schemes of interpolating boundary conditions.

SIGNIFICANCE AND EXPLANATION

The Dirichlet problem for the Poisson equation is the following: Given a function f and a function g, find a function u such that

$$\begin{aligned} u_{xx} + u_{yy} &= f \quad \text{on} \quad \Omega , \\ u &= g \quad \text{on} \quad \partial\Omega . \end{aligned}$$

Here Ω is a simply connected domain with boundary $\partial\Omega$.

The problem has wide applications in electrostatics, elasticity, temperature distributions and plasma physics. Its solution by finite difference or finite elements methods have received considerable attention. It is known that if Ω is a rectangle, then fast Fourier transform methods are very efficient in solving the linear system of equations arising from finite difference or finite element discretizations.

There seems to be no such short cut to the solution of these equations when Ω is a general region. In many conventional methods, the operation count is usually proportional to $N^{3/2}$ (N is the number of mesh points in Ω) while at least N computer storage is required. These methods are therefore undesirable when N is very large. In this paper we describe an algorithm and prove mathematically that the operation count of this algorithm can be proportional to $N \log N$. While some versions of our algorithm also require at least N computer storage, there is one version that requires less than $N/3$ computer storage.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

FAST POISSON SOLVERS ON GENERAL TWO DIMENSIONAL REGIONS FOR THE DIRICHLET PROBLEM

A. S. L. Shieh

§1. Introduction

Over the past ten years, very fast numerical methods have been developed to solve Poisson's or Helmholtz's equation on certain simple regions with Dirichlet, Neumann or periodic boundary conditions. See e.g. [2], [3], [8], [9], [12], [19] and [21]. These methods can only be used for regions and boundary conditions that allow for separation of the variables. Typical examples are Poisson's or Helmholtz's equations in Cartesian coordinates on rectangular regions with boundary conditions that do not change type along any of the sides of the rectangle. In these special cases, the operation count for solving the discrete problem is almost proportional to the number of mesh points.

The purpose of this paper is to establish similar results for the Poisson equation on general regions. In this work we are only concerned with finite difference schemes of first and second order accuracy for the Dirichlet problem on simply connected bounded domains with smooth boundaries. A formal discrete potential theory motivated by the classical potential theory is incorporated into the so-called capacitance matrix method. It is shown that by using the simplest construction of discrete dipoles in our Ansatz, it is possible to have an algorithm the operation count of which is proportional to $N^2 \log N$, where $h = 1/n$ is the mesh size. Some numerical results are given in section 9 and a brief survey of past work in this direction is given in section 8.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024 and the Energy Research and Development Administration.

§2. Certain results from classical potential theory.

We give only a very brief review of a few results of classical potential theory. For a detailed exposition see e.g. [10], [14], [25] and [28]. We define the potential V resulting from a charge distribution ρ on a smooth boundary curve $\partial\Omega$ by

$$V(x) = (1/\pi)\int_{\partial\Omega} \rho(\xi) \log\pi \, ds(\xi) \ .$$

Here $x = (x_1,x_2)$, $\xi = (\xi_1,\xi_2)$ and $\pi^2 = (x_1-\xi_1)^2 + (x_2-\xi_2)^2$. The Green's function $(1/2\pi) \log \pi$ which we shall denote by G^* satisfies

$$\Delta(1/2\pi) \log r = \delta(x) \ ,$$

where $\delta(x)$ is the delta function. Similarly the potential W of a dipole density μ on $\partial\Omega$ defined by

$$W(x) = (1/\pi)\int_{\partial\Omega} \mu(\xi) \, \partial G^*/\partial\nu_\xi \, ds(\xi) \ . \tag{2.1}$$

We adopt here the convention that the normal direction of $\partial\Omega$ is towards the exterior of the region Ω in which we want to solve our problem.

The interior Dirichlet problem can be reduced to a Fredholm integral equation of the second kind if we make the double layer Ansatz as follows. Let

$$u(x) = -(1/2\pi) \iint_{\Omega} f(\xi) \log r \, d\xi + (1/\pi) \int_{\partial\Omega} \mu(\xi) \, \partial G^*/\partial\nu_\xi \, ds(\xi) = u_s(x) + W(x) \ ,$$

for the solution of

$$\begin{aligned} -\Delta u &= f, \quad x \in \Omega \\ u &= g, \quad x \in \partial\Omega \ . \end{aligned} \tag{2.2}$$

It can be shown that the dipole density μ satisfy the following integral equation

$$\mu + (1/\pi) \int_{\partial\Omega} \mu(\partial G^*/\partial\Omega_\xi) ds = g - u_s\big|_{\partial\Omega} = \tilde{g} \ . \tag{2.3}$$

This is a well posed problem of the form

$$(I + K)\mu = \tilde{g} \ , \tag{2.4}$$

where K is a compact operator defined by the integral above.

If we instead attempt to use a single layer Ansatz for the Dirichlet problem we obtain a Fredholm integral equation of the first kind. It has the form

$$V(x) = g - u_s\big|_{\partial\Omega} \ , \quad x \in \partial\Omega \ ,$$

which is an ill posed problem.

To illustrate the distribution of the eigenvalues of the compact operator K in equation (2.4), we study the case when Ω is an ellipse with

$\gamma = (a-b)/(a+b)$ where a and b are the half axes of the ellipse. It is known (see e.g. p. 135 of [25]) that $K = K^T$ and

(2.5) $$\lambda_i(K+K^T) = 2\gamma^{i-1} \quad , \ i = 1,2,\ldots \ .$$

On the other hand, both the interior and exterior Neumann problems can also be reduced to Fredholm integral equations of the second kind if we make the single layer Ansatz. The charge density ρ for the exterior Neumann problem satisfies

$$(I + K^T)\rho = g^* \ ,$$

for some suitably chosen function g^* defined on $\partial\Omega$. The existence and uniqueness problems for the solution of equation (2.3) can therefore be determined from that of equation (2.4) and vice versa. Finally, we remark that the G^* in equations (2.1) and (2.3) can be replaced by the Green's function on a sufficiently large rectangle with zero Dirichlet boundary conditions.

§3. The capacitance matrix method

In this section we develop a similar, formal potential theory for the discrete problems arising from the original Dirichlet problem (2.2). See also Sections 3 and 4 of [29] for a similar discussion. We shall assume that uniform mesh sizes in both coordinate directions are used.

We replace the Laplace operator by the five-point formula. The fundamental solution $(1/2\pi)\log(1/r)$, used in Section 2, will be replaced by its discrete analogue, the discrete Green's function on the entire plane, which we shall denote by G. Properties and efficient methods of generating G and its undivided differences will be studied in Section 4. An efficient method of computing Gv for arbitrary $N \times N$ vectors v is also given in Section 4. We will denote by B the matrix representing the five-point discrete Laplacian $h^2\Delta_h$, using undivided differences, on the entire plane. We then divide the set of mesh points into three disjoint sets $\Omega_h, \partial\Omega_h$ and $(C\Omega)_h$. The set $\partial\Omega_h$ contains all the irregular mesh points in Ω, i.e. mesh points that do not have all four neighbours within the open set Ω. Ω_h is the set of regular mesh points inside Ω and $(C\Omega)_h$ contains the remaining, the exterior mesh points.

We then set up the matrix equation

$$Au = v \tag{3.1}$$

that we are solving as follows. We use the same discretization formula for both A and B on $\Omega_h \cup (C\Omega)_h$. For points in $\partial\Omega_h$, a linear combination of the discrete Laplacian and interpolation formulas of first or second order accuracy for the boundary conditions are used. The values of the solution at the exterior mesh points are always eliminated from the discrete Laplacian, centered at an irregular mesh point. This guarantees that A is a reducible matrix with no couplings to the exterior mesh points from the irregular mesh points. If P is a suitably chosen permutation matrix, then

$$P\,A\,P^T = \begin{bmatrix} A_{11} & 0 \\ & \\ A_{21} & A_{22} \end{bmatrix}$$

where A_{11} is the coefficient matrix for our discrete problem on $\Omega_h \cup \partial\Omega_h$. It is easily seen that the solution on $\Omega_h \cup \partial\Omega_h$ will not be influenced by either the solution or the data on $(C\Omega)_h$.

The matrix A differs from B by only m rows where m is the cardinal number of $\partial\Omega_h$. We can therefore write

$$A = B + U W^T,$$

where the matrices U and W have m columns. The matrix U represents an extension operator. It maps any mesh function defined only on $\partial\Omega_h$ into a function on all mesh points. Its transpose, U^T, is a trace operator mapping any mesh function defined for all mesh points into its restriction to $\partial\Omega_h$. We easily verify that

$$W^T = U^T(A-B).$$

We now describe our method for solving the discrete problem (3.1). Guided by the continuous analog we make the Ansatz,

$$u = Gv + GV\mu . \tag{3.2}$$

The vector Gv satisfies $BGv = v$. The m-vector μ is determined by solving a system of $m\times m$ linear equations derived below. The mesh function $V\mu$ should vanish on Ω_h. Each column of the matrix V represents a discrete dipole of unit strength. Let such a column corresponding to $P \in \partial\Omega_h$ be regarded as a mesh function, denoted by $V^T(P)$. We require that $(1/h_\alpha)V^T(P)\, u = [\partial u/\partial\nu](P) + O(h)$. Here $h_\alpha = h/\cos\alpha$ where α is the angle between the normal through the irregular mesh point P and the closest coordinate axis. In particular, if that the western and northwestern neighbours of P in $(C\Omega)_h$, then

$$[V^Tu](P) = u(P) - (1 - \tan\alpha)\, u(w) - (\tan\alpha) u(NW). \tag{3.3}$$

We now use our Ansatz and compute the residual vector,

$$\begin{aligned} Au-v &= (B + UW^T)(GF + GV\mu) - Fv \\ &= (V + UW^TGV)\mu + UW^TGFv. \end{aligned} \tag{3.4}$$

From the properties of U and V, it follows that the residuals are zero for all $x \in \Omega_h$. To derive a linear system of equations for the vector μ we multiply equation (3.4) by U^T. It is easy to verify that $U^TU = I_m$ and $U^TV = I_m$. Here I_m is the $m\times m$ identity matrix. We thus obtain

$$(I_m + W^TGV)\mu = -W^TGF . \tag{3.5}$$

This choice of μ make the residuals zero for all $x \in \partial\Omega_h$. Hence substitution of μ in equation (3.2) will provide us a solution on $\Omega_h \cup \partial\Omega_h$ if equation (3.5) is solvable. Note that the residuals will in general not be equal to zero for all $x \in (C\Omega)_h$. The matrix on the left-hand side of Equation (3.5) is the capacitance matrix C. We shall refer to Equation (3.5) as the capacitance matrix equation.

In the special case when $v = U U^T v$, we can simply make the Ansatz $u = B^{-1} V\mu$. It is easily seen that the residual $Au-v$ will again be zero at $x \in \Omega_h$. The capacitance matrix equation now becomes

$$(3.6) \qquad C\mu = U^T v .$$

If Equation (3.6) is solvable, then $Au = v$ will also be zero on $\partial\Omega_h$. The solvability of Equations (3.5) and (3.6) will be discussed in Section 6.

We now describe our choices of difference equations at the irregular mesh points. We approximate the boundary conditions by interpolation schemes of first or second order accuracy, which we shall refer to as schemes Ia, Ib and II respectively.

