JOURNAL OF COMPUTATIONAL PHYSICS 15, 46–54 (1974)

The Direct Solution of the Discrete Poisson Equation on the Surface of a Sphere

PAUL N. SWARZTRAUBER*

National Center for Atmospheric Research, Boulder, Colorado 80303

Received July 11, 1973

The solution of Poisson's equation on the surface of a sphere may not exist unless the right side of the equation is perturbed. A method for perturbing is described which then admits a least squares solution. This least squares solution is obtained by the Fourier method which is economical in both computational time and storage.

1. INTRODUCTION

A solution of Poisson's equation on the surface of a sphere may not exist and if it does exist it is not unique. Any solution is nonunique since that solution plus an arbitrary constant is also a solution. In many applications this nonuniqueness is not a concern. For example, in computing incompressible fluid flow, the pressure is determined by solving Poisson's equation. However, the gradient of the pressure is used to predict the course of the fluid, and therefore an additive constant may be selected arbitrarily without affecting the flow. Hence, instead of a single solution u, we are interested in an equivalence class $\hat{u}$ of solutions where $v \in \hat{u}$ if and only if $v = u +$ constant.

The nonexistence of a solution is a problem which requires additional consideration. For a solution to exist, the right side of Poisson's equation must satisfy an orthogonality condition which will be derived in Section 3. As a result of computational or observational errors this condition may not be satisfied. In this event, a reasonable alternative is to determine a least-squares solution [2, 4]. In this paper, we elect to perturb the right side of the equations so that the system becomes consistent and existing methods can then be used to obtain the solution. There are two factors to consider when perturbing the right side. First, the perturbation should be small so that the solution almost satisfies the unperturbed equations. Second, the perturbation should be the same at each point since it is probable that no a priori knowledge is available about the functional dependence of the errors. The

* Research sponsored by the National Science Foundation.

first factor is treated by perturbing the right side so that the solution to the resulting consistent system is a least squares solution to the unperturbed system. However, the least-squares solution depends on the choice of the vector norm. The second factor is treated by selecting this norm so that the resulting perturbation is a constant. It is shown at the end of Section 3 that the usual choice of the l_2 norm results in a perturbation which is not constant and approximates a function which is not differentiable at the poles. This latter property is undesirable as it can result in a solution which is not regular at the poles even though the right side of Poisson's equation is specified as regular. In time dependent flow problems, Poisson's equation is solved many times and the error induced by such irregularities can accumulate.

The discrete problem is described in Section 2 together with its matrix formulation. In Section 3 a necessary and sufficient condition for the existence of a solution is derived. This section also contains the method of perturbing the right side. Section 5 contains a direct method for solving the resulting consistent linear system of equations. There are presently two direct methods which compete for being the most efficient way to solve this system. The Buneman variant of the cyclic reduction algorithm and the Fourier series method [1, 3]. In a previous paper [5] it was shown how the Buneman algorithm could be adapted to solve Poisson's equation on a disk. In this paper we will discuss the Fourier series method applied to the surface of the sphere. It makes extensive use of the fast Fourier transform which is generally available in subroutine form. These direct methods are desirable both from a standpoint of speed and storage. The operation count for both is proportional to $mn \log n$ where m and n are the number of latitude and longitude points respectively. They require half the storage of iterative methods since the solution may be returned in the storage occupied by the right side of Poisson's equation.

2. The Discretization

We wish to determine an approximate solution of Poisson's equation defined on the surface of a sphere.

$$\frac{1}{\sin\theta}(\sin\theta u_\theta)_\theta + \frac{1}{\sin^2\theta} u_{\phi\phi} = f(\theta, \phi), \qquad 0 \leqslant \theta \leqslant \pi, \qquad 0 < \phi \leqslant 2\pi. \quad (2.1)$$

We place a net on the surface of the sphere by selecting integers m and n and defining net spacings $\Delta\theta = \pi/(m+1)$, $\Delta\phi = 2\pi/n$ and the net

$$\begin{aligned} \theta_i &= i\,\Delta\theta \qquad i = 0, \tfrac{1}{2}, 1, \ldots, m + \tfrac{1}{2}, m + 1, \\ \phi_j &= j\,\Delta\phi \qquad j = 1, 2, \ldots, n. \end{aligned} \quad (2.2)$$

