Convergence Rates of ADI Methods with Smooth Initial Error

By Robert E. Lynch* and John R. Rice**

1. Introduction. In this paper*** we investigate the Peaceman-Rachford Alternating Direction Implicit (ADI) method for solving elliptic partial difference equations with parameters chosen in such a way as to exploit smoothness properties of the initial error.

The first result of this type was obtained by Guilinger [4]. One interpretation of it is as follows. Consider the partial difference equation derived from the two-dimensional Poisson problem on a convex polygon by the use of the usual five-point approximation on a mesh with spacing h. Make a certain smoothness assumption on the initial error for the ADI method with one parameter ρ. Then, given $\epsilon > 0$, there is a number K and a parameter ρ, *both independent of* h (h sufficiently small), such that by the Kth iterate of the ADI scheme, the error has been reduced by a factor of at least ϵ. That is to say, the asymptotic rate of convergence of this ADI process with constant ρ is a constant independent of h. Guilinger uses a different normalization of the parameter than other authors; with the usual normalization, the parameter ρ does depend on h.

We study the effect of smooth initial conditions with the aid of tensor product analysis [5]–[7] and extend and clarify Guilinger's result. Our study is both theoretical and experimental. In the theoretical part we: (1) exhibit more clearly the role of the smoothness of the initial error; (2) relate the choice of parameters to a certain approximation problem; (3) establish the existence, uniqueness, and characterization results for this approximation problem; (4) obtain precise asymptotic results for the case of one parameter and the Poisson problem; and (5) establish a general, but imprecise, extension of Guilinger's result. We have also developed a computational algorithm for calculating the smooth optimum parameters; it is a modification of the one in [3] and the details of it are not included here.

The use of tensor products is successful because the problems considered are *separable*. In particular, this restricts our analysis to equations of special form (as in Eq. (2.2) below) and to rectangular regions. Thus, our results are weaker than Guilinger's in this respect. It has been conjectured by several people that there exist bounds on the convergence rate for the ADI method applied to the Poisson

Received June 8, 1966. Revised April 5, 1967.

* Part of the work was done at the General Motors Research Laboratories, Warren, Michigan. Also, support was given, in part, by the National Science Foundation through Grant GP-217 and by the Army Research Office (Durham) through Grant DA-ARO(D)-31-124-G388, at The University of Texas.

** Purdue University. Part of this work was done at the General Motors Research Laboratories, Warren, Michigan.

*** Some preliminary results were announced at the Society of Industrial and Applied Mathematics Symposium on Matrix Computation, Gatlinburg, Tennessee, April 13–18, 1964; other results were presented at the Summer Meeting of the American Mathematical Society, Ithaca, New York, September 3, 1965 (Notices Amer. Math. Soc. **12** (1965), 585).

problem on a convex region R in terms of the convergence rate on the smallest rectangular region containing R. If this conjecture is true, then our results are immediately applicable to the case considered by Guilinger.

In the numerical experiments we: (1) study the nature of the optimum parameters and corresponding deviations; (2) study the actual effect of various sets of parameters for the Poisson problem with different initial estimates of the solution; (3) compare the optimum parameters with discrete optimum parameters for the specific matrix of the problem; (4) study the error reduction as a function of initial conditions and parameter sets, as the mesh length tends to zero; and (5) study the effect on the error reduction of uncertainty in the value of $\alpha = \lambda_1/\lambda_N$, the ratio of the smallest to the largest eigenvalue associated with the problem.

A variety of conclusions are presented in the final section. The most important is that, with smooth initial error, the number of iterations necessary to achieve a given error reduction is independent of mesh length h. Other important conclusions are: (1) the Wachspress parameters (4.4) should be used unless precise information about the eigenvalues and initial error is available;* (2) the initial error should be as smooth as possible; (3) it is better to underestimate than to overestimate the ratio α for any of the sets of parameters; (4) an explanation of why the Birkhoff, Varga, and Young experiments [2] do not show that their optimum parameters are optimum.

For problems with smooth solutions, smooth initial error can be achieved by use of a proper interpolation procedure. We do not discuss this procedure in detail here. We give the following example of one possible approach. For Dirichlet boundary conditions on a rectangle, initial values in the interior can be obtained from a function with two continuous derivatives by using bicubic spline** interpolation (see de Boor [12] for interpolation formulas). Then, since both the initial estimate and the solution have two continuous derivatives, so does their difference. Smoother initial error can be obtained by use of higher degree polynomial spline functions. If one has such an interpolation formula for functions of one variable, then, of course, the two-variable formula on a square is obtained by forming the tensor product of the one-dimension formula with itself.

2. Problem Statement. Let R denote the unit square $\{(x, y) | 0 < x < 1, 0 < y < 1\}$, and let $\bar{R}$ denote its closure. Consider a second-order linear elliptic partial differential equation in two independent variables

$$\mathfrak{L}v = f, \qquad (x, y) \in R, \tag{2.1}$$

subject to given Dirichlet conditions on the boundary of R. We assume that the operator $\mathfrak{L}$ has the special form

$$\mathfrak{L} = \mathfrak{L}_x + \mathfrak{L}_y, \tag{2.2}$$

* Such information is available for the Dirichlet problem for the five-point approximation to the Laplacian operator on a rectangle. For this and similar cases for which this information is available, the smooth parameter theory yields better computational results, as is illustrated in Fig. 5.

** A bicubic spline interpolation function is a piecewise bicubic function with two continuous derivatives in the interior of the rectangle which takes on the prescribed values at the boundary mesh points. Thus in each mesh rectangle, $x_m \leqq x \leqq x_{m+1}$, $y_n \leqq y \leqq y_{n+1}$, the function has the form $\Sigma_{i,j=1}^{3} a_{ij}x^i y^j$.

where $\mathfrak{L}_x$ [$\mathfrak{L}_y$] is a second-order linear differential operator with derivatives with respect to x [y] only and with coefficients which depend on x [y] only. A difference approximation to $\mathfrak{L}_x$ [$\mathfrak{L}_y$] can be obtained by subdividing the unit interval by mesh points $0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1$ [$0 = y_0 < y_1 < \cdots < y_M < y_{M+1} = 1$] and by replacing derivatives in $\mathfrak{L}_x$ [$\mathfrak{L}_y$] with divided differences. The solution of the ordinary nonhomogeneous differential equation $\mathfrak{L}_x w = f_1$ [$\mathfrak{L}_y w = f_2$] with w given at 0 and 1 is approximated at mesh points x_i [y_j] by the solution of a linear system $Aw_1 = b$ [$Bw_2 = c$] where w_1 [w_2] is a vector and A [B] is an $N \times N$ [$M \times M$] matrix. A system of linear equations with solution which approximates the solution of (2.1) at mesh points (x_i, y_j) in R can then be obtained by means of tensor products (Kronecker products) [5], [6] as

$$(I_M \otimes A + B \otimes I_N)u = g\,, \tag{2.3}$$

where I_M, I_N are identity matrices of order M and N, respectively. The components of the vector u are taken as approximations to the values of v at mesh points (x_i, y_j); these components are ordered so that the $([i-1]N + j)$th component corresponds to $v(x_i, y_j)$. The vector g in (2.3) involves values of the right side of (2.1) as well as boundary values. For more details and for a description of the use of tensor products, see [5], [6] and [7].