We start with Scheme II. Let $P \in \partial\Omega_h$ and P^* be its closest point on $\partial\Omega$. Let W, E, N and S be the western, eastern, northern and southern neighbours of P on the mesh respectively. We assume that the local orientation of the boundary is such that either both W and N are in $(C\Omega)_h$ or only W is in $(C\Omega)_h$. Assume that both W and N are in $(C\Omega)_h$. Let d_1 denote h_1/h where h_1 is the distance, along a mesh line parallel with the x_1-axis, between the mesh point P and the boundary $\partial\Omega$. Hence $d_1 \in (0,1]$. The Dirichlet data at this point on $\partial\Omega$ is denoted by u_W. The values of d_2 and u_N are similarly defined. We then approximate u_W and u_N by $(1/2)[(1+d_1)u(W) + (1-d_1)u(E)]$ and $(1/2)[1+d_2)u(N) + (1-d_2)u(S)]$ respectively. By combining the above with the five-point formula for the Laplacian and eliminating $u(W)$ and $u(N)$ between them, we obtain

$$(3.6) \qquad \begin{aligned} 4u(P) &- [2d_1/(1+d_1)]u(E) - [2\,d_2/(1+d_2)]\,u(S) \\ &= h^2 f(P) + [2/(1+d_1)]u_W + [2/(1+d_2)]u_N . \end{aligned}$$

If only W is in $(C\Omega)_h$, then we obtain

$$(3.7) \qquad \begin{aligned} 4u(P) &- [2\,d_1/(1+d_1)]u(E) - u(N) - u(S) \\ &= h^2 f(P) + [2/(1+d_1)]u_W . \end{aligned}$$

We now describe the two variants of Scheme I, namely Scheme Ia and Ib. In Scheme Ib, if both W and N are in $(C\Omega)_h$, we obtain

$$(3.8) \qquad 4u(P) - u(S) - u(E) = h^2 f(P) + u_W + u_N .$$

If only W is in $(C\Omega)_h$, we obtain

(3.9) $$2[4u(P) - u(S) - u(E) - u(N)] = 2\ [h^2 f(P) + u_W] .$$

The scaling factor 2 is largely artifical and is put in only for the convenience of theoretical estimates in Sections 5 and 6. In Scheme Ia, we seek to eliminate this scaling factor while retaining the theoretical convenience. We require that Equation (3.8) should be used regardless of whether both W and N are in $(C\Omega)_h$ or only W is in $(C\Omega)_h$. The matrices A_{11} for all the above three schemes are of positive type. Hence, the results in [4] or [13] apply and all these schemes are convergent.

There is an important alternative to the above approach. Instead of the discrete Green's function of the entire plane, we may use the discrete analog of the Green's function on a sufficiently large square S with zero boundary conditions as our G in equations (3.2) and (3.5). In this case $G = B_D^{-1}$; $A = B_D + UW^T$. Here B_D denotes the matrix representing the discrete Laplacian $h^2\Delta_h$ on S and zero boundary values on the grid points of ∂S. The residual Au-v will again be zero on $\Omega_h \cup \partial\Omega_h$ if $C = U^T A B_D^{-1} V$ is nonsingular.

Finally we come to the central question as to whether the capacitance matrix equation (3.5) is closely related to the Fredholm integral equation (2.2)? It is known (see e.g. [16]) that the conjugate gradient method converges superlinearly for Fredholm integral equations of the second kind. In our experiments we normally fail to observe superlinear convergence. To understand this fully, we split up the matrices C into two parts as follows.

$$C = B_h + K_h .$$

The matrices B_k are defined by

(3.10) $$\begin{aligned} B_h(P,Q) &\equiv C(P,Q), \quad \text{if} \quad d(P,Q) \leq \sqrt{h} \\ &\equiv 0 , \quad\quad \text{otherwise} . \end{aligned}$$

They are therefore the near diagonal parts of C ; and the matrices K_h are the remaining parts, the off diagonal parts of C .

It will be shown in Section 6 that for Schemes La and Ib, and, after a suitable scaling, for schemes II, the matrices K_h are closely related to the compact integral operator K in Equations (2.2) or (2.3). The matrices B_h, however, will not in general be formal approximations to the identy operator. In fact, the algebraic row sums of B_h need not always be equal to one.

It is, however, shown in [16] that for operator equations with symmetric positive definite operators of the form $B + K$ with B symmetric positive definite and K symmetric compact, the conjugate gradient method will converge linearly with asymptotic rate of convergence governed only by the spectral condition number of B. We therefore proceed to study the special condition numbers of B_h in Section 5 and the distribution of singular values of K_h and C in Section 6. We shall show in Section 7 that the asymptotic convergence of the conjugate gradient method for solving the capacitance matrix equations will depend essentially on the spectral condition number of B_h.

We now discuss briefly two different methods of implementing our algorithm and the operation count involved. We use the conjugate gradient method to solve

$$C^T C\mu = C^T b,$$

where b denotes the right hand side of the capacitance matrix equation. The solution u is then computed from (3.2). In the first method, we generate the capacitance matrix explicitly. Assume that the G in equations (3.2) and (3.5) is the discrete Green's function on the entire plane. Because of translational invariance it is only necessary to compute G with the second parameter fixed at the origin. It is shown in section 4 that only one call of fast Poisson solved on a sufficiently large rectangle is needed to generate G and only two calls of a similar solver is needed to compute the final solution and the right hand side b. The operation count of the algorithm is therefore constant $N^2 \log N + 2 C_0 m^2 + O(m)$, where C_0 is the number of iterations needed to achieve a certain accuracy. If the G in equation (3.5) is B_D^{-1}, it is desirable to use the second method where the solution μ is computed by an iterative implicit method first appearing in [15]. The operation count for computing μ is proportional to $C_0(m+m_1)N$, where m_1 is the number of nonzero entries in the matrix V provided that a special fast solver is used in the process. See section 4 of [31] for details. It will be shown in section 7 that C_0 cannot exceed constant $\log m$ if $G = B_D^{-1}$ is used in (3.5) for all domains with sufficiently smooth boundaries and that C_0 is uniformly bounded in some special cases if the discrete Green's function on the entire plane is used in (3.5). The total operation count of our algorithm therefore does not exceed constant $N^2 \log N$.

§4. Properties and fast generation of G, the discrete Green's function on the entire plane, and the efficient computations of Gv, $GV\mu$ and W^TGv.

A discrete fundamental solution of the five-point Laplacian Δ_h with respect to the origin is a mesh function γ that satisfies

$$(4.1)\qquad \Delta_h\gamma(\mu h) = \begin{cases} h^{-2} & \text{if} \quad \mu = 0\,, \\ 0 & \text{if} \quad \mu \neq 0\,, \end{cases}$$

where μ has integer components μ_1 and μ_2.

Clearly γ is unique up to an arbitrary linear function. The constants involved will be chosen so that we have a proper discrete analog of the logarithmic potential. The resulting discrete fundamental solution will then be our discrete Green's function G.

It is established in [27] that if $g(r,s)$ denotes

$$(4.2)\qquad (2/\pi)\int_0^\pi [1-\cos(s\lambda)\ \exp(-|r|\mu)]/\sinh\mu\ d\lambda$$

where

$$(4.3)\qquad \cos\lambda + \cosh\mu = 2$$

$$(4.4)\qquad \mu/\lambda \to 1 \quad \text{as} \quad \mu \to 0,$$

then the function G defined by

$$(4.5)\qquad G(rk,sh) = (1/4)\ g(r,s) + (1/2\pi)\log h - (1/4\pi)(\log 8 + 2\gamma_1)$$

is the desired Green's function of the entire plane. Here γ_1 is the Euler's constant; r and s integers and h is the mesh size.

It is shown in [27] that

$$(4.6)\qquad g(0,0) = 0,\quad g(0,1) = 1$$

$$(4.7)\qquad g(r,s) = g(s,r) = g(-s,r) = g(s,-r),$$

$$(4.8)\qquad g(r,r) = (4/\pi)\ [1 + 1/3 + \ldots + 1/(2r-1)]\ ,$$

$$(4.9)\qquad g(r,s) = (1/\pi)\log(s^2+r^2) + (1/\pi)(\log 8 + 2\gamma_1) + o(1/r),$$

$$(4.10)\qquad g(r,s) - g(r,t) = (1/\pi)\log[(r^2+s^2)/(t^2+r^2)] + o(1/r).$$

We have found it necessary to obtain sharper estimates for the remainder term in (4.10) when $t = s+1$ and a similar estimate for $g(r+1,s) - g(r,s)$.

Theorem 4.1. Let r,s and t be nonnegative integers with $r \geq 8$; $s = t-1$. Then

$$(4.11a)\qquad G(rh,sh) - G(rh,th) = (1/4\pi)\log[(s^2+r^2)/(t^2+r^2)] + R_1(r,s) + R_2(r,s) + R_3(r,s),$$

where

(4.11b) $R_1(r,s) \equiv (1/24\pi) \int_0^\infty (\lambda^2 - r\lambda^3)\sin[(s + 1/2)\lambda]$

(4.11c) $R_2(r,s) \equiv -(1/2\pi) \int_0^\infty \{[(1/30)\lambda^4 + (7/96)r\lambda^5 + (1/288)r^2\lambda^6] \cdot e^{-r\lambda}\sin[(2s+1)\lambda/2]\}d\lambda,$

(4.11d) $|R_3(r,s)| \leq (0.37)r^{-1}e^{-1.35r^{2/3}} + (167)r^{-7}$.

Proof. We have

$$(\pi/2)[g(r,s)-g(r,t)] = \int_0^\pi [\cos(t\lambda) - \cos(s\lambda)]e^{-r\mu}[\sinh\mu]^{-1}\, d\lambda \ ,$$

and it is known that

$$\int_0^\infty [\cos(t\lambda) - \cos(s\lambda)]e^{-r\lambda}\lambda^{-1}d\lambda = (1/2)\log[(s^2+r^2)/(t^2+r^2)].$$

Since

(4.12) $$\int_\pi^\infty |\cos(t\lambda) - \cos(s\lambda)|\ e^{-r\lambda}\lambda^{-1}d\lambda < (2/\pi)r^{-1}e^{-r\pi},$$

it suffices to estimate

$$\int_0^\pi [\cos(s\lambda) - \cos(t\lambda)][e^{-r\lambda}\lambda^{-1} - e^{-r\mu}(\sinh\mu)^{-1}]d\lambda.$$

The integrand in the above expression will be denoted by J. Let $\varepsilon = (1.5)r^{-1/3}$, $r \geq 8$. We have

(4.13) $$\int_0^\varepsilon J\ d\ \lambda = \int_0^\varepsilon [\cos(s\lambda)-\cos(t\lambda)]e^{-r\lambda}[2\ \sin(\lambda/2)]^{-1}F(\lambda)d\lambda \ ,$$

where

(4.14) $$F(\lambda) \equiv 2[\sin(\lambda/2)]\lambda^{-1} - 2\ \sin(\lambda/2)[\sinh\mu]^{-1}e^{(\lambda-\mu)r} .$$

By (4.3) and (4.4), we have

(4.15) $$\sinh\mu = 2[\sin(\lambda/2)][1+\sin^2(\lambda/2)]^{1/2} ,$$

(4.16) $$e^{-\mu} = 2 - \cos\lambda - 2[\sin(\lambda/2)][1 + \sin^2(\lambda/2)]^{1/2} .$$

It is easily verified that for $0 \leq \lambda \leq \varepsilon$,

(4.17) $$e^{(\lambda-\mu)} = 1 + (1/12)(\lambda^3+\lambda^5) + (1/288)\lambda^6 + C_1(\lambda)\lambda^7,$$

with $|C_1(\lambda)| \leq 0.035$, $0 \leq \lambda \leq 1$. Hence,

$$e^{(\lambda-\mu)r} = 1 + (1/12)(r\lambda^3 + r\lambda^5) + (1/288)r^2\lambda^6 + C_2(r,\lambda) .$$

Here $|C_2(r,\lambda)| \leq (0.04)r\lambda^7 + (0.01)r^2\lambda^8 + (0.001)r^3\lambda^9$. Therefore,

$$F(\lambda) = (1/12)[\lambda^2 - r\lambda^3 - (7/8)r\lambda^5] - (1/30)\lambda^4 - (1/288)r^2\lambda^6 + C_3(r,\lambda).$$

Here $|C_3(r,\lambda)| \leq (0.01)\lambda^6 + (0.051)r\lambda^7 + (0.0104)r^2\lambda^8 + (0.001)r^3\lambda^9$.