We wish to determine values v_{ij} which approximate $u(\theta_i, \phi_j)$. To this end we require that the v_{ij} satisfy finite difference approximation to (2.1).

$$\frac{1}{\Delta\theta^2 \sin\theta_i}[\sin\theta_{i+1/2}(v_{i+1,j} - v_{i,j}) - \sin\theta_{i-1/2}(v_{i,j} - v_{i-1,j})]$$

$$+ \frac{1}{\Delta\phi^2 \sin^2\theta_i}(v_{i,j+1} - 2v_{i,j} + v_{i,j-1}) = f_{i,j}, \tag{2.3}$$

for $i = 1, 2,..., m$ and $j = 1, 2,..., n$.
If we define

$$a_i = \frac{\sin\theta_{i-1/2}}{\Delta\theta^2 \sin\theta_i},$$

$$b_i = \frac{\sin\theta_{i+1/2}}{\Delta\theta^2 \sin\theta_i}, \tag{2.4}$$

$$d_i = \frac{1}{\Delta\phi^2 \sin^2\theta_i}.$$

Then (2.3) can be written

$$a_i v_{i-1,j} - (a_i + b_i) v_{i,j} + b_i v_{i+1,j} + d_i(v_{i,j-1} - 2v_{i,j} + v_{i,j+1}) = f_{i,j}, \tag{2.5}$$

for $i = 1, 2,..., m$ and $j = 1, 2,..., n$ where $f_{i,j} = f(\theta_i, \phi_j)$. As a result of the periodicity in ϕ we have $v_{i,0} = v_{i,n}$ and $v_{i,n+1} = v_{i,1}$. We will denote by v_N the value common to all $v_{0,j}$, $j = 1, 2,..., n$ and v_S as the value of $v_{m+1,j}$ for $j = 1, 2,..., n$. Also we will denote $f_{0,j}$ by f_N and $f_{m+1,j}$ by f_S. We then have $nm + 2$ unknowns, namely v_N, v_S and $v_{i,j}$ for $i = 1, 2,..., m$ and $j = 1, 2,..., n$. Equation (2.5) represents mn equations. An additional equation is obtained by integrating (and subsequently discretizing) Eq. (2.1) over the spherical segment $\theta < \Delta\theta/2$.

$$\frac{4}{n\Delta\theta^2}\left(\sum_{j=1}^{n} v_{1,j} - nv_N\right) = f_N. \tag{2.6}$$

Similarly at $\theta = \pi$ we obtain

$$\frac{4}{n\Delta\theta^2}\left(\sum_{j=1}^{n} v_{m,j} - nv_S\right) = f_S, \tag{2.7}$$

which together with (2.5) and (2.6) provide $mn + 2$ equations. We can write the complete linear system of equations as

$$Av = f. \tag{2.8}$$

The vector v has length $mn + 2$ and is of the form $v^T = (v_N, v_1^T, v_2^T, \ldots, v_n^T, v_S)$ with $v_j^T = (v_{1,j}, v_{2,j}, \ldots, v_{m,j})$. The vector f has the form $f^T = (f_N, f_1^T, f_2^T, \ldots, f_n^T, f_S)$, where $f_j^T = (f_{1,j}, f_{2,j}, \ldots, f_{m,j})$.

If we let $\beta = 4/(n\Delta\theta^2)$ then A is an $mn + 2$ by $mn + 2$ matrix of the form

$$A = \left[\begin{array}{c|c|c} -\beta n & w^T & 0 \\ \hline x & B & y \\ \hline 0 & z^T & -\beta n \end{array}\right] \tag{2.9}$$