For a given set of positive parameters, ρ_k, $k = 1, 2, \cdots$, the Peaceman-Rachford ADI iteration scheme [8] for obtaining a solution of (2.3) can be written as [3], [5]

$$u^{(0)} \quad \text{given}\,,$$

$$\begin{aligned} [I_M \otimes (A + \rho_{k+1}I_N)]u^{(k+1/2)} &= g - [(B - \rho_{k+1}I_M) \otimes I_N]u^{(k)}\,, \\ [(B + \rho_{k+1}I_M) \otimes I_N]u^{(k+1)} &= g - [I_M \otimes (A - \rho_{k+1}I_N)]u^{(k+1/2)}\,. \end{aligned} \tag{2.4}$$

Let $e^{(k)} = u^{(k)} - u$ denote the error of the kth iterate, it satisfies

$$e^{(0)} = u^{(0)} - u\,,$$

$$\begin{aligned} [I_M \otimes (A + \rho_{k+1}I_N)]e^{(k+1/2)} &= -[(B - \rho_{k+1}I_M) \otimes I_N]e^{(k)}\,, \\ [(B + \rho_{k+1}I_M) \otimes I_N]e^{(k+1)} &= -[I_M \otimes (A - \rho_{k+1}I_N)]e^{(k+1/2)}\,. \end{aligned} \tag{2.5}$$

We assume that A and B have complete sets of normalized eigenvectors p_i and q_j, respectively. Then, the matrices P and Q whose columns are p_i and q_j are such that

$$P^{-1}AP = \Lambda(A)\,, \qquad Q^{-1}BQ = \Lambda(B)\,, \tag{2.6}$$

where $\Lambda(A)$, $\Lambda(B)$ denote diagonal matrices of eigenvalues λ_i, μ_j of A, B, respectively. It is easily verified that the eigenvectors of $(I_M \otimes A + B \otimes I_N)$ are $q_i \otimes p_j$ and that $(\mu_i + \lambda_j)$ are the associated eigenvalues. The error $e^{(m)}$ can be expanded in terms of these eigenvectors as

$$e^{(m)} = \sum_{i=1}^{M} \sum_{j=1}^{N} a_{ij}^{(m)} q_i \otimes p_j\,. \tag{2.7}$$

By substitution into (2.5) and simplification, the error $e^{(m)}$ can be expressed in terms of the coefficients of the initial error as

$$e^{(m)} = \sum_{i=1}^{M} \sum_{j=1}^{N} a_{ij}^{(0)} \prod_{k=1}^{m} \left[\frac{\lambda_j - \rho_k}{\lambda_j + \rho_k} \frac{\mu_i - \rho_k}{\mu_i + \rho_k} \right] q_i \otimes p_j . \tag{2.8}$$

We use two measures of error. The uniform norm is

$$\|e^{(m)}\|_\infty = \max_{k,l} |(e^{(m)})_{k,l}|$$

in which $(e^{(m)})_{k,l}$ denotes the error at the point (x_k, y_l). The discrete L_2-norm for $e^{(m)}$ on an equally spaced grid: $x_{i+1} - x_i = 1/(M+1)$ and $y_{j+1} - y_j = 1/(N+1)$, is

$$\|e^{(m)}\|_2 = \left[\sum_{k=1}^{M} \sum_{l=1}^{N} (e^{(m)})_{k,l}^2 \Big/ (M+1)(N+1) \right]^{1/2} .$$

Heretofore, for given m, "optimum parameters," $\rho_1, \rho_2, \cdots, \rho_m$, have been defined [2], [6] as those numbers which minimize the maximum value of

$$\left| \prod_{k=1}^{m} \left(\frac{x - \rho_k}{x + \rho_k} \frac{y - \rho_k}{y + \rho_k} \right) \right|$$

for $\lambda_1 \leqq x \leqq \lambda_N$, $\mu_1 \leqq y \leqq \mu_M$. Denoting the value of this minimax by d_m, one obtains a bound on $e^{(m)}$ as

$$(\|e^{(m)}\|_2)^2 = \sum_{i,j} [a_{ij}^{(m)}]^2 \leqq d_m^2 NM \max_{i,j} |a_{ij}^{(0)}|^2 . \tag{2.9}$$

But the variation of the coefficients $a_{ij}^{(0)}$ with i and j has been neglected; equivalently, the assumption has been made that each value $a_{ij}^{(0)}$ has equal importance. If some additional information concerning these coefficients is available, then a different set of parameters can be obtained which yields a smaller upper bound on $||e^{(m)}||_2$.

To study this, we impose the following condition on the initial error.

SMOOTHNESS ASSUMPTION. *There are two given positive and bounded functions* $\omega_1(i)$, $\omega_2(j)$ *such that*

$$|a_{ij}^{(0)}| \leqq \omega_1(i)\omega_2(j) . \tag{2.10}$$

To show why (2.10) is a condition on the smoothness of the initial error and to illustrate the nature of suitable functions ω we discuss the following.

Example. Suppose $-\mathcal{L}$ is the Laplace operator: $\mathcal{L}v = -\partial^2 v/\partial x^2 - \partial^2 v/\partial y^2$, and that the mesh is uniform, $x_i = y_i = i/(N+1)$, $i = 1, \cdots, N$, and that divided central differences are used to obtain the standard five-point approximation to $\mathcal{L}v = f$. Then, the eigenvalues λ_i and normalized eigenvectors p_i of the matrix A $(=B)$ are known [6]:

$$\lambda_i = [\sin^2(\theta_i/2)]/[\sin^2(\theta_N/2)] , \qquad \theta_i = i\pi/(N+1) , \qquad i = 1, \cdots, N , \tag{2.11}$$

$$p_i = [2/N]^{1/2} (\sin\theta_i, \sin 2\theta_i, \cdots, \sin N\theta_i) ,$$

where the matrix A is normalized so that its largest eigenvalue, λ_N, is unity and the eigenvectors are normalized with respect to the norm $||\cdot||_2$.

Since we want to use Fourier series and because the *difference equation is separable*, it is *natural* to expand $e^{(0)}$ in terms of the tensor products, $q_i \otimes p_j$, of the

eigenvectors.* Suppose the solution u has a convergent Fourier expansion and consider the components of the initial approximation, $u^{(0)}$, to be values of a given function which also has a convergent Fourier series on the interior of $\bar{R}$. For large N, the coefficients of the low frequency terms of the finite Fourier expansion of $e^{(0)}$ (on the mesh) will be nearly the same as the corresponding coefficients of the Fourier expansion of $e^{(0)}$ (on R). There is a well-known relation between the smoothness of a function and the asymptotic behavior of its Fourier coefficients. In particular, [10, p. 50] if a periodic function of one variable has a continuous $(p-1)$th derivative and if its pth derivative has bounded variation, then its Fourier coefficients a_j are $O(1/j^{p+1})$ as j tends to infinity (this also holds for $p = 0$ [9, p. 18]). Now, suppose that the components of $u^{(0)}$ are taken to be equal to values of a constant function on the interior of $\bar{R}$ and that, as usual, $e^{(0)} = 0$ at boundary points. Then one would have that, at worst, $a_{ij}^{(0)} = O(1/ij)$, because $e^{(0)}$ would be discontinuous only at the boundary of $\bar{R}$. Furthermore, if the error were smoother, then $a_{ij}^{(0)}$ tends to zero faster as i and j tend to infinity. For example, if the error were $xy(1-x)(1-y)$, then $a^{(0)} = O(1/[ij]^3)$.