Hence,

(4.18) $$\int_0^\varepsilon J d\lambda = 2\pi[R_1(r,s) + R_2(r,s)] + Z_1(r,s) + Z_2(r,s),$$

where (4.19)

(4.19) $$|Z_1(r,s)| < \int_0^\infty e^{-r\lambda}C_3(r,\lambda)d\lambda$$
$$< 1047\ r^{-7} .$$

(4.20) $\quad |Z_2(r,s)| \; < \; (1/6)r^{-1}\, e^{-1.5r^{2/3}}.$

It remains to estimate $\int_{\varepsilon}^{\pi} J\, d\lambda$.

Clearly,

(4.21) $\quad |\int_{\varepsilon}^{\pi} J d\lambda| \; \leq \; \int_{\varepsilon}^{\pi} (e^{-r\lambda} + e^{-r\mu}) d\lambda$.

It can be shown that $e^{-\mu}$ us a decreasing function of λ for $0 < \lambda < \pi$. See e.g. p. 33 of [33]. It is also shown there that

(4.22)
$$\begin{aligned} e^{-\mu} &< e^{-0.76}, & \lambda &= 0.8 \\ &< e^{-0.91}, & \lambda &= 1 \\ &< e^{-1.07}, & \lambda &= 1.2 \\ &< e^{-1.31}, & \lambda &= \pi/2. \end{aligned}$$

By (4.17),

$$e^{-\mu} = e^{-\lambda}(1 + C^*(\lambda)\lambda), \; |C^*(\lambda)| < 0.1, \; 0 < \lambda < 0.8,$$

so that

$$e^{-r\mu} \leq e^{-0.9\lambda r}, \quad 0 < \lambda < 0.8.$$

Hence,

$$\int_{0.8}^{\pi} e^{-r\mu}\, d\lambda < (0.32)e^{-0.76r} ,$$

and

(4.23) $\quad \int_{\varepsilon}^{\pi} [e^{-r\mu} + e^{-r\lambda}] d\lambda < (2.15)r^{-1}e^{-1.35r^{2/3}} - r^{-1}\, e^{-\pi r}$.

By combining (4.12), (4.19), (4.20) and (4.22), we see that

(4.24) $\quad 2\pi\, |R_3(r,s)| \; \leq \; 2.3\; r^{-1}e^{-1.35r^{2/3}} + 1047\; r^{-7}$.

The theorem then follows from (4.5), (4.13), (4.18)-(4.21) and (4.23-(4.24).

Theorem 4.2. Let r,s and t be nonnegative integers with $r \geq 8$; $t = r+1$. Then

(4.25a) $\quad G(rh,sh) - G(th,sh) = (1/4\pi)\log[(r^2+s^2)/(t^2+s^2)] + S_1(r,s) + S_2(r,s) + S_3(r,s),$

where

(4.25b) $\quad S_1(r,s) = \int_0^{\infty} (1/24\pi)[\lambda^2 - r\lambda^3 + \lambda^2 e^{-\lambda/2}]e^{-(r+1/2)\lambda}\cos s\lambda\,(s\lambda)$

(4.25c) $\quad S_2(r,s) = - \int_0^{\infty} (1/2\pi)[(1/30)\lambda^4 + (7/96)r\lambda^5 + (1/288)r^2\lambda^6 + (1/12)\lambda^4 e^{-\lambda/2}] \cdot$
$\cdot\; e^{-(r+1/2)\lambda}\cos s\lambda(s\lambda)$

(4.25d) $\quad |S_3(r,s)| \; \leq \; (1.9)r^{-6} + (206)r^{-7} + (0.5)r^{-1}\, e^{-1.35r^{2/3}}$.

Proof. We have

$$(\pi/2)[g(r,s)-g(t,s)] = \int_0^{\pi} [e^{-t\lambda}-e^{-r\lambda}]\lambda^{-1}\cos(s\lambda)d\lambda + \int_0^{\pi} J^{*}d\lambda,$$

where

$$J^{*} \equiv \cos(s\lambda)[e^{-t\mu}(\sinh\mu)^{-1} - e^{-t\lambda}\lambda^{-1} - e^{-r\mu}(\sinh\mu)^{-1} + e^{-r\lambda}\lambda^{-1}] .$$

It is known that

(4.26) $$\int_0^{\infty} [e^{-t\lambda}-e^{-r\lambda}]\lambda^{-1} \cos(s\lambda)d\lambda = (1/2)\log[(r^2+s^2)/(t^2+s^2)] .$$

Clearly,

(4.27) $$\int_{\pi}^{\infty} |e^{-t\lambda}- e^{-r\lambda}|\lambda^{-1}d \leq (1/\pi)r^{-1}e^{-r\pi}.$$

It therefore suffices to estimate $\int_0^{\pi} J^{*}d\lambda$. Let

(4.28) $$J^{*} = J_1^{*} + J_2^{*} ,$$

where

(4.29) $$J_1^{*} \equiv \cos(s\lambda)[e^{-r\lambda}\lambda^{-1}-e^{-r\mu}(\sinh\mu)^{-1}](1-e^{-\mu}),$$

(4.30) $$J_2^{*} \equiv \cos(s\lambda)[e^{-\mu}-e^{-\lambda}]e^{-r\lambda}\lambda^{-1} .$$

Let $\varepsilon \equiv (1.5)r^{-1/3}$. By (4.16),

$$(1-e^{-\mu})[2 \sin(\lambda/2)]^{-1} = -\sin(\lambda/2) + [1 + \sin^2(\lambda/2)]^{1/2} \leq 1.$$

But

$$-\sin(\lambda/2) + [1 + \sin^2(\lambda/2)]^{1/2} = e^{- /2} + b_1(\lambda)\lambda^3 ,$$

where

$$|b_1(\lambda)| \leq 1/12, \quad 0 \leq \lambda \leq \varepsilon .$$

Hence, by (4.14)

(4.31) $$\int_0^{\varepsilon} J_1^{*} = \int_0^{\varepsilon} \cos(s\lambda)e^{-(r+1/2)\lambda} F(\lambda)^{\prime} + E_1(r,s) ,$$

where

(4.32) $$|E_1(r,s)| \leq \int_0^{\infty} (1/12)\lambda^3 F(\lambda)e^{-r} d\lambda \leq (5.84)r^{-6} + 364r^{-8} + 10^5r^{-10} .$$

By (4.17),

(4.33) $$\int_0^{\varepsilon} J_2^{*} = \int_0^{\varepsilon} [(1/12)(\lambda^2+\lambda^4) + (1/288)\lambda^5 + C_1(\lambda)\lambda^6]\cos(s\lambda)e^{-t\lambda}d\lambda,$$

with $|C_1(\lambda)| \leq 0.035$. It is easily seen (see also p. 38 of [30]) that

(4.34) $$|\int_{\varepsilon}^{\pi} J^{*} d\lambda| \leq r^{-1} e^{-(1.5)r\ 2/3} + (2.15)r^{-1}e^{-1.35r\ 2/3} - r^{-1} e^{-r\pi}.$$

The theorem easily follows from (4.19), (4.20) and (4.26-(4.34).

Theorems 4.1 and 4.2 provide accurate estimates for distant values of the undivided differences of G. One can use these estimates as Dirichlet conditions for a fast Poisson solver to generate all the values that are needed to set up the capacitance matrix C. One can also use integer arithmetic as in [27] to construct a table of values of G for $r,s \le 7$. See e.g. Table II on p. 292 of [27] or Table I on p. 41 of [30]. By using the above tables and Theorems 4.1 and 4.2, we obtain the following.

Theorem 4.3. Let r and $t = s+1$ be positive integers. Then

$$G(sh,rh) - G(th,sh) = (1/4\pi)\log[(s^2+r^2)/(t^2+r^2)] + R(a,r) \tag{4.35}$$

where

$$|R(s,r)| < (0.34)\min\{s^{-3},r^{-3}\}. \tag{4.36}$$

Moreover,

$$\max\{\sum_{r=2}^{7} \max_{s\le r} |R(s,r)|,\ \sum_{s=2}^{7} \max_{r\le s} |R(s,r)|,\ \sum_{s=r}^{7} \max_{r\le s} |R(s,r)|\} \le 0.01. \tag{4.37}$$

We next investigate the monotone behavior of the undivided differences of G in certain directions. Let

$$G_x(i,j) \equiv G((i+1)h,jh) - G(ih,jh)$$

$$G_y(i,j) \equiv G(ih,(j+1)h) - G(ih,jh)\ .$$

By the five-point formula and symmetry,

$$G_{xx}(i-1,j) = G_{yy}(i,j-1) \qquad i\ne 0 \text{ or } j\ne 0; \tag{4.38}$$

$$2G_x(0,j) = -G_{yy}(0,j-1) \qquad j\ne 0\ . \tag{4.39}$$

Theorem 4.4 Let r and s be nonnegative integers. Then $G_x(s,r)$, $G_y(s,r)$, $-G_{xy}(s,r)$ and $G_{yy}(1,r) - G_{yy}(0,r)$ are always positive; and $G_{xx}(s-1,r)$ is always nonnegative for $r \ge s$.

Proof. Except for the result on $G_{xx}(s-1,r)$, the proof for all the other results are similar. We first estimate the values of the expression for $s = 0$, $r > 8$ using Theorem 4.1 or 4.2. By symmetry, the results hold for $s \ge 8$, $r = 0$ as well. We then verify with the aid of Table II on p. 292 of [27] that the same results hold for $s = 0$, $r \le 8$ and $s \ge 8$, $r = 0$. Since the five point formula is satisfied at all points rh and sh with $r > 0$ and $s > 0$, an application of discrete maximum principle immediately yields the desired result.

The proof for $G_{xx}(s-1,r)$ is as follows. By symmetry and (4.38), we note that

$$G_{xx}(s,r) \equiv 0 \text{ for } s = r\ .$$

By symmetry,

$$G_{xx}(s-1,r) = 2G_x(0,r) \qquad \text{for} \quad s = 0 \; .$$

An application of the discrete maximum principle therefore completes the proof of the theorem.

This concludes our discussion for undivided differences of G. We now proceed to obtain better estimates for the remainder terms in (4.8) and (4.9).

The following theorem is an immediate consequence of (4.8) and some well known result of asymptotic series. See also p. 325 of [5].