The vectors w, x, y, and z are of length mn and are of the form

$$w^T = (w_1^T, w_1^T, \ldots, w_1^T), \qquad \text{where} \qquad w_1^T = (\beta, 0, \ldots, 0)$$

has m components and is repeated n times. Similarly

$$\begin{aligned} x^T &= (x_1^T, x_1^T, \ldots, x_1^T), \quad \text{where} \quad x_1^T = (a_1, 0, \ldots, 0), \\ y^T &= (y_1^T, y_1^T, \ldots, y_1^T), \quad \text{where} \quad y_1^T = (0, \ldots, 0, b_m), \\ z^T &= (z_1^T, z_1^T, \ldots, z_1^T), \quad \text{where} \quad z_1^T = (0, \ldots, 0, \beta). \end{aligned} \tag{2.10}$$

The matrix B has order mn and is of the form

$$B = \begin{bmatrix} C - 2D & D & & & & & D \\ D & C - 2D & D & & & & \\ & \ddots & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & D & C - 2D & D & \\ D & & & & D & C - 2D & \end{bmatrix}, \tag{2.11}$$

where D is a diagonal matrix of order m, $D = \operatorname{diag}(d_1, d_2, \ldots, d_m)$ and C is an $m \times m$ tridiagonal matrix

$$C = \begin{bmatrix} -(a_1 + b_1) & b_1 & & & \\ a_2 & -(a_2 + b_2) & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{m-1} & -(a_{m-1} + b_{m-1}) & b_{m-1} \\ & & & a_m & -(a_m + b_m) \end{bmatrix}, \tag{2.12}$$

3. Least-Squares Solutions

In this section we will first give necessary and sufficient conditions on f for a solution to exist. In the event a solution does not exist we then show how to perturb f so that the resulting consistent system has a solution which is a least-squares solution to the unperturbed system.

THEOREM 1. *The linear system $Av = f$ has a solution if and only if $f^T h = 0$ where*

$$h^T = \left(\frac{n}{4}\sin\theta_{1/2}\,,\, h_1{}^T, h_1{}^T,\ldots, h_1{}^T, \frac{n}{4}\sin\theta_{m+1/2}\right)$$

and (3.1)

$$h_1{}^T = (\sin\theta_1\,,\,\sin\theta_2\,,\ldots,\,\sin\theta_m).$$

Proof. It can be shown that $A^T h = 0$ and therefore the rank of A is less than $mn + 2$. However, if we delete the first row and column of A then the resulting matrix is irreducibly diagonally dominant and therefore nonsingular [6, p. 23]. Hence the rank of A is $mn + 1$ and h spans the null space of A^T. This completes the proof since we know that $Av = f$ has a solution if and only if $f^T h = 0$ for every h such that $A^T h = 0$.

In practice, as a result of computational or observational errors, $f^T h \neq 0$, and the system does not have a solution. In this event, an acceptable option is to determine a least-squares solution. We will perturb f so that a solution to the resulting consistent system is a least-squares solution to the original inconsistent system. Further, the norm will be selected so that the perturbation is a scalar multiple of the vector $e^T = (1, 1,\ldots, 1)$.

If we define the matrix $H = \text{diag}\,(h)$ then H is positive definite, and we can define the inner product

$$(f, g)_H = f^T H g \tag{3.2}$$

and the induced H norm

$$\| u \|_H^2 = u^T H u. \tag{3.3}$$

THEOREM 2. *If v is a solution to the consistent system, $Av = g$ where*

$$g = f - \frac{(e, f)_H}{(e, e)_H}\, e \tag{3.4}$$

then v minimizes $\| Av - f \|_H$.

Proof.

$$A^T H A v = A^T H g,$$

$$A^T H A v = A^T H \left(f - \frac{(e,f)_H}{(e,e)_H} e \right),$$

$$A^T H A v = A^T H f.$$

A derivation of the normal equations shows that v minimizes $\| Av - f \|_H$ if and only if v is a solution to the consistent system $A^T H A v = A^T H f$.

Remark 1. v determines an equivalence class $\hat{v}$ of solutions since $v + \alpha e$ is also a solution for any scalar α.

Remark 2. The matrix HA is symmetric, hence for arbitrary vectors g and f, $(g, Af)_H = (Ag, f)_H$. Therefore the matrix A represents a self-adjoint operator under the inner product defined by (3.2).