Observe that to determine the smoothness at the boundary, it must be specified how functions which are zero on the boundary of $\bar{R}$ are extended to be periodic functions on the whole (x, y)-plane. This extension is made by observing that as N tends to infinity, components of the vectors $p_i/\sqrt{2}$ tend to values of $\sin(i\pi x)$ so that the error (2.7) is expanded in terms of products of sines; consequently, any function F on R is to be extended to a function on the whole (x, y)-plane which is antisymmetric about $x = 0$ and $y = 0$ and periodic: $F(x+2, y+2) = F(x, y)$.

Below we refer specifically to the following three initial errors on the unit square $\bar{R}$:

$$
\begin{aligned}
&e_1^{(0)}(x,y) = 1\,, \\
&e_2^{(0)}(x,y) = C_2 \min[x, y, 1-x, 1-y]\,, \qquad (x,y) \in R\,; \\
&e_3^{(0)}(x,y) = C_3 xy(1-x)(1-y)\,,
\end{aligned} \tag{2.12}
$$

$$
e_i^{(0)}(x,y) = 0\,, \qquad (x,y) \in \bar{R} - R\,, \qquad i = 1,2,3\,.
$$

The normalizing factors are $C_2 = 2$, $C_3 = 16$ for the uniform norm and $C_2 = 6$, $C_3 = 36$ for the L_2-norm. The Fourier expansions of these initial errors are given by

$$
\begin{aligned}
e_1^{(0)}(x,y) &= \frac{16}{\pi^2} \sum_{\text{odd } m,n} \frac{1}{mn} \sin(m\pi x) \sin(n\pi y)\,, \\
e_2^{(0)}(x,y) &= C_2 \frac{4}{\pi^2} \sum_{\text{odd } n} \frac{1}{n^2} \sin(n\pi x) \sin(n\pi y)\,, \\
e_3^{(0)}(x,y) &= C_3 \frac{24}{\pi^6} \sum_{\text{odd } m,n} \frac{1}{m^3 n^3} \sin(m\pi x) \sin(n\pi y)\,.
\end{aligned} \tag{2.13}
$$

* For the example of the Laplace operator, $q_i = p_i$. Note that the subsequent analysis is *not* restricted to the case of the five-point approximation to the Laplacian, but it applies to any separable difference equation approximation to a separable differential equation.

The first two of these satisfy the Smoothness Assumption $\omega_1(i) = \omega_2(i) = 1/i$, and the third satisfies $\omega_1(i) = \omega_2(i) = 1/i^3$.

Note that the assumption that the $|a_{ij}^{(0)}|$ are of equal importance, as in (2.9), is equivalent to the Smoothness Assumption $\omega_1(i) = \omega_2(i) = 1$ and, moreover, if $a_{ij}^{(0)} = 1$, the sine series with these coefficients is not convergent.

Returning to the more general formulation, we have, with the Smoothness Assumption, that

$$(\|e^{(m)}\|_2)^2 \leqq \sum_{i=1}^{M} \left[\omega_1(i) \prod_{k=1}^{m} \frac{\mu_i - \rho_k}{\mu_i + \rho_k}\right]^2 \sum_{j=1}^{N} \left[\omega_2(j) \prod_{k=1}^{N} \frac{\lambda_j - \rho_k}{\lambda_j + \rho_k}\right]^2$$

$$\leqq MN \max_i \left[\omega_1(i) \prod_{k=1}^{m} \frac{\mu_i - \rho_k}{\mu_i + \rho_k}\right]^2 \max_j \left[\omega_2(i) \prod_{k=1}^{m} \frac{\lambda_j - \rho_k}{\lambda_j + \rho_k}\right]^2 ; \tag{2.14a}$$

$$\|e^{(m)}\|_\infty \leqq NM \max_i \left| \omega_1(i) \prod_{k=1}^{m} \frac{\mu_i - \rho_k}{\mu_i + \rho_k} \right| \max_j \left| \omega_2(j) \prod_{j=1}^{m} \frac{\lambda_j - \rho_k}{\lambda_j + \rho_k} \right| . \tag{2.14b}$$

We can now attempt to determine "optimum parameters," $\rho_1, \cdots, \rho_m$ which minimize the expressions on the right side of (2.14). Clearly, the "optimum parameters" in [2] and [6] are the parameters for the special case that ω_1 and ω_2 are taken as constants and $(\omega_1\omega_2)^2 = \max_{ij}|a_{ij}^{(0)}|^2$. One should keep in mind that here, as in many cases, one minimizes an *upper bound* for $||e^{(m)}||_2$ and not $||e^{(m)}||_2$ itself.

In order to proceed with the analysis, we *assume* that the eigenvalues λ_j, μ_i are real, positive, and distinct and that the system (2.3) has been normalized* so that

$$\max [\lambda_N, \mu_M] = 1 . \tag{2.15}$$

The eigenvalues are functions of i and j; because of the assumption that they are distinct, the inverse functions exist:

$$j = \lambda^{-1}(\lambda_j) , \qquad i = \mu^{-1}(\mu_i) . \tag{2.16}$$

In particular, for the example cited above (see (2.11)),

$$\lambda^{-1}(s) = [2(N + 1)/\pi] \arcsin [(s\nu)^{1/2}] , \tag{2.17}$$

where $\nu = \sin^2 [N\pi/2(N + 1)]$ is the normalization factor.

Set

$$\alpha = \min [\lambda_1, \mu_1] \tag{2.18}$$

and extend λ^{-1}, defined on the discrete set $\lambda_1 < \lambda_2 < \cdots < \lambda_N$, to be a continuous strictly monotonic increasing function on the interval $\alpha \leqq s \leqq 1$; also, extend ω, defined on 1, 2, $\cdots$, N, to be a continuous positive and bounded function on the interval $1 \leqq t < \infty$. Denote by $E_2^{(m)}$ the following,

$$E_2^{(m)} = \max_{\alpha \leqq z \leqq 1} \left[\omega_2(\lambda^{-1}(z)) \left| \prod_{k=1}^{m} \frac{z - \rho_k}{z + \rho_k} \right| \right] . \tag{2.19}$$

Then, by defining $E_1^{(m)}$ with μ in place of λ, we have

* Guilinger [4] used the normalization $\min [\lambda_1, \mu_1] = 1$. This does make certain aspects of the approximation problem easier to treat, but we use the more usual normalization (2.15).

$$(2.20) \qquad (||e^{(m)}||_2)^2 \leqq E_1{}^{(m)}E_2{}^{(m)}, \qquad ||e^{(m)}||_\infty \leqq NME_1{}^{(m)}E_2{}^{(m)}.$$

Thus, the question of "optimum parameters" under the Smoothness Assumption leads to the approximation problem of determining parameters $\rho_1, \rho_2, \cdots, \rho_m$ which minimize $E_2{}^{(m)}$. We now treat this problem.