<u>Theorem 4.5</u> Let r be any positive integer. Then

$$G(rh,rh) = (1/4\pi)\log(2r^2h^2) + (1/48\pi)r^{-2} + R^7 \; , \tag{4.40}$$

where

$$|R'| \leq (7/1920\pi)r^{-4}. \tag{4.41}$$

<u>Theorem 4.6</u> Let $r > s$; r and s are nonnegative integers. Then

$$G(rh,sh) = (1/4\pi)\log[(s^2+r^2)h^2] - (1/24\pi)(r^2+s^2)^{-1} + (1/3\pi)r^2s^2(r^2+s^2)^{-3} + L(r,s) \tag{4.42}$$

where

$$|L(r,s)| \leq (2.51/\pi)r^{-4} + (154/\pi)r^{-6} + (1/\pi)\log r \; e^{-1.35r^{2/3}} \; . \tag{4.43}$$

<u>Proof</u>. As in the proof of Theorem 4.1, we have

$$\int_0^{\varepsilon} J \, d\lambda = \int_0^{\varepsilon} [\cos(s\lambda) - \cos(t\lambda)] e^{-r\lambda} [2 \sin(\lambda/2)]^{-1} F(\lambda) d\lambda.$$

Here $\varepsilon = (1.5)r^{-1/3}$, t may be any nonnegative integer. It can be shown that

$$[2 \sin(\lambda/2)]^{-1}F(\lambda) = (1/12)(\lambda - r\lambda^2) - (11/144)r\lambda^4 - (43/1440)\lambda^3 - (1/288)r^2\lambda^5 + L_1(r,\lambda).$$

Here $|L_1(r,\lambda)| \leq (0.012)\lambda^5 + (0.06)r\lambda^6 + (0.012)r^2\lambda^7 + (0.0012)r^3\lambda^8$. But

$$\int_0^{\infty} e^{-ax} \; x^n \cos(bx)dx = r! \; [(a-ib)^{n+1} + (a+ib)^{n+1}]/2(a^2+b^2)^{n+1} \; .$$

Hence,

$$\int_0^{\infty} e^{-r\lambda}\cos(s\lambda)(1/12)(\lambda - r\lambda^2) = -(1/12)(r^2+s^2)^{-1} + (2/3)r^2s^2(r^2+s^2)^{-3} = (1/24)r^{-2}, \tag{4.44}$$

when $r = s$.

We also have

$$(1/72)\int^{\infty} [11r\lambda^4 + (43/10)\lambda^3 + (1/2)r^2\lambda^5 + 144|L_1(r,\lambda)|]e^{-r\lambda}d\lambda \leq 5r^{-4} + 308r^{-6}. \tag{4.45}$$

Since $|\sinh\mu|^{-1} \leq \lambda^{-1}$ for $\lambda \leq 1$, we have

$$\left| \int_{\varepsilon}^{\pi} J d\lambda \right| \leq \int_{\varepsilon}^{\pi} \lambda^{-1}(e^{-r\lambda} + e^{-r\mu})d\lambda.$$

But $e^{-\mu} \leq e^{-0.9\lambda}$, $0 < \lambda \leq \varepsilon$. Hence,

(4.46) $$\left|\int_{\varepsilon}^{\pi} J\, d\lambda\right| \leq 2(e^{-1.5r^{2/3}} + e^{-1.35r^{2/3}}) \left| \int_{\varepsilon}^{\pi} \lambda^{-1}\, d\lambda \right|$$
$$\leq (2/3)\log r\ [e^{-15r^{2/3}} + e^{-1.35r^{2/3}}] .$$

By combining (4.40), (4.41), (4.44)-(4.46) and using the technique in the proof of Theorem 4.1 to estimate the remaining remainder terms, we see that Theorem 4.6 holds.

Theorem 4.6 also provides a means of generating $G(rh,sh)$ by means of a fast Poisson solver on a rectangle using the first three terms on the right hand side of Equation (4.42) as approximations for distant values of the Dirichlet data.

We now describe an efficient method developed in [24] of computing Gv for any vector v defined on a square mesh $S_h \supset \Omega_h$, with boundary mesh ∂S_h. Let U_S and $U_{\partial S}$ denote the extension operators from S_h and ∂S_h respectively to all mesh points that are defined the same way as U. We are actually computing $U_S^T\, G\, U_S\, v$. We first solve the system of equations

$$U_S^T B U_S \phi = v \quad \text{on} \quad S$$
$$\phi = 0 \quad \text{on} \quad S$$

for the potential ϕ. We then extend ϕ by zero to all mesh points. We represent ϕ as

(4.47) $$\phi = GU_S v + GU_{\partial S}\rho$$

where ρ is an unknown vector defined on the mesh points on ∂S to be determined. It is easy to see that

$$\rho = U_{\partial S}^T B\phi .$$

The vector $U_S^T\, G\, U_{\partial S}\rho$ can easily be computed by one fast Poisson solver on S with $U_{\partial S}^T G U_{\partial S}\rho$ as the Dirichlet data on ∂S. Because of the sparsity of the vector $U_{\partial S}$, the Dirichlet data can be computed at a cost of constant N^2. $U_S^T\, G\, U_S v$ is then computed from (4.47).

We now describe a method of computing all three of the vectors Gv, $Gv\mu$ and $W^T Gv$ using only two calls of fast Poisson solvers. This may appear to be impossible since $W^T Gv$, the right hand side of the capacitance matrix equation, must be determined first and the computation of Gv alone requires two calls of fast Poisson solvers. We can, however, first compute and store the vectors φ and ρ in equation (4.47). Clearly,

$$W^T Gv = W^T \phi - W^T GU_{\partial S}\rho.$$

But we need only to compute $GU_{\partial S}\rho$ at those mesh points that $W^T Gv$ is defined. Hence, the computation of $W^T GU_{\partial S}\rho$ requires only constant n^2 operations. On the other hand, $GU_{\partial S}\rho$ and $GU\mu$ can be computed simultaneously with one call of fast Poisson solver with $U_{\partial S}^T G(U_{\partial S}\rho + U\mu)$ as the Dirichlet data on S . Hence our algorithm of computing all three of the vectors W^TGv, $GU\mu$ and Gv requires only calls of fast Poisson solvers plus constant times n^2 operations.

In the methods described so far, a fast solver on $2N \times 2N$ mesh points is needed to generate the discrete Green's function or its undivided differences on a $N \times N$ mesh S_h . An alternative method is to first generate the undivided differences of G on a $N/2 \times N/2$ mesh using a fast solver on $N \times N$ mesh points. The values on the remaining mesh points of S_h are computed by using (4.11) and (4.25). An accuracy of eight decimal digits is guaranteed by Theorems 4.1 and 4.2 if $N \geq 60$. A somewhat less accurate but easier to program method is to generate the values of G on a $N/2 \times N/2$ mesh and compute the values on the rest of S_h by using (4.42) and (4.43). An accuracy of five decimal digits is guaranteed by Theorem 4.6 if $N \geq 60$.

We shall assume in the next two sections that the G used in the capacitance matrix equation is the discrete Green's function on the entire plane. The main results in sections 5 and 6, however, will also hold if $G = B_D^{-1}$ is used in equations (3.2) and (3.5). See section 5 of [31] for a discussion in this respect.

5. Spectral bounds of B_h .

We shall show in this section that B_h is uniformly well conditioned in the spectral norm as $h \to 0$. The following well known lemma is crucial to the proof of our main result.

Lemma 5.1. Let the symmetric part of a matrix A satisfy

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

Then

$$A^T A \geq \delta^2 I.$$

Theorem 5.1.

$$\begin{aligned} (0.25) I &\leq B_h^T B_h \leq (7.29) I && \text{for scheme I.a} \\ (0.25) I &\leq B_h^T B_h \leq (13.7) I && \text{for scheme I.b} \\ (0.04) I &\leq B_h^T B_h \leq (5.57) I && \text{for scheme II} \end{aligned}$$

for all sufficiently small $h > 0$.

Proof. We shall first prove that the following holds for scheme Ib.

$$I \leq B_h + B_h^T . \tag{5.1}$$

Let $B_s \equiv B_h + B_h^T$. We shall show that

$$\min_{P \in \partial\Omega_h} \{B_s(P,P) - \sum_{Q \in \partial\Omega_h, Q \neq P} |B_s(P,Q)|\} \geq 1 \tag{5.2}$$

so that (5.1) holds because of a well known Gerschgorin theorem. The inequality

$$I \leq B_h^T B_h$$

will then follow from Lemma 5.1.

Let $P \in \partial\Omega_h$. Assume that the local orientation of the boundary near P is such that for any point $P' \in \partial\Omega_h$ in that neighbourhood, either W' and N', the western and northern neighbours of P', are both in $(C\Omega)_h$ or W' alone is in $(C\Omega)_h$. Let $\alpha_{P'}$ be the angle $\leq \pi/4$ that the normal through P' makes with the x_1 axis in the east-west direction. By (3.5) and (3.9), we have, for $P \neq Q$,

$$B_h(P,Q) = 2[G(W_P;W_Q) - G(W_P;Q)] + 2 \tan\alpha_P [G(W_P;NW_Q) - G(W_P;W_Q)] \tag{5.3}$$

if P has only one neighbour in $(C\Omega)_h$. Here Y_P denotes the immediate neighbour on the mesh for any point P in the Y direction. Similar expressions to (5.3) are easily obtained when N_P is also in $(C\Omega)_h$ or when $P \equiv Q$. If $P \equiv (jh,kh)$, $Q \equiv (mh,nh)$, then because of translational invariance, $G(P;Q) \equiv G(|j-m|h, |k-n|h)$.

Assume that $0 < \alpha_P < \pi/4$ and that $(\partial\Omega_h)_{loc}$, which denotes a subset of $\partial\Omega_h$ that contains a $\sqrt{h}$ neighbourhood of P , can be partitioned into blocks as follows. Let

$$I_0 = \{(0,h),\ldots,(0,M_1h)\}$$

$$I_k = \{(kh,M_kh+h),\ldots,(kh,M_{k+1}h)\}, \quad k = 1,\ldots,K_1 ,$$

$$I_k = \{(-kh,-M_{-(k-1)}h,\ldots,(-kh,-M_{-k}h+h)\}, \quad k = 1,\ldots,K_2.$$

Then,

$$(\partial\Omega_h)_{loc} = \bigcup_{K=-K_2}^{K_1} I_k , \quad P \in I_0 .$$

Note that $M_k - M_{k-1}$ gives the number of points in I_{k-1}, $k = 1,\ldots,K_1$ while $M_{-k} - M_{-(k-1)}$ gives the number of points in I_{-k} , $k = 1,\ldots,K_2$; $M_0 \equiv 0$.

Let P_j denote the point with x_2-coordinate jh. From (5.3) and Theorem 4.3, it is easily verified that $\sum_{Q\neq P} |B_s(P,Q)|$ will remain essentially unchanged for sufficiently smooth $\partial\Omega$ if $\tan \alpha_Q$ is replaced throughout by $\tan\alpha_P$. Let $a \equiv \tan\alpha_P$. Let $P \equiv P_i$ and $G(i,j) \equiv G(ih,jh)$. We shall assume that $P \neq P_{M_1}$ unless otherwise stated. We easily verify that for $P_j \in I_0$,

(5.4) $$B_s(P,P) = 3+a,$$

(5.5) $$B_s(p,P_j) = -4(1+a)\ G_x(0,|i-j|), \qquad j \neq i,\ j \neq M_1 ,$$

(5.6) $$B_s(P,P_{M_1}) = -2(1+a)\ G_x(0,M_1-i) + (1-a)\ G_{yx}(0,M_1-i).$$

By (4.39), we see that for $P_j \in I_0$,

(5.7) $$\sum_{i\neq j} B_s(P,P_j) = -2(1+a)[2\ G_y(0,0) - G_y(0,M_1-i) - G_y(0,i-1)]$$
$$+ 2(1+a)\ G_x(0,M_1-i) + (1-a)\ G_{yx}(0,M_1-i).$$

For $P_j \in I_k$, $k = 1,\ldots,K_1$, we have

(5.8) $$B_s(P,P_j) = 2(1+a)\ G_{yy}(k,j-i-1), \quad j \neq M_{k+1}$$

(5.9) $$B_s(P,P_j) = (1+a)\ G_{yy}(k,j-i-1) + (1-a)\ G_{yx}(k,j-i), \quad j = M_{k+1} .$$

(5.10) $$\sum_{P_j\in I_k} B_s(P,P_j) = -2(1+a)[G_y(k_1M_k-i) - G_y(k,M_{k+1}-i)] - (1+a)G_{yy}(k,M_{k+1}-i-1)$$
$$+ (1-a)\ G_{yx}(k,M_{k+1}-i) .$$

By Theorem 4.4, we see that each $B_s(P,P_j)$, $j \neq i$, $j \leq 1$, is negative. Hence

(5.11) $$\sum_{k=1}^{K_1} \sum_{P_j\ I_k} B_s(P,P_j) = 2(1+a)[G_y(1,M_1-i) + \sum_{k=1}^{K_1} G_{yx}(k,M_{k+1}-i)]$$
$$-(1-a) \sum_{k=1}^{K_1} G_{yx}(k,M_{k+1}-i) + (1+a) \sum_{k=1}^{K_1} G_{yy}(k,M_{k-1}-i-1)$$
$$< 2(1+a)\ G_y(1,M_1-i) .$$

Similarly,

(5.12) $$\sum_{k=1}^{K_1} \sum_{P_j\in I_{-k}} |B_s(P,P_j)| < 2(1+a)\ G_y(1,i-1) + (1-a)\ G_y(0,i-1).$$

By combining (5.7), (5.11) and (5.12), we see that

$$(5.13) \qquad \sum_{i\neq j} |B_s(P,P_j)| < 4(1+a)\,G_y(0,0) + (1-a)\,G_y(0,i-1) < (5+3a\,G_y(0,0) = 5/4 + 3/4a \,.$$

Here we have used (4.5), (4.6) and Theorem 4.4. Hence, by (5.4) and (5.13), we see that (5.2) holds for $P \in I_0$ $P \neq P_{M_1}$.