Remark 3. For arbitrary f and g

$$\| g - f \|_H^2 = \left\| g - f + \frac{(e,f)_H}{(e,e)_H} e \right\|_H^2 - 2 \frac{(e,f)_H}{(e,e)_H} (e,g)_H + \frac{(e,f)_H^2}{(e,e)_H}, \tag{3.5}$$

Hence, subject to $(e, g)_H = 0$, this expression has a minimum value of $(e,f)_H^2/(e,e)_H$ for g defined by (3.4).

Remark 4. v is the least-square solution (l_2 norm) to the weighted system

$$H^{1/2} A = H^{1/2} f.$$

Remark 5. There is considerable freedom of choice for the perturbation of f. For example let P be any symmetric positive definite matrix and define $\epsilon = P^{-1} h$. Then a least-squares solution (P norm) is given as a solution to the consistent system

$$Av = f - \frac{(\epsilon, f)_P}{(\epsilon, \epsilon)_P} \epsilon,$$

where $(\epsilon, f)_P = \epsilon^T P f$.

Hence P can be selected to provide almost any functional dependence for ϵ. However, usually the functional dependence of the errors in f will not be known and therefore it is reasonable to perturb f by e which has no functional dependence.

If the usual l_2 norm is selected, then $\epsilon = h$ which is not constant and also approximates a function which is not regular at the poles.

We close this section with a summary of the calculations used to perturb f.

I. Compute the inner product of f with e

$$(e,f)_H = n \sin \theta_{1/2}/4(f_N + f_S) + \sum_{i=1}^{m} \sin i\,\Delta\theta \sum_{j=1}^{n} f_{ij}\,. \tag{3.6}$$

II. Compute $(e, e)_H$

$$(e, e)_H = n \sin \theta_{1/2}/4 + n(1 + \cos \theta_{1/2})/\sin \theta_{1/2}\,. \tag{3.7}$$

III. Define

$$\alpha = \frac{(e,f)_H}{(e, e)_H}\,,$$

then the elements of the vector g are computed from

$$\begin{aligned} g_{ij} &= f_{ij} - \alpha \qquad i = 1, 2,..., m; \quad j = 1, 2,..., n. \\ g_N &= f_N - \alpha. \\ g_S &= f_S - \alpha. \end{aligned}$$

4. Solutions of the Consistent System

In this section we assume that $h^T g = 0$ and hence the system $Av = g$ has a solution. The approach will be to deflate the system to a nonsingular system which can be solved by existing methods. For completeness, the Fourier series technique is described.

Let E denote the matrix which is obtained by replacing the first row of the identity matrix by the vector h^T. E is nonsingular since its determinant is $n \sin \theta_{1/2}/4$. Now if we multiply $Av = g$ by E we obtain

$$EAv = Eg. \tag{4.1}$$

The first element of Eg is $h^T g = 0$. The remaining elements are unchanged from the corresponding elements of g. Also the first row of EA is just $h^T A$ which is zero. The remaining rows are unchanged from those of A. To solve (4.1) we can arbitrarily set $v_N = 0$, then the system of order $mn + 1$, consisting of (4.1) with the first equation and the first variable v_N deleted, is nonsingular and can therefore be solved by a variety of methods. The solution to this system of reduced order, augmented with $v_N = 0$ is a solution of (4.1) and therefore of $Av = g$ since E is nonsingular. This deflation is equivalent to specifying v_N to be zero and then solving the system of reduced order. Note that Eq. (2.6) with f_N replaced by g_N, will be satisfied since the system is singular and (2.6) can be expressed as a linear com-

bination of the remaining equations which are satisfied. The nonuniqueness is demonstrated by the fact that v_N may be arbitrarily selected.

We will now describe the Fourier series method: The elements of the vector g can be expressed in the form

$$g_{i,j} = \sum_{k=0}^{L} G_{k,i}^{(1)} \cos k\phi_j + G_{k,i}^{(2)} \sin k\phi_j), \tag{4.2}$$

where for each i, the coefficients $G_{k,i}^{(1)}$ and $G_{k,i}^{(2)}$ may be obtained from a one dimensional fast Fourier transform (FFT) in the variable ϕ. If n is even then $L = n/2$; if n is odd then $L = (n-1)/2$.