3. Approximation Theory. For simplicity of notation, set

$$(3.1) \qquad E_m(P, z) = \omega(z) \prod_{k=1}^{m} \frac{z - \rho_k}{z + \rho_k}, \qquad \alpha \leqq z \leqq 1,$$

where P stands for the set of parameters

$$(3.2) \qquad P = \{\rho_k | 0 < \rho_k < \infty, k = 1, \cdots, m\}.$$

The Chebyshev approximation problem is: *Given $f(x)$ continuous on $[\alpha, 1]$, determine a parameter set P^* such that*

$$\max_{\alpha \leqq z \leqq 1} |E_m(P^*, z) - f(z)| \leqq \max_{\alpha \leqq z \leqq 1} |E_m(P, z) - f(z)|$$

for all P. The functions $E_m(P^*, z)$ are, by definition, *best approximations* to $f(z)$ and the *deviation d_m of $E_m(P^*, z)$* is, by definition,

$$d_m = \max_{\alpha \leqq z \leqq 1} |E_m(P^*, z) - f(z)|.$$

A function $h(z)$ is said to *alternate k times* on a set Z if there is a subset $\{z_i | z_i \in Z, z_1 < z_2 < \cdots < z_{k+1}\}$ such that

$$|h(z_i)| = \max_{z \in Z} |h(z)|, \qquad h(z_i) = -h(z_{i+1}), \qquad i = 1, \cdots, k.$$

The points in $\alpha \leqq z \leqq 1$ where the value $|E_m(P^*, z) - f(z)|$ is equal to the deviation are called extremal points. We consider these extremal points z_i to be ordered by $z_i < z_{i+1}$.

The theory of Chebyshev approximation by rational functions is well developed [6]–[8] and can be applied directly here to establish the following.

THEOREM 1. *If $\omega(z)$ is continuous and positive on $\alpha \leqq z \leqq 1$, then, for any P, $E_m(P, z)$ is locally unisolvent of degree m.*

COROLLARY 1. (A) *There is at most one best approximation.* (B) *A necessary and sufficient condition that $E_m(P^*, z)$ is the best approximation to $f(z)$ is that $E_m(P^*, z) - f(z)$ alternate at least m times on $\alpha \leqq z \leqq 1$.*

The domain of the parameter P is not closed; hence, we cannot conclude immediately that every continuous function possesses a best approximation. Nevertheless, because the function $|(z - \rho)/(z + \rho)|$ is a monotonic decreasing function of ρ for $z > \rho$ and a monotonic increasing function of ρ for $z < \rho$, we can conclude that a best approximation does exist for $f \equiv 0$.

COROLLARY 2. *There exists a best approximation to the function $f(z) \equiv 0$, $\alpha \leqq z \leqq 1$, and all the parameters ρ_k associated with P^* are such that $\alpha \leqq \rho_k \leqq 1$.*

Together, these two corollaries state that there exists one and only one set of parameters, P^*, which minimize $E_2{}^{(m)}$ defined in (2.19); moreover, the alternation of $E_m(P^*, z)$ is a property which characterizes these parameters.

4. Asymptotic Results. In this section we obtain precise asymptotic results for the optimum parameters and deviations as the mesh length tends to zero for the special case of one parameter. We also show for the multiple parameter case that if the initial error satisfies a certain Smoothness Assumption then the number of iterations required to reduce the error to less than a given $\epsilon > 0$ is independent of the mesh length h as h tends to zero.

We now treat a specific case for one parameter which is applicable to the Example in Section 2. We assume

$$\begin{aligned} &\text{(i) } m = 1\,, \\ &\text{(ii) } \alpha = [\pi/2(N+1)]^2\,, \\ &\text{(iii) } \lambda^{-1}(z) = [2(N+1)/\pi]\arcsin[(z)^{1/2}]\,, \qquad \alpha \leqq z \leqq 1\,, \\ &\text{(iv) } \omega(t) = Ct^q\,, \qquad q \leqq -1\,, \qquad C = [\lambda^{-1}(\alpha)]^{-q}\,, \qquad \text{i.e., } \omega(\lambda^{-1}(\alpha)) = 1\,. \end{aligned} \tag{4.1}$$

Note that the eigenvalue ratio λ_1/λ_N in the Example in Section 2 differs from α by $O(1/N^4)$ as N tends to infinity (see (2.11)).

THEOREM 2. *Let ρ^* denote the parameter of best approximation. If the assumptions* (4.1) *hold, then*

(i) $\operatorname{Lim}_{N\to\infty} E_1(P^*, 1) = 0$.

(ii) $E_1(P^*, \alpha) < 1$.

(iii) $z = \alpha$ *is the extremal point z_1 and the second extremal point z_2 satisfies*

$$4\rho^*[z(1-z)]^{1/2}\sin^{-1}(z^{1/2}) + q(z^2 - \rho^{*2}) = 0\,.$$

(iv) $\alpha < \rho^* < z_2 = -\rho^*(2 + [q^2+4]^{1/2})/q + O(\rho^*)^2$.

(v) $\operatorname{Lim}_{N\to\infty} \rho^* = 0$.

(vi) ρ^* *satisfies*

$$\alpha - \rho^* + K[\rho^*/\alpha]^{q/2}[\alpha + \rho^*] + O(\alpha^2) = 0\,,$$

where

$$K = \frac{z_2 - 1}{z_2 + 1} z_2^{q/2} > 0\,.$$

(vii) *If $q = -1$, then $\rho^* = 1.6\alpha + O(\alpha^2)$ and the deviation, $d_1(q)$, is $d_2(-1) = .23 + O(\alpha)$.*

(viii) *If $q = -2$, then $\rho^* = 1.3\alpha + O(\alpha^2)$ and the deviation, $d_2(q)$, is $d_2(-2) = .13 + O(\alpha)$.*

(ix) $\operatorname{Lim}_{q\to-\infty} \alpha/\rho^* = 1$, $\operatorname{Lim}_{q\to-\infty} d_1(q) = 0$.

Proof. (i) Since

$$E_1(P^*, x) = \omega(\lambda^{-1}(x))\frac{x - \rho^*}{x + \rho^*}$$

and $(x - \rho^*)/(x + \rho^*) < 1$, $\omega(\lambda^{-1}(1)) = C[2(N+1)/\pi]^q$, it follows that

$$E_1(P^*, 1) \sim (N+1)^q$$

and, since $q \leqq -1$, the result is established.

(ii) We have $\omega(\lambda^{-1}(\alpha)) = 1$ and the choice of $\rho = 1/2$ leads to $||E_1(1/2, x)||_\infty < 1$ for $\alpha \leqq x \leqq 1$.

(iii) We note that $E_1(\rho^*, x)$ has only one zero, ρ^*, and only one extremum between α and infinity. We find the coordinate z_2 of this extremum by differentiation. A manipulation leads to

$$4\rho^*(z_2(1 - z_2))^{1/2} \arcsin(z_2^{1/2}) + q(z_2^2 - \rho^{*2}) = 0 .$$

(iv), (v), and (vi) We consider the equation

$$E_1(\rho^*, \alpha) = -E_1(\rho^*, z_2) . \tag{4.2}$$

There is only one choice of ρ^* and z_2 which satisfies (4.2) and (iii). We show there is a choice of ρ^* and z_2 which satisfies these and for which ρ^*/α, z_2/α are bounded for large N. These values then must be the desired quantities.

With z_2/α bounded we can replace $\arcsin(z_2^{1/2})$ with $(z_2)^{1/2} + O(\alpha^2)$ and $(1 - z_2)^{1/2}$ with $1 + O(\alpha)$ in (iii). Then, with terms $O(\alpha)$ omitted, the solution of the resulting quadratic equation is

$$z_2 = \frac{2\rho^*}{4\rho^*/3 - q} \{1 \pm [1 - q(4\rho/3 - q)]^{1/2}\} .$$

Recall $q \leqq -1$ so that we need to take the plus sign on the square root. Since ρ^* is small, this relation is the one in (iv).