The proof for $P \equiv P_{M_1}$ is quite similar and is sketched as follows.

$$(5.14) \qquad B_s(P,P) = 2+2(1-a)\,G_{xy}(0,0) \,.$$

$$(5.15) \qquad \sum_{P_j\in I_0,\, j\neq i} |B_s(P,P_j)| = (1+a)[G_y(0,0) - G_y(0,M_1-1)] + (1-a)[G_x(0,1) - G_x(0,M_1)];$$

$$(5.16) \qquad \sum_{k=1} \sum_{P_j\in I_k} |B_s(P,P_j)| = (1-a)[G_x(0,0) + \sum_{k=1} \{G_{xx}(k-1,M_{k+1}-M_1) - G_{yx}(k,M_{k+1}-M_1)\}]$$
$$+ (1+a)[G_y(1,0) + \sum_{k=1} \{G_{yx}(k,M_{k+1}-M_1) + G_{yy}(k,M_{k+1}-M_1-1)\}];$$

$$(5.17) \qquad \sum_{k=1} \sum_{P_j\in I_{-k}} |B_s(P,P_j)| = (1+a)[G_y(1,M_1) + \sum_{k=1} \{G_{yx}(k,M_1+M_k) + G_{yy}(k,M_1+M_{-k}-1)\}]$$
$$+ (1-a)[G_x(1,M_1) + \sum_{k=1} \{G_{xx}(k,M_1+M_{-k}) - G_{yx}(k-1,M_1+M_{-(k-1)}-1)\}].$$

By (4.38) and Theorem 4.4, we see that

$$(5.18) \qquad \sum_{k=1} \sum_{P_j\in I_k} |B_s(P,P_j)| \leq (1-a)\,G_x(0,0) + (1+a)\,G_y(1,0).$$

Similarly, by using the identity (5.20), we have

$$(5.19) \qquad \sum_{k=1} \sum_{P_j\in I_{-k}} |B_s(P,P_j)| \leq (1+a)G_y(1,M_1) + (1-a)[G_x(1,M_1) - G_{yx}(0,M_1-1)] \,.$$

$$(5.20) \qquad G_{yy}(k,M_1+M_{-k}-1) \equiv G_{yy}(k+1,M_1+M_{-k}-1) + G_{yx}(k,M_1+M_{-k}-1) - G_{yx}(k,M_1+M_{-k}) \,.$$

Hence,

$$(5.21) \qquad \sum_{k=1} |B_s(P,P_j)| \leq 2G_x(0,0) + 2G_x(0,1) + aG_y(1,M_1) \,.$$

By combining (5.14) and (5.21), we have

$$(5.22) \qquad B_s(P,P) - \sum_{i\neq j} |B_s(P,P_j)| \geq 2 + 2(1-a)G_{xy}(0,0) - 2G_x(0,0) - 2G_x(0,1) - aG_y(1,M_1).$$

Clearly, the right hand side of (5.22) attains its maximum at $a = 0$. Hence, (5.2) holds for $P \equiv P_{M_1}$. The proof for other choices of P is similar and will not be repeated. We note that we have assumed that $B_s(P,Q) \neq 0$ for any $Q \in (\partial\Omega_h)_{loc}$. This assumption will not affect our estimate (5.2) because each $B_s(P,Q)$ is either zero or negative for $P \neq Q$.

Finally we remark that the schemes Ia and Ib described in this work are essentially dual to the schemes I.N.a and I.N.b described in [31] in the following sense. If we maintain that $a \equiv \tan\alpha_p$ does not change its value for the entire row or column of B_h

corresponding to P, then $B_h(P,Q)$ is the same for both schemes Ia and I.N.a. or for both schemes Ib and I.N.b. Therefore we refer the reader to [31] for the proof of Theorem 5.1 for scheme I.a. We now proceed to prove the following inequality for scheme II.

(5.23) $$\min_P \{B_s(P,P) - \sum_{Q \neq P} |B_s(P,Q)|\} \geq 0.41 .$$

Let $P \in \partial\Omega_h$. We assume the same local configuration of irregular mesh points near P as before. By (3.5) and (3.7), we have for $d_2 \geq 1$, $P \neq Q$,

(5.24) $$B_h(P,Q) = -G_x(W_P;W_Q) - e_1 G_x(E_P;W_Q) + a[G_y(W_P;W_Q) + e_1 G_y(E_P;W_Q)]$$

where

(5.25) $$e_1 \equiv (1-d_1)/(1+d_1), \quad a \equiv \tan \alpha_P .$$

Here $G_x(\cdot;\cdot)$ and $G_y(\cdot;\cdot)$ denote the forward undivided differences in the x_1 and x_2 direction with respect to the second variable of any mesh function G respectively. A similar expreswion involving e_2 which is similarly defined or a constant one should be added respectively to the right hand side if $d_2 < 1$ or if $P \equiv Q$. Let e_{1j} and e_{2j} denote the corresponding e_1 and e_2 respectively.

Let $P \equiv P_i$. We first estimate $\sum_{|i-j| \geq 3} \{|B_h(P,P_j)| + |B_h^T(P,P_j)|\}$.

Let $B_h^*(P,Q)$ be defined by Equation (5.24) with G replaced by its continuous analog which we shall denote by G^*. $B_h^{T*}(P,Q)$ is similarly defined. By (5.24) and (5.25), we have for $d_2 > 1$ and $P \neq Q$

(5.26) $$\begin{aligned} B_h^*(P,Q) &= [2/(1+d_1)][-G^*(W^*;Q) + (1-a)G^*(W^*;W_Q) + a\, G^*(W^*;NW_Q)] + R_{WE} \\ &= [2/1+d_1)][G^*(W^*;W') - G^*(W^*;Q)] + R_{WE} + R_Q , \end{aligned}$$

where

(5.27) $$\begin{aligned} R_{WE} \equiv [\partial/(1+d_1)]-[&\sum_{k=2}^{n-1} (h^k/k!)[(d_1-1)^k + (d_1+1)^k] \\ &\cdot[(\partial/\partial x_1)^k\{-G^*(\cdot;Q) + (1-a)G^*(\cdot;W_Q) + aG^*(\cdot;NW_Q)\}](W^*)]\} \\ &+ \text{nth order remainder term.} \end{aligned}$$

(5.28) $$R_Q \equiv [h^2 a(1-s)/(1+d_1)]\{ a[(\partial/\partial x_2)^2 G^*(W^*;\cdot)](NW_0) + (1-a)[(\partial/\partial x_2)^2 G^*(W^*;\cdot)](NW_1)\}.$$

Here W* is the point were the Dirichlet data U_W is given. W' is a point on the mesh line connecting W_Q and NW_Q and at a distance ah from W_Q. NW_0 and NW_1 are respectively points that can be anywhere on the mesh lines between W' and W_Q and between W' and NW_Q. We shall assume that d_{2Q}, the analogue of d_2 for the point Q, is also greater than 1. In that case, we have a similar expression for $B_h^{T*}(P,Q)$ as that in (5.26).

We now proceed to estimate $B_h^{T*}(P,Q) + B_h^*(P,Q)$. It is easily verified that if P,Q are two points on $\partial\Omega$ with $d(P,Q) \leq h^\gamma$, $0 < \gamma \leq 1$ and ta is the tangent at Q to $\partial\Omega$, then

$$d(P,ta) \leq (K_{max} + o(1))h^{\partial\gamma}.$$

Here K_{max} is the maximum absolute value of the curvatures of $\partial\Omega$. Hence, without loss of generality, we may assume that W^* lies on the tangent to $\partial\Omega$ through Q^*, the point where the normal through Q intersects with $\partial\Omega$ and vice versa. Let r denote $d(W^*,Q)$ and r' denote $d(W^*,W')$. It is easily seen that $r' \geq r$ if $d(Q^*,Q) \leq h(1+a^2)/2$. We separate our discussion into four cases. The first case is when $r' > r$ and $d(P,P^*) < h(1+a^2)/2$. The maximum of $\log(r'/r)$ then occurs when Q coincides with Q^*. In this case, $r'^2 - r^2 = (1+a^2)h^2$. Hence, if $P \equiv P_i$, $Q \equiv P_j$, $i \neq j$, then

$$0 \leq G^*(W^*;W') - G^*(W^*;Q) \leq (1/4\pi)|j-i|^{-2}.$$

By (5.28),

$$-(1/8\pi)|i-j|^{-2} \leq R_Q \leq 0 \;; \qquad \text{if } j > i$$

$$-(1/8\pi)|i-j-1|^{-2} \leq R_Q \leq [a(1-a^2)/\sqrt{2}(1+d_1)\pi](1+a^2)^{-3/2}\,|j-i-1|^{-3} \;; \qquad \text{if } i > j.$$

Similarly, by (5.27),

$$(1+d_1)\,|R_{WE}| \leq (2/\pi)\,|j-i|^{-4}.$$

Hence, for $|j-i| \geq 3$,

$$(5.29) \qquad (1+d_1)|B_h^*(P,Q)| \leq (1/2\pi)|j-i|^{-2} + (1/5\pi)|j-i-1|^{-3} + (2/\pi)|j-i|^{-4}, \qquad \text{if } i > j;$$
$$\leq (1/2\pi)|j-i|^{-2} + (2/\pi)|j-i|^{-4}, \quad \text{if } j>i.$$

Since $d(P,P^*) < h(1+a^2)/2$, the estimate for $B_h^{T*}(P,Q)$ is the same as that given in (5.29). Hence,

$$(5.30) \qquad |B_h^{T*}(P,Q)| + |B_h^*(P,Q)| \leq (1/\pi)|j-i|^{-2} + (4/\pi)|j-i|^{-4} + (1/5\pi)|j-i-1|^{-3}.$$

The second case is when $r \leq r'$ and $d(P,P^*) < h(1+a^2)/2$. Let d_{1Q} denote $d(Q^*,Q)/h$. Then $d_{1Q} \geq 1/2$. Hence by (5.29),

$$|B_h^{T*}(P,Q)| \leq (1/3\pi)|j-i|^{-2} + (2/15\pi)|j-i-1|^{-3} + (4/3\pi)|j-i|^{-4}, \qquad \text{if } j > i\;;$$
$$\leq (1/3\pi)|j-i|^{-2} + (4/3\pi)|j-i|^{-4}, \qquad \text{if } i > j.$$