We will determine a solution in the form

$$v_{i,j} = \sum_{k=0}^{L} (V_{k,i}^{(1)} \cos k\phi_j + V_{k,i}^{(2)} \sin k\phi_j), \tag{4.3}$$

where the coefficients $V_{k,i}^{(1)}$ and $V_{k,i}^{(2)}$ are evaluated in the following manner. If we substitute (4.2) and (4.3) into the finite difference equations (2.3) (with f_{ij} replaced by g_{ij}) and equate coefficients of $\cos k\phi_j$ and $\sin k\phi_j$ then

$$a_i V_{k,i-1}^{(l)} - (a_i + b_i) V_{k,i}^{(l)} + b_i V_{k,i+1}^{(l)} - \lambda_{k,i} V_{k,i}^{(l)} = G_{k,i}^{(l)}, \tag{4.4}$$

for

$$l = 1, 2; \qquad i = 1, 2, \ldots, m \qquad \text{and} \qquad k = 0, 1, \ldots, L,$$

where

$$\lambda_{k,i} = 2d_i(1 - \cos k\Delta\phi).$$

For each k and l, (4.4) represents a tridiagonal system of m equations in $m+2$ unknowns $V_{k,i}^{(l)}$ $i = 0, 1, \ldots, m+1$. Therefore (4.4) must be augmented with additional equations. Since (4.3) is constant at the poles we have

$$V_{k,0}^{(l)} = V_{k,m+1}^{(l)} = 0 \tag{4.5}$$

for $l = 1, 2$ and $k = 1, 2, \ldots, L$. For $k = 0$ the system (4.4) is augmented by $V_{0,0}^{(l)} = v_N$ which can be arbitrarily specified and

$$n\beta V_{0,m}^{(1)} - n\beta V_{0,m+1}^{(1)} = G_{0,m+1}^{(1)}, \tag{4.6}$$

which is obtained by substituting (4.2) and (4.3) into (2.7) with f_S replaced by g_S. Once the $V_{k,i}^{(l)}$ are determined by solving the tridiagonal systems (4.4) augmented by (4.5) and (4.6), then the solution $v_{i,j}$ is obtained by a fast Fourier synthesis of (4.3). Since the Fourier transforms require on the order of $n \ln n$ operations, the total operation count is proportional to $mn \ln n$.

We close this section with a table of computational results. The calculations were performed on the Control Data 7600 and times are in milliseconds. The error is the maximum absolute value of the difference between the solution of the finite difference equations and the computed solution.

Computational Results

n	m	Time	Error
32	15	13	4.33×10^{-13}
64	31	56	2.28×10^{-12}
128	63	208	4.39×10^{-11}

ACKNOWLEDGMENTS

I wish to thank Dr. Roland Sweet for many helpful ideas and also Dr. Fred Dorr whose thoughtful suggestions resulted in a much improved manuscript.

REFERENCES

1. B. L. BUZBEE, G. H. GOLUB, AND C. W. NIELSON, On direct methods for solving Poisson's equations, *SIAM J. Numer. Anal.* 7 (1970), 627–656.
2. J. F. DALPHIN AND V. LOVASS-NAGY, Best least squares solutions to finite difference equations using the generalized inverse and tensor product method, *J. Assoc. Comput. Mach.* **20** (1973), 279–289.
3. F. W. DORR, The direct solution of the discrete Poisson equation on a rectangle, *SIAM Rev.* **12** (1970), 248–263.
4. V. LOVASS-NAGY, D. L. POWERS, AND F. D. ULLMAN, A modified ADI method for computing the "best least-squares solution of an incompatible system $(A \times I + I \times B)x = g$," *Linear Algebra and Its Applications* **7** (1973), 179–185.
5. P. N. SWARZTRAUBER AND R. A. SWEET, The direct solution of the discrete Poisson equation on a disk, *SIAM J. Numer. Anal.* **10** (1973), 900–907.
6. R. S. VARGA, "Matrix Iterative Analysis," Prentice-Hall, Englewood Cliffs, NJ, 1962.