We substitute this value into (4.2) and, after considerable manipulation, obtain

$$1 - \rho^*/\alpha + K(\rho^*/\alpha)^{q/2} + K(\rho^*/\alpha)^{q/2+1} = 0 \tag{4.3}$$

where

$$K = (z_2 - 1)(z_2)^{q/2}/(1 + z_2) + O(\alpha) .$$

It is clear that (4.3) can only be satisfied if ρ^*/α is bounded and hence if (v) holds. We see that (4.3) has a root by evaluating the left side for $\rho^* = \alpha$ (where it is positive) and for ρ^*/α large (with ρ^* still small compared to q). Thus, for N sufficiently large, there is a choice of ρ^* and z_2 which satisfies (4.2) and (iii), and ρ^*/α, z_2/α are bounded.

The final three points are verified by direct calculation from the preceding relationships.

It is interesting to note that one can obtain arbitrarily high accuracy with *one iteration* of ADI, but one needs a *"sufficiently smooth" initial error*, i.e. $-q$ sufficiently large *and* $\omega(\alpha) = 1$. This reflects the fact that it is nontrivial to construct initial estimates of u so that $e^{(0)}$ is "sufficiently smooth."

We now treat the problem of achieving a given accuracy with a fixed number of iterations, independent of the number of mesh points. The following assumptions are sufficient to establish such a result. These assumptions are satisfied by many separable partial difference equation problems which approximate separable partial differential equation problems of mathematical physics, such as the Example of Section 2.

(4.4) i. The matrices A and B in (2.3) are positive definite; A is $N \times N$ and B is $M \times M$, and each has a set of distinct eigenvalues: $\lambda_1^{(N)} < \lambda_2^{(N)} < \cdots < \lambda_N^{(N)}$ and $\mu_1^{(M)} < \mu_2^{(M)} < \cdots < \mu_M^{(M)}$, respectively.

(4.4) ii. For any fixed index i, the ratio $\lambda_i^{(N)}/\lambda_1^{(N)}$ is uniformly bounded for all meshes considered, i.e. for all N.

(4.4) iii. There is a continuous function $\Lambda^{-1}(t)$, $1 \leqq t < \infty$, and a positive constant $\sigma \leqq 1$ such that $\sigma i \leqq \Lambda^{-1}(\lambda_i^{(N)}/\lambda_1^{(N)}) \leqq i$ for all N.

(4.4) iv. The sets of positive parameters used, $\rho_1^{(N)}, \rho_2^{(N)}, \cdots$, are such that there are positive integers s, J independent of N and $1 \leqq J$ (N sufficiently large) such that for each $r = 0, 1, \cdots$, there is a positive integer $p < s$ such that $\lambda_1^{(N)} < \rho_{sr+p}^{(N)} < \lambda_J^{(N)}$.

(4.4) v. For each mesh considered, the coefficients $a_{ij}^{(0)}$ (which depend on M and N) of the initial error satisfy the Smoothness Assumption (2.10) for fixed functions $\omega_1(i)$, $\omega_2(i)$, $i = 1, 2, \cdots$, which satisfy the following:

Given $\epsilon > 0$, there is a positive integer K_0 and constant γ such that

$$\sum_{i=1}^{\infty} \omega_1^2(i) \leqq \gamma < \infty , \qquad \sum_{j=K_0}^{\infty} \omega_2^2(j) < \epsilon .$$

THEOREM 3. *With the assumptions* (4.4), *given* $\epsilon > 0$, *there is a positive integer* L *such that* $||e^{(L)}||_2 < \epsilon$ *uniformly for all meshes considered.*

Proof. (We suppress the superscripts (N) on λ_j and ρ_k.) We have the following bound on $||e^{(l)}||_2$ from (2.14a) and (4.2v) if $\rho_k > 0$

$$(\|e^{(l)}\|_2)^2 \leqq \gamma \sum_{j=1}^{N} \left[\omega_2(j) \prod_{k=1}^{l} \frac{\lambda_j - \rho_k}{\lambda_j + \rho_k} \right]^2 .$$

From (4.4v) it follows that there is a K_0 such that

$$\sum_{j=K_0}^{\infty} \omega_2^2(j) < \epsilon^2/2\gamma .$$

Set $K = \max \{[K_0/\sigma] + 1, J\}$ (for σ see (4.4iii) and for J see (4.4iv)); then

$$(\|e^{(l)}\|_2)^2 \leqq \gamma \sum_{j=1}^{K} \omega_2^2(j) \prod_{k=1}^{l} \left[\frac{\lambda_j - \rho_k}{\lambda_j + \rho_k} \right]^2 + \frac{\epsilon^2}{2} .$$

For ρ_{sr+p} such that $\lambda_1 < \rho_{sr+p} < \lambda_K$,

$$\max_{1 \leqq j \leqq K} \left| \frac{\lambda_j - \rho_{sr+p}}{\lambda_j + \rho_{sr+p}} \right| \leqq \max \left(\left| \frac{\lambda_1 - \rho_{sr+p}}{\lambda_1 + \rho_{sr+p}} \right| , \left| \frac{\lambda_K - \rho_{sr+p}}{\lambda_K + \rho_{sr+p}} \right| \right) \leqq \eta ,$$

where

$$\eta = (\lambda_K - \lambda_1)/(\lambda_K + \lambda_1) ,$$

and η is positive and, by (4.4ii), it is less than unity, independent of N. Therefore,

$$\prod_{k=1}^{l} \left(\frac{\lambda_j - \rho_k}{\lambda_j + \rho_k} \right)^2 \leqq \eta^{[l/s]}$$

and for some value L of l we have

$$\eta^{[L/s]} < \epsilon^2 \Big/ \left[2\gamma \sum_{j=1}^{K} \omega_2^2(j) \right] ,$$

which implies that $||e^{(L)}||_2 < \epsilon$ and concludes the proof.

The estimates made in the proof are extremely crude, and a smaller value of

L can be obtained with a judicious choice of the parameters ρ_k. However, we have not been able to analyze the multiple-parameter problem sufficiently well to obtain results similar to Theorem 2. The experimental results reported in Section 7 indicate that analogous results are true.

We also obtain a similar result for the uniform norm of the error.

THEOREM 4. *With the assumptions* (4.4), *but replace the inequalities in* (4.4v) *with*

$$\sum_{i=1}^{\infty} \omega_1(i) \leqq \gamma < \infty , \quad \sum_{j=K_0}^{\infty} \omega_2(j) < \epsilon , \tag{4.5}$$

given $\epsilon > 0$, *there is a positive integer* L *such that* $||e^{(L)}||_\infty < \epsilon$ *uniformly for all meshes considered.*

The proof is similar to that for Theorem 3 and we omit it.

Remark. Neither the unit error function e_1 nor the pyramid function e_2 in (2.12) satisfies (4.5). But, for e_2, $a_{ij}^{(0)} = \delta_{ij}/i^2$ and the proof of Theorem 4 can be easily modified to treat this special case; the result is the same and $||e^{(L)}||_\infty < \epsilon$ for the example of Section 2 provided parameters which satisfy (4.4iv) are used.

The Wachspress parameters $\rho_i^{(W)}$, $i = 1, \cdots, m$, are given by

$$\rho_{m-k+1}^{(W)} = \lambda_N \alpha^{(k-1)/(m-1)} , \qquad k = 1, \cdots, m . \tag{4.6}$$

Since $\rho_1 = \alpha$, the Wachspress parameters satisfy assumption (4.4iv) with $s = m$ when cycles of them are used: $\rho_1^{(W)}, \cdots, \rho_m^{(W)}, \rho_1^{(W)}, \cdots, \rho_m^{(W)}, \rho_1^{(W)}, \cdots$. Because of this, and by the preceding remark, we have the following.