On the other hand, the maximum of $\log r/r'$ occurs when Q^* coincides with W_2. In this case, $r^2 - r'^2 = (1-a^2)h^2$. Hence,

$$G^*(W^*;W') - G^*(W^*;Q) \leq (1/4\pi)[(1-a^2)/(1+a^2)]|j-i|^{-2}, \qquad \text{if } j > i\;;$$
$$\leq (1/4\pi)[(1-a^2)/(1+a^2)]|j-i-a|^{2}, \qquad \text{if } i > j\;;$$

and

(5.31) $$|B_h^*(P,Q)| \leq [(1/2\pi) + (1/8\pi)]|j-i|^{-2} + (2/\pi)|j-i|^{-4} \quad \text{if} \quad j > i ;$$
$$\leq (1/2\ )|j-i|^{-2} + (1/8\pi)|j-i-1|^{-2} + (2\pi)|j-i|^{-4}, \quad \text{if} \quad i > j .$$

Therefore,

(5.32) $$|B_h^{T*}(P,Q)| + |B_h^*(P,Q)| < [(1/3\pi) + (1/2\pi) + (1/8\pi)]|j-i|^{-2} + (4/3\pi)]|j-i|^{-4}$$
$$+ (2/15\pi)|j-i-1|^{-3}, \quad \text{if} \quad j > i ;$$
$$\leq [(1/2\pi) + (1/3\pi)]|j-i|^{-2} + (1/8\pi)|j-i-1|^{-2}$$
$$+ [(4/3\pi) + (2/\pi)]|j-i|^{-4}, \quad \text{if} \quad i > j .$$

The third case is when $r < r'$ and $d(P,P^*) \geq h(1+a^2)/2$. The estimate for $|B_h^*(P,Q)| + +|B_h^{T*}(P,Q)|$ is the same as that for the second case.

The fourth case is when $r > r'$ and $d(P,P^*) \geq h(1+a^2)/2$. Both d_1 and d_{1Q} are not less than $1/2$. Therefore by (5.31) and the above observation,

(5.33) $$|B_h^{T*}(P,Q)| + |B_h^*(P,Q)| \leq (2/3\pi)|j-i|^{-2} + (1/12\pi)|j-i|^{-2}$$
$$+ (1/12\pi)|j-i|^{-2} + (8/3\pi)|j-i|^{-4} .$$

By comparing (5.30), (5.32) and (5.33), we see that

$$\max |B_h^{T*}(p,Q)| + |B_h^*(P,Q| \leq (1/\pi)|j-i|^{-2} + (4/\pi)|j-i|^{-4} + (1/5\pi)|j-i-1|^{-3} \quad \text{if} \quad j > i$$
$$\leq (1/\pi)|j-i|^{-2} + (1/8\pi)|j-i|^{-2} + (1/5\pi)|j-i-1|^{-3} + (4/\pi)|j-i|^{-4}$$
$$\text{if} \quad i > j .$$

Hence, if $d_2, d_{2Q} > 1$, then

$$\sum_{|j-i|\geq 3} |B_h^{T*}(P,P_j| + |B_h^*(P,P_j)| \leq (1/\pi) \sum_{k=3} \{2/k^2 + 8/k^4 + 2/5k^3 + 1/8k^2\}$$
$$\leq 0.328 .$$

Similarly, it can be shown that if both d_2 and d_{2Q} are not greater than 1, then

(5.34) $$\sum_{|j-i|\geq 3} |B_h^{T*}(P,P_j)| + |B_h^*(P,P_j)| \leq 0.677.$$

By Theorems 4.1-4.3 and Table I on p. 41 of [30],

$$\sum_{|i-j|\leq 3} \{|B_h(P,P_j) - B_h^*(p,P_j)| + |B_h^T(P,P_j)|\} \leq 0.04 .$$

Hence

(5.35) $$\sum_{|j-i|\geq 3} \{|B_h(P,P_j)| + |B_h^T(P,P_j)|\} \leq 0.717 .$$

It remains to estimate $B_s(P,P) - \sum|B_s(P,P_j)|$, $|i-j| \leq 2$, $i \neq j$. Without loss of generality, we may assume that both d_1 and d_2 are less than 1. We shall assume that $P \equiv P_i \equiv P_{M_1}$ with $M_1 \geq 3$ and $M_2 \geq 5$. The case when $i \neq M_1$ can be treated in a similar manner. We have

(5.36) $$B_s(P,P) = 2 + 2(1-a)\, G_{xy}(0,0) + g_0(a) ,$$

where

$$g_0(a) \equiv 2\, e_1[G_x(1,0) + G_y(2,0)] + 2\, e_2[G_x(0,1) + a\, G_y(1,1)].$$

For any $P_j \in I_0$, $j \neq M_1$, we have

$$B_s(P,P_j) = g_1(a,j) + g_2(a,j) ,$$

where

$$g_1(a,j) \equiv G_{xy}(0,i-j) + a\, G_{yy}(0,i-j-1) - a\, G_y(1,i-j) - G_x(0,i-j) ,$$

$$g_2(a,j) \equiv e_1[G_x(1,i-j) - a\, G_y(2,i-j-1)] + e_{ij}[G_x(1,i-j) + a\, G_y(2,i-j)] + e_2[G_x(0,i-j-1) - a\, G_y(1,i-j-2)] .$$

For any $P_j \in I_1$, we have

$$B_s(P,P_j) = g_3(a,j) + g_4(a,j) ,$$

where

$$g_3(a,j) \equiv -G_x(0,j-i-1) + a\, G_{yy}(1,j-i-1) + a\, G_y(0,j-i-1)$$

$$g_4(a,j) \equiv e_1[G_x(0,j-i) + a\, G_y(1,j-i)] + e_2[-Gx(0,j-i+1) + a\, G_y(0,j-i+1)] + e_{ij}[G_x(2,j-i) - a\, G_y(3,j-i-1)].$$

By Theorem 4.4, $g_1(a,j)$ is negative. It is easily verified that $g_2(a,j)$ is nonnegative for $0 \leq i-j \leq 2$; $g_3(a,j)$ is negative for $a \leq 1/2$; and $g_4(a,j)$ is positive. Moreover, for $a \leq 1/2$,

(5.37) $$\sum_{i-j=1}^{2} |g_1(a,j) + g_2(a,j)| + \sum_{j-i=1}^{2} |g_3(a,j) + g_4(a,j)| < g_0(a) + \sum_{i-j=1}^{2} |g_1(a,j)| + \sum_{j-i=1}^{2} |g_3(a,j)|.$$

Hence, for $a < 1/3$, $M_1 \geq 3$, $M_2 \geq 5$, we have from (5.37) that the following holds.

(5.38) $$\begin{aligned} & B_s(P,P) - \sum\{|B_s(P,P_j)| \ , \ i \neq j, \ |i-j| \leq 2\} \\ & \geq 2+2(1-a)G_{xy}(0,0) + \sum_{i-j=1}^{2} g_1(a,j) + \sum_{j-i=1}^{2} g_3(a,j) \\ & \geq 2+2(1-a)G_{xy}(0,0) - G_x(0,0) - 3G_x(0,1) - G_x(0,2) + G_x(0,3) \\ & \geq 1.17. \end{aligned}$$

By considering all possible configurations of P_j, $i-j \leq 2$, $a \geq 1/3$, it can be shown that the constant 1.17 is always majorized by the left hand side of (5.38). It is easy to see that (5.38) also holds when $P_i \neq P_{M_1}$. Hence by (5.35), we see that (5.23) holds.

This established the lower spectral bounds of B_h for all schemes. To complete the proof of the theorem, we note that the spectral norm of $B_h^T B_h$ is majorized by $(1/2)(\|B_h^T\|_\infty + \| B_h\|_\infty)^2$. It therefore suffices to prove the following two inequalities

$$\| B_h^T\|_\infty + \| B_h\|_\infty \leq 7.4 \qquad \text{if scheme I.b is used} \tag{5.39}$$

$$\| B_h^T\|_\infty + \| B_h\|_\infty \leq 4.72 \qquad \text{if scheme II is used.} \tag{5.40}$$

We first prove that (5.39) holds. Without loss of generality we may assume that $P \in \partial\Omega_h$ has only one western neighbour W in $(C\Omega)_h$. Let $P \equiv (0,0)$ and $P_j \equiv (x,y)$, be in $\partial\Omega_h$. Then $x = ay + b$, $|b| \leq (1+o(1))h$ if $d(P,P_1) < \sqrt{h}$. And

$$B_h^*(P,P_j) = -(1/2\pi)\log[(x+h)^2+y^2] + (a/2\pi)\log[x^2+(y+h)^2] + [(1-a)/2\pi)\log(x^2+y^2)$$
$$= (1/2\pi)[2h(x-ay) + (a-1)h^2]/r^2 + R,$$

where $r = d(P,P_j)$ and $|R| \leq (1/2\pi)[(2x+h)^2 + (2y+h)^2]h^2/2r^4$. It is easily verified that

$$\sum_{P\neq P_j} |B_h^*(P,P_j)| \leq 1.25 \qquad \text{if} \quad |y| \geq 2h \tag{5.41}$$

By Theorems 4.1-4.3 and the Table I on p. 41 of [30],

$$B_h(p,P) = 2 - (1/2)(1-a) \; ; \tag{5.42}$$

$$\sum_{|y|<2h} |B_h(P,P_j)| \leq 0.28 + 0.55a/2 \; ; \tag{5.43}$$

$$\sum_{|y|\geq 2h} |B_h(P,P_j) - B_h^*(P,P_j)| < 0.06 \, . \tag{5.44}$$

By (5.41)-(5.44),

$$\sum |B_h(P,P_j)| < 3.7.$$

It is easily seen that the above inequality also holds when B_h is replaced by B_h^T. We have therefore completed the proof of (5.39) .

Let $P \equiv P_i$. By (5.36),

$$|B_s(P,P)| \leq 2+2(1-a)G_{xy}(0,0) + g_0(a) \, . \tag{5.45}$$

By (5.37) and (5.45),

$$\sum_{|i-j|\leq 2} |B_h(P,P_j)|+|B_h^T(P,P_j)| < 2+2g_0(a) + 2G_x(0,0) + 6G_x(0,1) \tag{5.46}$$
$$\leq 4.$$

By (5.35) and (5.46), we see that (5.40) holds.

6. Singular values of K_h and C .

We shall show that all except a few singular values of C lie in the interval $[d_1 - \varepsilon, d_2 + \varepsilon]$, $\varepsilon > 0$, where d_1 and d_2 are the spectral bounds of B_h. This is accomplished by first proving that the singular values of K_h cluster around that of a compact operator K . Our main result then follows as an immediate consequence of a well known result in matrix theory which will be stated below as Lemma 6.8. We first need some definitions from modern analysis. Let X denote a Banach space throughout this section.

Definition. A subset $S \subset X$ is sequentially compact if any sequence in S contains a convergent subsequence with limit in X .

Definition. A family of operators K_n on X is collectively compact if the set $\{K_m f : \|f\| < 1, f \in X, m = 1,2,\ldots\}$ is sequentially compact in X .