COROLLARY. *Suppose cycles of Wachspress parameters are used to reduce an initial error* e_i, $i = 1, 2,$ *or* 3 *of* (2.12) *in the example of Section* 2. *Then, given* $\epsilon > 0$ *there are integers* L_i, $i = 1, 2, 3,$ *such that* $||e_i^{(L_i)}||_2 < \epsilon$, $i = 1, 2, 3,$ *independent of* N, *and there are integers* M_2, M_3 *such that* $||e_i^{(M_i)}||_\infty < \epsilon$, $i = 2, 3,$ *independent of* N.

5. Nature of Parameters and Deviations. For a given smoothness assumption, optimum parameters can be computed efficiently with the Remes algorithm [9] by a modification of a procedure of de Boor and Rice [3]. The computation is made easier by a change of independent variable:*

$$z = \alpha^{1-v^2} , \qquad \alpha \leqq z \leqq 1 . \tag{5.1}$$

But, even when v is used instead of z, caution must be used to obtain fairly accurate initial estimates of the optimum parameters so that the iteration converges. Table 1 contains parameters for weight functions $\omega(i)$ equal to 1, $1/i$, $1/i^2$, and $1/i^3$ for a single value of α and for several values of m.

The results of Theorems 2, 3, and 4 can be explained heuristically as follows for the smoothness condition $\omega(i) = i^{-q}$ with $q \geqq 1$ and for initial values of u at mesh points chosen to be values of given function f defined on R. The initial error is expressed as

* The change of variables can also be written as $y = \alpha^{-v^2}$, $y = z/\alpha$, $1 \leqq y \leqq 1/\alpha$. In effect, Guilinger [4] uses the variable y.

$$e^{(0)} = \sum_{i,j=1}^{N} a_{ij}^{(0)} q_i \otimes p_j$$

and as N tends to infinity $a_{ij}^{(0)}$ tends to the Fourier coefficients of the initial error $u - f$. Because of the Smoothness Assumption, for large N the major contribution to the error occurs in the modes for which both i/N and j/N are small. In the case of the example in Section 2, for fixed i, $|\lambda_i - \lambda_1|$ (see (2.11)) tends to zero as N tends to infinity, i.e., the eigenvalues coalesce. For large N, a single parameter ρ_1 suitably chosen close to λ_1 makes the factor $(\lambda_i - \rho_1)/(\lambda_i + \rho_1)$ very small for all those i which contribute most to the error. Thus, the use of ρ_1 causes a large decrease in the amplitudes of those Fourier components; the other error components are already small because of the Smoothness Assumption.

TABLE 1

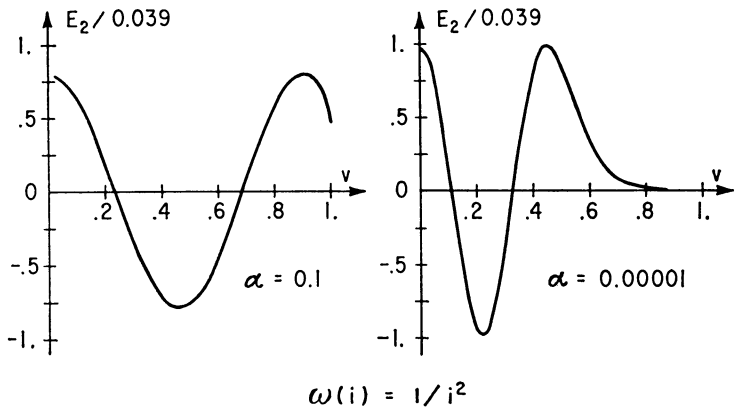
*Optimum Parameters, α = .00154 37125**

m	$\omega(i) = 1$**	$\omega(i) = 1/i$	$\omega(i) = 1/i^2$	$\omega(i) = 1/i^3$
		(Deviations in parentheses)		
1	.15691 8	.00996 970	.00803 294	.00738 883
	(.92)	(.24)	(.13)	(.090)
2	.02242 57	.00802 725	.00709 381	.00678 300
	1.09800	.04644 93	.02123 81	.01500 02
	(.56)	(.10)	(.039)	(.020)
4	.00938 813	.00707 865	.00663 058	.00647 621
	.05890 57	.01777 43	.01139 43	.00948 425
	.41801 3	.07253 30	.02956 80	.01954 20
	2.62282	.49016 8	.11947 5	.05948 92
	(.16)	(.028)	(.0064)	(.0022)
6	.00753 579	.00673 313	.00647 416	.00637 289
	.02242 57	.01238 41	.00933 635	.00823 053
	.08161 90	.03156 24	.01764 89	.01320 49
	.30168 6	.09620 75	.04036 27	.02527 28
	1.09800	.35457 7	.11239 7	.05796 41
	3.26752	1.73149	.44512 2	.18037 2
	(.046)	(.0084)	(.0015)	(.00040)
8	.00692 432	.00654 842	.00639 328	.00632 062
	.01416 34	.01007 07	.00841 113	.00766 372
	.03620 90	.01997 74	.01362 89	.01094 59
	.09609 81	.04486 95	.02534 29	.01774 41
	.25623 1	.10959 2	.05229 88	.03188 12
	.68003 3	.29209 8	.12053 0	.06363 19
	1.73852	.86612 3	.32516 3	.14683 2
	3.55607	2.73555	1.16852	.44809 5
	(.013)	(.0024)	(.00044)	(.00009 1)

* For the five-point approximation to the Laplacian on a unit square this value of α corresponds to $N + 1 = 1/h = 40$.

** For $m = 1$, 2, and 4, these values differ from those listed in [2]; we believe our values to be correct to the number of digits listed.

Graphs of $E_m(P^*, z)$ versus z indicate the same behavior for $m > 1$. For example, Fig. 1 shows the error function $E_2(P^*, z)$ for $\omega(i) = 1/i^2$, for two values of α, plotted as a function of v (see (5.1)); thus the z-axis has been stretched near $z = \alpha$ to show the oscillations there. Observe that in contrast to the case $\omega(i) = 1$, even for $\alpha = .1$ ($N = 5$ in the Example of Section 2), the point $z = 1$ is not an extremal point. The rightmost extremal point, z_3, is in the *interior* of $[\alpha, 1]$. Note that for $z > z_3$ the error function $E_2(P^*, z)$ decreases because of the Smoothness Assumption. Note also that in Fig. 1 the vertical scales of the two graphs are the same; the deviation decreases only slightly as α decreases. Graphs of $E_m(P^*, z)$ for $m > 2$ are similar, but instead of having two zeros as in Fig. 1, they have m zeros.



One of the most significant features of the functions $E_m(P^*, z)$, as functions of v, is the regular oscillation from the first extremal point $z_1 = \alpha$ to the last, z_m. Indeed, the extremal points are almost exactly linearly distributed along the v-axis. We do not elaborate on this distribution, but only point out that $y_m = z_m/\alpha$ is a well-behaved function of m. Table 2 gives values of y_m for several values of m and α for $\omega(i) = 1/i^2$.