We shall first assume that either scheme I.a or scheme I.b is used. We start by constructing a family of operators $\{K_m\}$ from $\{K_h\}$ in the same way that is done in Section 5 of [32]. For completeness, we briefly sketch this construction in the following. Define K_m: $C[0,1] \to C[0,1]$ by

(6.1) $$[K_m] f(t) = \sum_{j=1}^{m} k(t,t_j) f(t_j), \qquad t_j \in [0,1];\ f \in C[0,1],$$

where

(6.2) $$k(t,t_j) = K_h(P_i,P_j) + [(t-t_i)/(t_{i-1}-t_i)][K_h(P_{i-1},P_j) - K_h(P_i,P_j)]\ ,\quad t_{i-1} \le t \le t_i\ .$$

$C[0,1]$ is the Banach space of continuous functions on $[0,1]$. The t_i, $i = 1,\ldots,m$ are defined as follows. Let ϕ,ψ be a smooth parametrization of $\partial\Omega$. Then $(\phi(t_i),\psi(t_i))$ is the closest point on $\partial\Omega$ to $P_i \in \partial\Omega_h$ which is on the normal through P_i. When t is very close to 0 or 1 , the $k(t,t_j)$ in (6.2) should be adjusted slightly. See [32] for the details. We can construct by the same procedure a family of operators $\{K'_m\}$ from $\{K_h^T\}$. Let $K_s \equiv K'_m K_m$.

Lemma 6.1. The nonzero eigenvalues of K_h, $K_h+K_h^T$ and $K_h^T K_h$ coincides with that of K_m, $K_m+K'_m$ and K_s respectively.

Proof: See e.g. Lemma 5.2 in [32].

Lemma 6.2. Let P and Q be two points in $\partial\Omega_h$ with $d(P,Q) = h^\beta$, $\beta \le 1/2$. Let P^* and Q^* the closest points on $\partial\Omega$ to P and Q respectively. Then

$$K_h(P,Q) = 2[\partial G^*/\partial\nu_{Q^*}](P^*;Q^*) h \sec\alpha_Q + O(h^{2-2\beta})\ .$$

Proof: Essentially the same as that of Lemma 5.4 in [31].

__Lemma 6.3.__ The families of operators $\{K_m\},\{K'_m\}$ and $\{K_s\}$ are collectively compact on $C[0,1]$.

__Proof:__ Essentially the same as that of Lemma 5.5 in [32].

__Lemma 6.4.__ $K_m f \to Kf$, $K'_m f \to K^T f$ and $K_s f \to K^T K f$ for each $f \in C[0,1]$ where K is the compact integral operator defined by

$$(Kf)(t_P) = 2 \int_{\partial\Omega} [\partial G^*/\partial\nu_Q](P;Q) f \, ds_Q$$

where $P \equiv (\phi(t_P),\psi(t_P))$.

__Proof:__ Essentially the same as that of Lemma 5.6 in [32] .

In order for the above theorems to apply in the case when scheme II is used, we scale the matrix K_h in that case as follows. The rows of K_h that correspond to irregular mesh points that have one or two neighbours in $(C\Omega)_h$ are multiplied with $(1+d_1)$ or $(1+d_1)(1+d_2)(1+d_1+d_2)^{-1}$ respectively. It is easily verified that Lemmas 6.3 and 6.4 hold for scheme II if K_m is constructed from the scaled K_h. It will be shown after theorem 6.1 that such a scaling is not e-sential and our main results will hold even without it.

__Lemma 6.5.__ Let $\{K_n\}$ be collectively compact on X ; $K_n f \to Kf$ for each $f \in X$. Given $\varepsilon > 0$, let μ_i, with algebraic multiplicities m_i, $i = 1,\dots,N$ be the eigenvalues of K with absolute values greater than or equal to $\varepsilon > 0$. Then there exist positive numbers N^* and $\varepsilon^* < \varepsilon$ such that for all $n \geq N^*$ each ε^* neighbourhood of μ_i contains exactly m_i eigenvalues of K_n while all the other eigenvalues of K_n lie in an ε-neighbourhood of zero.

__Proof:__ This is an immediate consequence of Theorem 4.8 on p. 65 of [1]. See also Chapter 4 of [30]. By combining Lemmas 6.1, 6.3, 6.4 and 6.5, we easily have the following.

__Theorem 6.1__ Given $\varepsilon > 0$, there exists a positive integer N such that for all $h > 0$, all except N singular values of K_h lie in $[0,\varepsilon]$.

__Lemma 6.6.__ Let $C = AB$, where A, B and C are arbitrary matrices with singular values $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_m \geq 0$, $\beta_1 \geq \beta_2 \geq \dots \geq \beta_m \geq 0$ and $\gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_m \geq 0$ respectively, then

$$\gamma_{i+j+1} \leq \alpha_{i+1}\beta_{j+1}\ , \qquad i,j \text{ positive integers.}$$

__Proof:__ See e.g. Exercise 28 on p. 89 of [23] . An immediate consequence of Lemma 6.6 is that Theorem 6.1 holds in the case when scheme II is used even if the matrices C or K_h are not scaled by the scaling described just before Lemma 6.5.

__Lemma 6.7.__ If $D = A+B$, where A and B are as in Lemma 6.6, and $\delta_1 \geq \delta_2 \geq \dots \geq \delta_m \geq 0$ are the singular values of D , then

$$\delta_{i+j+1} \leq \alpha_{i+1} + \beta_{j+1}\ , \qquad i,j \text{ positive integers.}$$

Proof: See e.g. Exercise 30 on p. 89 of [23].

Theorem 6.2 Let d_1 and d_2 be the spectral bounds of B_h. Then given $\varepsilon > 0$, there exists a positive integer N independent of h such that all except N singular values of C lie in $[d_1-\varepsilon, d_2+\varepsilon]$.

Proof: An immediate consequence of Lemma 6.7 and Theorem 6.1. See also Theorem 5.3 in [32]. In the following, $\| \ \|$ shall denote either the spectral norm of a matrix or the Euclidean norm of a vector.

Lemma 6.8 Let U^* be the extension operator from $\Omega_h \cup \partial\Omega_h$ ot all mesh points that is defined the same way as U. Suppose that $U^{*T} G V\mu \neq 0$ for any nonzero m-vector μ defined on $\partial\Omega_h$. Then C is nonsingular. Moreover, if $\| U^* G V\mu \| \geq c_1 \|\mu\| / \| A_{11} \|$ for any m-vector μ then $\| C^{-1} \| \leq \kappa(A_{11})/C_1$, where $\kappa(A_{11})$ is the spectral condition number of A_{11} with respect to the norm $\| \ \|$

Proof: Let $Au = v = UU^Tv$ with U^Tv in the range of C be the equation we are solving. From Section 3, we see that $u = GV\mu$ is a solution of $Au = v$ if μ satisfies Equation (3.6). Suppose C is singular so that there exist two distince solutions μ_1 and μ_2 of Equation (3.6). Let $\mu_0 \equiv \mu_1 - \mu_2$. Then $AGV\mu_0 = 0$. Because of the reducible structure of A, $A_{11}U^{*T}GV\mu_0 = 0$. This contradicts the assumptions that A_{11} is nonsingular and $U^{*T}GV\mu_0 \neq 0$. Moreover, if $\| U^{*T}GV\mu \| \geq c_1 \|\mu\| / \|A_{11}\|$, then

$$\| A_{11}^{-1} \| \, \| U^Tv \| \geq \| U^{*T}GV\mu \| \geq c_1 \|\mu\| / \|A_{11}\| .$$

The lemma easily follows.

Definition. A scheme of interpolating boundary conditions is said to be admissible if its corresponding coefficient matrix A_{11} of the discrete problem is nonsingular and $\kappa(A_{11}) \leq$ constant h^{-2}.

Lemma 6.9 Let C^* and A^* denote respectively the capacitance matrix and the coefficient matrix of the discrete problem for a certain scheme of interpolating boundary conditions. Suppose that both C^* and A^* are nonsingular. Then C is nonsingular for any admissible scheme of interpolating boundary conditions. Moreover, if $\| C^{*-1} \| \leq c_2$ and $\|A^*\| \leq c_3 \|A_{11}\|$, then $\| C^{-1} \| \leq c_2 c_3 \kappa(A_{11})$.

Proof: We first claim that if both C^* and A^* are nonsingular, then there exists no $\mu \neq 0$ such that $U^{*T}GV\mu = 0$. Suppose this is not so, then there exists a $\mu^* \neq 0$ such that $U^{*T}GV\mu^* = 0$. Let $v^* = UU^Tv^*$ where $C^*\mu^* = U^Tv^*$. Since C^* is nonsingular, $U^{*T}v^* \neq 0$. But

$$U^{*T}v^* = U^{*T}AGV\mu^* = A^*U^{*T}GV\mu^* = 0 .$$

This proves our claim. By Lemma 6.8, C is nonsingular. Suppose now that $\| C^{*-1} \| \leq c_2$ and $\| A^* \| \leq c_3 \| A_{11} \|$. Let $Au = v = UU^Tv$. Clearly,

$$\| \mu \| \leq \| C^{*-1} \| \, \| U^Tv \| \leq \| C^{*-1} \| \, \| A^* \| \, \| U^{*T}GV\mu \| .$$

Hence,

$$\| U^{*T}GV\mu \| \geq \| \mu \| / c_2c_3 \| A_{11} \| .$$

The lemma easily follows from Lemma 6.8.

Definition Ω is said to be in $\mathfrak{J}(\beta)$ if the associated integral operator K defined by Equations (2.2)-(2.3) is such that $K + K^T > -\beta I$.

Lemma 6.10. All ellipses with thickness $b/a > 1/3$ are in $\mathfrak{J}(1)$. Here a and b are respectively the major and minor axes of the ellipses.

Proof: An immediate consequence of (2.5).

Theorem 6.3 Let $G = B_D^{-1}$ be the discrete Green's function used in equations (3.2) and (3.5). Then the capacitance matrix C is nonsingular and $\| C^{-1} \| \leq$ constant h^{-q} for some positive integer q independent of h .

Proof: It suffices to find a pair $< C^*, A^* >$ that satisfies the hypothesis of lemma 6.9. Assume that the difference equations are already preordered in such a way so that

$$B = \begin{pmatrix} B_{11} & B_{12} & 0 \\ B_{21} & B_{22} & B_{23} \\ 0 & B_{32} & B_{33} \end{pmatrix} ,$$

where the first, second and last rows of B in block form correspond to the coefficient matrices of the difference equations on $\Omega_h, \partial\Omega_h$ and $(C\Omega)_h$ respectively.

Suppose that in forming A we use a zero order interpolation of Dirichlet data on $\partial\Omega$ at $\partial\Omega_h$ to obtain the equations on $\partial\Omega_h$. Partition V, A and G in the same way as B. We obtain

$$(6.3) \qquad V = \begin{pmatrix} 0 \\ V_2 \\ V_3 \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} B_{11} & B_{12} & 0 \\ 0 & I & 0 \\ 0 & B_{32} & B_{33} \end{pmatrix}$$

so that the capacitance matrix which we now denote by C_D satisfies

(6.4) $$C_D = G_{22} V_2 + G_{23} V_3 \ .$$

Let

$$A_N = \begin{pmatrix} B_{11} & B_{12} & 0 \\ 0 & V_2^T & V_3^T \\ 0 & B_{32} & B_{33} \end{pmatrix}, \quad A_{1N} = \begin{pmatrix} V_2^T & V_3^T \\ B_{32} & B_{33} \end{pmatrix} .$$

It is easily seen that A_{1N} is the coefficient matrix of a discrete exterior Neumann problem, with the normal derivative approximated by the first order scheme described on p. 203 of [13]. Hence A_N is nonsingular since both B_{11} and A_{1N} are nonsingular and A_N is reducible. Clearly,

(6.5) $$A_N = B + UV^T - UU^T B \ .$$

Let u_N be the solution of

(6.6) $$A_N u_N = f_N \ ,$$

where f_N is any mesh function that vanishes outside $\partial\Omega_h$. Suppose we make the Ansatz that

(6.7) $$u_N = B^{-1} U\rho$$

where ρ satisfies

(6.8) $$U^T A_N B^{-1} U\rho = U^T f_N \ .$$

By (6.5), (6.7) and (6.8), it is easily seen that (6.6) is satisfied. Let $C_N \equiv U^T A_N B^{-1} U$. It is clear that C_N is nonsingular. By (6.4) and (6.5), we have

$$C_N^T = D_D$$

so that C_D is nonsingular. Moreover, using an argument similar to the proof of lemma 6.8, we have

$$\| C_D^{-1} \| = \| C_N^{-1} \| \leq h^{-q}$$

for some positive integer q independent of h . The Theorem easily follows.