TABLE 2

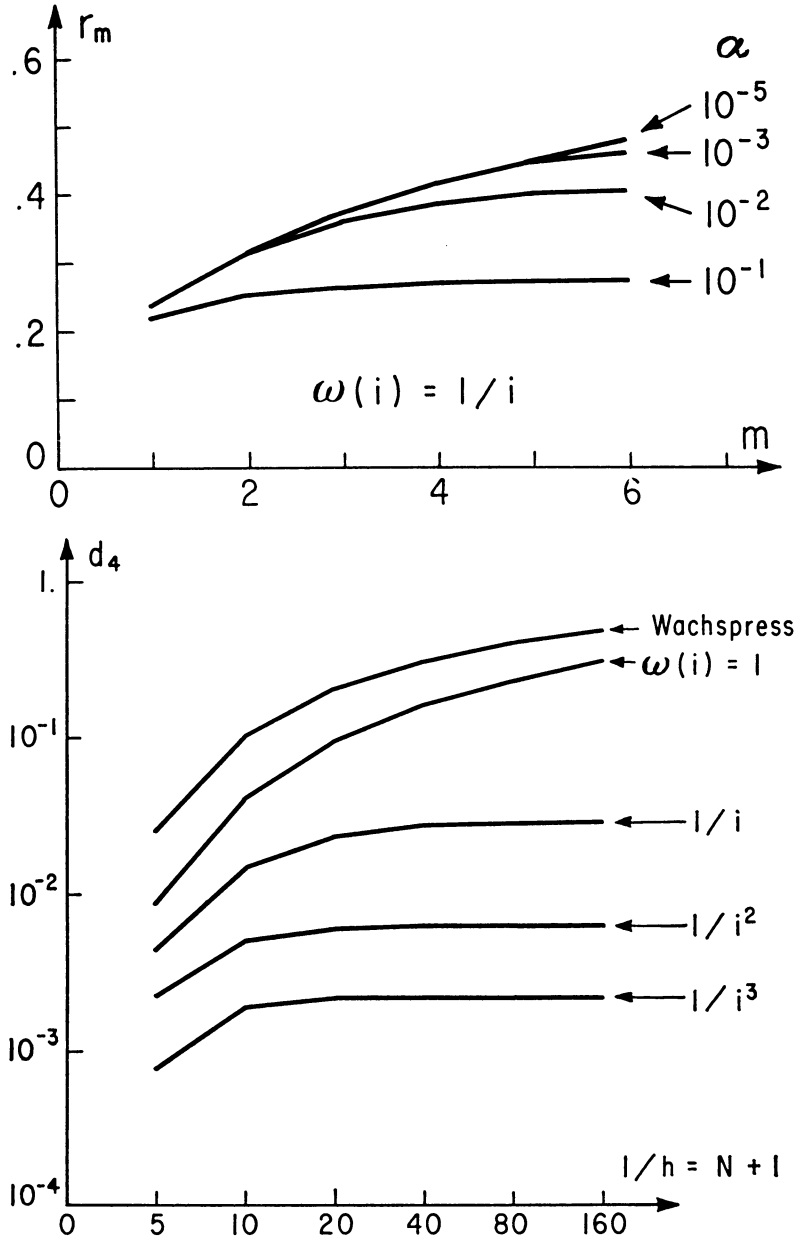
Ratio of $y_m = z_m/\alpha$ of Rightmost Extreme Point to α for $\omega(i) = 1/i^2$

α	$m = 1$	2	3	4	5	6	9
1	1.0	1.0	1.0	1.	1.	1.	1.
.01	3.0	9.6	23.5	45.	68.	88.	97.
.001	3.2	10.3	26.	59.	115.	218.	700.
~ 0	3.2	10.3	26.	59.	123.	..	..

The dependence on m, α, and the Smoothness Assumption of the deviation is summarized by the graphs in Figs. 2A and 2B. In Fig. 2A, the average deviation per ADI sweep, $r_m = [d_m]^{1/m}$, is plotted versus m for several values of α for $\omega(i) = 1/i$. The principal facts illustrated by this graph are: (1) The asymptotic value, as $\alpha \downarrow 0$, of r_m is reached quickly. (2) There is a diminishing return in increasing the number of parameters; that is to say, the reduction in the error *per iteration* decreases with the number of parameters. Similar graphs with Smoothness

Assumption $\omega(i) = i^q$ for different values of $q < 0$ show that, apparently, the asymptotic limit as $\alpha \downarrow 0$ is reached faster as $(-q)$ increases.

In Fig. 2B the deviation is plotted versus $N + 1 = 1/h$ (the relationshhip to α is given in (2.12) for the Example in Section 2) for several values of q for $m = 4$. This shows that for $q = -1$, -2, and -3, the deviation is *bounded below unity* independent of N. For comparison, the deviation for four Wachspress parameters is also plotted.



6. Limitations in Use of Parameters. We stress four limitations in using parameters based on some smoothness assumption.

First, as is well known, there is *not* a simple relationship between the deviations in the approximation problem and the amount that the error will be reduced in an actual calculation. This is because the deviations together with other quantities enter into an *upper bound* which was derived by some rather crude estimates. Thus, while the (smooth) "optimum parameters" are definitely superior to parameters which assume $\omega(i) = 1$, we cannot conclude that they are the only efficient set of parameters for smooth initial errors.

Second, application of m sweeps of ADI with optimum parameters $\rho_1, \rho_2, \cdots, \rho_m$ changes the distribution of the error components.* If initially $|a_{ij}^{(0)}| \leqq \omega_1(i)\omega_2(j)$ where $\omega_1(1) = \omega_2(1) = 1$, then after m sweeps, one has $|a_{ij}^{(m)}| \leqq d_m(1)d_m(2)$. Because the (smooth) optimum parameters are concentrated near α, one still expects that $|a_{ij}^{(m)}| \leqq K\omega_1(i)\omega_2(j)$, but the coefficient K will, in general, not be $d_m(1)d_m(2)$. Furthermore, for small i/N, j/N the error components are more nearly equally important than for the initial error. Consequently, two successive applications of optimum parameters: $\rho_1, \rho_2, \cdots, \rho_m, \rho_1, \rho_2, \cdots, \rho_m$ *are unlikely to* reduce the error twice as much as one application, so that one does *not necessarily* have $||e^{(2m)}|| \leqq (d_m(1)d_m(2))^2$.

Let $G_{m,j}$ denote the error function after j cycles of m parameters; then

$$G_{m,j}(P, x) = E_m(P, x) \prod_{i=1}^{m} [(x - \rho_i)/(x + \rho_i)]^{j-1}$$

and the deviation $d_{m,j}$ after j cycles of optimum parameters is

$$d_{m,j} = \max_x |G_{m,j}(P^*, x)| .$$

For example, with $\alpha = 0.00001$ and $m = 2$, $G_{2,1} = E_2$ is shown in Fig. 1 and $d_{2,1}$ is taken on by $G_{2,1}$ at three values of x (or v). Graphs of $G_{2,j}$, $j = 2$ and 5 are shown in Fig. 3, and $d_{2,j}$ is taken on at only one value of x; note that in Fig. 3 the scales on the two y-axes are different. Values of $d_{2,j}$ and the points x_j for which $d_{2,j} = |G_{2,j}(P^*, x_j)|$ are listed in Table 3.