<u>Theorem 6.4</u> Let $\Omega \in \mathfrak{J}(1)$. Assume that the G in equation (3.5) is the discrete Green's function of the entire plane. Then $\| C^{-1} \| \leq$ constant h^{-2} as $h \to 0$ for any admissible scheme of interpolating boundary conditions. Moreover, $\| C^{-1} \| \leq$ constant as $h \to 0$ if either scheme I.a or scheme I.b is used or if scheme II is used and $\Omega \in \mathfrak{J}(0.4)$.

<u>Proof</u>: Let $\Omega \in \mathfrak{J}(1)$. Assume that either scheme I.a or scheme I.b is used for interpolating the boundary conditions. By (5.1), $B_h + B_h^T \geq I$. By assumption, there exists an $\varepsilon > 0$ such that $K + K^T \geq -I + \varepsilon$. By lemmas 5.1 and 6.3-6.5, we see that for sufficiently small h ,

$K_h + K_h^T \geq -I + \varepsilon/2$. Hence, $C + C^T \geq \varepsilon/2$ and $\| C^{-1} \| \leq$ constant as $h \to 0$. Similarly, it can be shown that $\| C^{-1} \| \leq$ constant if scheme II is used and $\Omega \in \mathfrak{F}(0,4)$. By lemma 6.9, $\| C^{-1} \| \leq$ constant $\cdot\ h^{-2}$ for any admissible scheme of interpolating boundary conditions.

§7. Convergence of conjugate gradient iteration.

Let b denote the right hand side of the capacitance matrix equation multiplied by C^T. Let Q denote C^TC. We are concerned with solving $Q_\rho = b$ by the conjugate gradient method. Detailed exposition of the method can be found e.g. in [11], [16], [17], [18] and [26]. A brief description of the method plus a simple extension of the known results in the above references can be found in Section 6 of [32]. It will be assumed that the readers are familiar with the results in [32].

Let ρ_k denote the vectors approximating the solution ρ generated by the conjugate gradient process. Let R denote the set of real numbers and L_m denote the set of m vectors. Let $Z = R^2 \to R$ and $E : L_m \to R$ be defined respectively by

$$Z(a,b) = \{(1 - \sqrt{a})/(1 + \sqrt{a})\}^b;$$

$$E(\rho_k) \equiv (1/2)(\rho_k - \rho)^T Q(\rho_k - \rho) .$$

It is shown in [32] that the following holds.

Theorem 7.1 Let K and K_1 be the spectral condition numbers of Q and $B_h^TB_h$ respectively. Let d^* and d' denote the smallest and largest eigenvalues of $B_h^TB_h$ resepctively. Then given $\varepsilon > 0$, there exists a positive integer independent of k and h such that

$$E(\rho_k)/E(\rho_0) \le \min\{4Z(\kappa,2k),\ 4Z(\kappa_1-2\varepsilon/d',2k-2N)\chi(\lambda)\} .$$

Here $\chi(\lambda) = \max_{d^* \le \lambda \le d'} \Pi|1-\lambda/\lambda_i|$, $i = 1,\dots,N$, where λ_i, $i = 1,\dots,N$ are the N eigenvalues of Q that lie outside of $[d^*-\varepsilon, d'+\varepsilon]$.

Corollary 7.1 Let $\mathfrak{G} = B_D^{-1}$ be used in equation (3.5). The number of iterations needed to reduce $E(\rho_k)/E(\rho_0)$ to a given accuracy can grow no faster than constant·log m as $h \to 0$.

Proof. By Theorem 6.3, $|\chi(\lambda)| \le h^{-kN}$, where k is a constant independent of h. The corollary is therefore an easy consequence of Theorem 7.1.

Corollary 7.2 Let $\Omega \in \mathfrak{J}^*(\beta)$, $\beta = 1$ if either scheme I.a or I.b is used; $\beta = 0.4$ if sheme II is used. Then the number of iterations needed to reduce $E(\rho_k)/E(\rho_0)$ to a given accuracy stays constant as $h \to 0$ if the G in (3.5) is the discrete Green's function on the entire plane.

Proof. By Theorem 5.1 and Theorem 6.4, C is uniformly well conditioned in the spectral norm. The corollary is therefore an immediate consequence of Theorem 7.1.

§8 Survey of previous work on caoacitance matrix methods.

R. W. Hockney in [20] and [21] described a method of this type which can be used for the solution of the interior Dirichlet problem for Laplace's equation. His capacitance matrices are always positive definite symmetric. His method thus corresponds to a single layer Ansatz for the Dirichlet problem. Buzbee, Dorr, George and Golub used a similar method in [8]. They made the Ansatz

$$u = B^{-1}v + B^{-1} U W\rho \ ,$$

when B is nonsingular. Here W is a $m \times m$ nonsingular metrix. The choice $W = I$ gives the Woodbury formula.

Proskurowski and Widlund introduced the double layer Ansatz in [29]. The algorithm used in their work differs from the one used here only in the discrete Green's function G and the W^T matrix. No theoretical analysis was presented in [29]. In [2]] the author analyzed the method for the Neumann problem. The algorithm used in [32] is similar to the one used by George in [15] which corresponds to solving the single layer Ansatz of the Dirichlet problem in an iterative imbedding fashion.

§9 Numerical experiments

The results in this section were obtained on the CDC 7600 at Lawrence Berkelev Laboratory. The model problem is the Laplace equation on ellipses with $\gamma = (a-b)/(a+b)$, where a,b are the half axes. $E(u) = \|u-u^*\|_\infty$ where u^* is the true solution of $-\Delta u = 0$ on Ω, $u = 1$ on $\partial\Omega$. The mesh size $h = 1/32$. The number of iterations of the conjugate gradient method is denoted by n $N(R)$ denotes the normalized norm of the residuals which is the L_2 norm of the residual divided by the square root of points in Ω. The numbers given for $E(u)$ are acturally upper bounds that describe the number of accurate digits only. The capacitance matrix is generated explicitly and the discrete Green's function on the plane is used in (3.5).

TABLE I

		Scheme	I.a	Scheme	I.b	Scheme	II
n	γ	N(R)	E(u)	N(R)	E(u)	N(R)	E(u)
4	0.2	—	—	—	—	3.9-04	1.0-03
5	0,2	—	—	—	—	2.1-04	1.0-03
4	1	1.5-04	1.0-03	8.7-03	1.0-02	3.5-04	1.0-03
5	1	1.0-04	1.0-03	4.2-03	1.0-02	1.6-04	1.0-03

In Table I we see that typically it takes four iterations to achieve three digits accuracy. The operation count of the conjugate gradient routine is therefore approximately $64n^2$. The total operation count (not counting that of setting up the matrix c) is therefore approximately $5n^2\log n + 80n^2$ for the Laplace's equation and $10n^2\log n + 120n^2$ for the Poisson equation.

References

1. P. M. Anselone, Collectively compact operator, Prentice Hall, New Jersey, 1971.

2. R. E. Bank, Marching algorithms for elliptic boundary value problems, Ph.D. Thesis, 1975, Harvard University.

3. R. E. Bank and D. J. Rose, Extrapolated fast direct algorithms for elliptic boundary value problems, TR-7-16, Harvard University.

4. J. H. Bramble and B. E. Hubbard, Approximation of derivatives by finite difference analogues of the Dirichlet problem for elliptic equations, Contr. Diff. Eqns., 2(1963).

5. A. Bromwich, An introduction to the theory of infinite series, McMillian & Co., 1926.

6. O. Buneman, Analytic inversion of the five-point Poisson operator, J. of Comp. Phys. 8, (1971), 500-505.

7. B. L. Buzbee and F. W. Dorr, The direct solution of the biharmonic equation on rectangular regions and the Poisson equation on irregular regions, SIAM J. Numer. Anal., 11 (1974), 753-763.

8. B. L. Buzbee, F. W. Dorr, J. A. George and G. H. Golumb, The direct solution of the discrete Poisson equation on irregular regions, SIAM J. Numer. Anal., 8 (1971), 722-736.

9. B. L. Buzbee, G. H. Golub and C. W. Nielson, On direct methods for solving Poisson's equation, SIAM J. Numer. Anal., 7 (1970), 627-656.

10. R. Courant and D. Hilbert, Methods of mathematical physics, Interscience, 1953.

11. J. W. Daniel, The conjugate gradient method for linear and nonlinear operator equations, SIAM J. Numer. Anal.,

12. D. Fischer, G. Golub, O. Hald, C. Leiva and O. Widlund, On Fourier-Toeplitz methods for separable elliptic problems, Math. Comp., 28 (1974), 349-368.

13. G. E. Forsythe and W. R. Wasow, Finite differences methods for partial differential equations, Wiley, 1960.

14. P. R. Garabedian, Partial differential equations, Wiley, 1964.

15. J. A. George, The use of direct methods for the solution of the discrete Poisson equation on nonrectangular regions, Comp. Sci. Dept. Report 159, Stanford University.

16. R. M. Hayes, Iterative methods of solving linear problems on Hilbert space, Nat. Bur. of Standards, Applied Math. Series, 39 (1954), 71-103.

17. M. R. Hestenes, The conjugate gradient method for solving linear systems, Proc. Symp. Appl. Math., 6 (1956), 83-102.

18. M. R. Hestenes and E. Stiefel, Method of conjugate gradients for solving linear systems, J. Res. Nat. Bur. Standards, 49 (1952), 409-436.

19. R. W. Hockney, A fast direct solution of Poisson's equation using Fourier analysis, J. Assoc. Comp. Mach., 12 (1965), 95-113.

20. R. W. Hockney, Formation and stability of virtual electrodes in a cylinder, J. Appl. Phys., 39 (1968), 4166-4170.

21. R. W. Hockney. POT4- A fast direct Poisson solver for the rectangle allowing some mixed boundary conditions and internal electrodes, IBM Research, R.C. 28%, 1970.

22. R. W. Hockney, The potential calculation and some applications, Methods in computational physics, 9 (1970), Academic Press.

23. A. S. Householder, The theory of matrices in numerical analysis, Blaisdell, 1964.

24. R. A. James, The solution of Poisson's equation for isolated source distribution, to appear.

25. L. V. Kantorovich and V. I. Krylov, Approximate methods of higher analysis, Interscience, 1958.

26. D. G. Luenberger, Introduction to linear and nonlinear programming, Addison-Wesley, 1973.

27. W. H. McCrea and F. J. W. Whipple, Random paths in two or three dimensions, Proc. Roy. Soc. Edinburg, Sect. A, 60, 281-298.

28. I. G. Petrovsky, Partial differential equations, Interscience, 1954.

29. W. Proskurowski and O. Widlund, On the numerical solution of Helmholtz equation by the capacitance matrix method, Math. Comp., 30 (1976), 433-168.

30. A. Shieh, Fast Poisson solvers on nonrectangular regions, Ph.D. Thesis, N.Y.U., 1976

31. A. Shieh, On the convergence of the conjugate gradient method for singular capacitance matrix equations, MRC TSR #1730, Math. Research Center, University of Wisconsin, Madison Wisconsin, 1977.

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

TECHNICAL INFORMATION DEPARTMENT
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720