TABLE 3

Deviations $d_{2,j}$ and Extremal Points x_k after j Cycles of Two Optimum Parameters with $\omega(i) = 1/i^2$ and $\alpha = 0.00001$

j	$d_{2,j}$	x_k
1	.0396	.0000100, .0000175, .000103
2	.0198	.000200
3	.0133	.000287
4	.00997	.000370
5	.00797	.000481

Since the distribution of the error components is not known after the first m sweeps, then if the error is not sufficiently small after m sweeps it seems to be

* This observation was made by Dr. Bernard Kripke in the spring of 1964.

advisable to follow a cycle of m (smooth) optimum parameters by a cycle of parameters obtained on the assumption that all error components are equally important. We have used, and recommend for use, the Wachspress parameters (see (4.6)).

Third, for a given problem, instead of using the optimum parameter P^* which solves the approximation problem of Section 3, one would like to use the *discrete* optimum parameter set, $Q^* = \{\sigma_i\}_1^m$, which minimizes $F_m(Q)$ defined by

$$F_m(Q) = \max_{z_i \in \Lambda} |F_m(Q, z_i)|, \tag{6.1a}$$

$$F_m(Q, z_i) = \omega(\lambda^{-1}(z_i)) \prod_{k=1}^{m} \frac{z_i - \sigma_k}{z_i + \sigma_k}, \tag{6.1b}$$

where Λ denotes the set of eigenvalues of the matrix A (or B). To illustrate the difference between P^* and Q^* again consider the example of Section 2 and take $N = 39$ (39×39 interior mesh points, $h = 1/40$). Table 4 contains values of σ_i for several m; compare these with values in Table 1. Also, compare the parameters with the eigenvalues listed in Table 5. For small m, ρ_i and σ_i are nearly equal, but as m increases, this is not so. Specifically, for $m = 8$, $\omega(i) = i^{-3}$, there are four values of ρ_i between λ_1 and λ_2, and the deviation of the optimum parameter error curve is 16 times larger than that of the discrete optimum parameter curve. Fig. 4 shows the error curves $E_8(P^*, z)$ and $F_8(Q^*, z)$ plotted as functions of v (see (4.1)). Finally, note that for $m = N$, the discrete optimum parameters are given by $\sigma_i = \lambda_i$ so that m sweeps with these parameters annihilates the error [5].

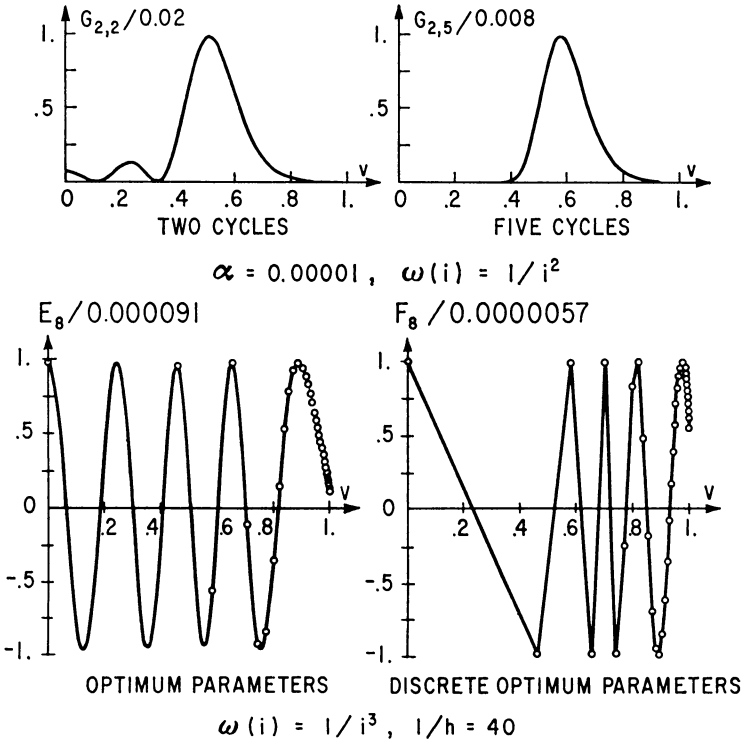


TABLE 4

Discrete Optimum Parameters for the 5-Point Approximation to the Laplacian on a Square with $N + 1 = 1/h = 40$ ($\alpha = .00154\ 37125$)

m	$\omega(i) = 1$	$\omega(i) = 1/i$	$\omega(i) = 1/i^2$	$\omega(i) = 1/i^3$
		(*Deviations in parentheses*)		
1	.15691 8	.00989 250	.00797 088	.00708 149
	(.92)	(.23)	(.13)	(.069)
2	.02242 03	.00783 639	.00663 236	.00632 127
	1.09826	.05296 70	.03530 12	.03087 79
	(.56)	(.094)	(.026)	(.0083)
4	.00936 885	.00661 830	.00623 107	.00617 708
	.05928 24	.03145 10	.02688 07	.02551 20
	.41968 2	.11205 0	.07194 46	.06487 66
	2.62539	.72808 2	.27648 6	.17133 1
	(.16)	(.021)	(.0027)	(.00045)
6	.00703 647	.00627 923	.00617 786	.00616 671
	.03239 73	.02694 39	.02523 42	.02477 74
	.10037 0	.06917 40	.06129 54	.05781 12
	.35699 2	.18452 5	.12638 9	.11294 4
	1.20984	.60862 3	.33018 3	.22988 9
	3.32967	2.33703	1.16917	.66998 5
	(.038)	(.0044)	(.00044)	(.00004 5)
8	.00637 169	.00618 796	.00616 753	.00616 552
	.02710 33	.02525 64	.02476 50	.02465 05
	.06766 94	.06035 81	.05721 27	.05590 44
	.15012 8	.11847 0	.10890 3	.10315 5
	.36153 5	.25281 8	.19338 7	.17406 5
	.86523 7	.57420 8	.40350 4	.31064 7
	1.98346	1.41132	.94676 9	.66710 5
	3.63452	3.2394	2.5054 7	1.73554
	(.0075)	(.00075)	(.00006 9)	(.00000 57)

TABLE 5

Eigenvalues of the A Matrix for the 5-Point Approximation to the Laplacian on a Square with $N + 1 = 1/h = 40$

.00616 533	.70110 4	2.15692	3.52081
.02462 33	.82442 9	2.31287	3.61803
.05526 02	.95500 3	2.46689	3.70528
.09788 70	1.09202	2.61803	3.78201
.15224 1	1.23463	2.76537	3.84776
.21798 7	1.38197	2.90798	3.90211
.29472 0	1.53311	3.04500	3.94474
.38196 6	1.68713	3.17557	3.97538
.47918 8	1.84308	3.29890	3.99383
.58578 6	2.00000	3.41421	

Fourth, since the largest contribution to the error normally occurs in the low-frequency Fourier components, one would expect to get even greater error reduction by always choosing α as one of the values of ρ_i, that is, make λ a root of the error curve instead of an extremal point.

7. Experimental Studies on the Square. The Peaceman-Rachford ADI scheme, (2.4), was used to solve the difference equations formed by the five-point equal-spaced difference approximation to Laplace's equation. For this, the A $(=B)$ matrix is tridiagonal with elements $a_{ii} = 2/h^2$, $a_{ij} = -1/h$, $|i - j| = 1$, $i = 1, \cdots, N$, $h = 1/(N + 1)$. As in the Birkhoff, Varga, Young experiments [2], boundary conditions

$$u(x, y) = 0, \qquad \text{for } x = 0, 1, \quad 0 \leqq y \leqq 1, \text{ and } 0 \leqq x \leqq 1, \quad y = 0, 1, \tag{7.1}$$

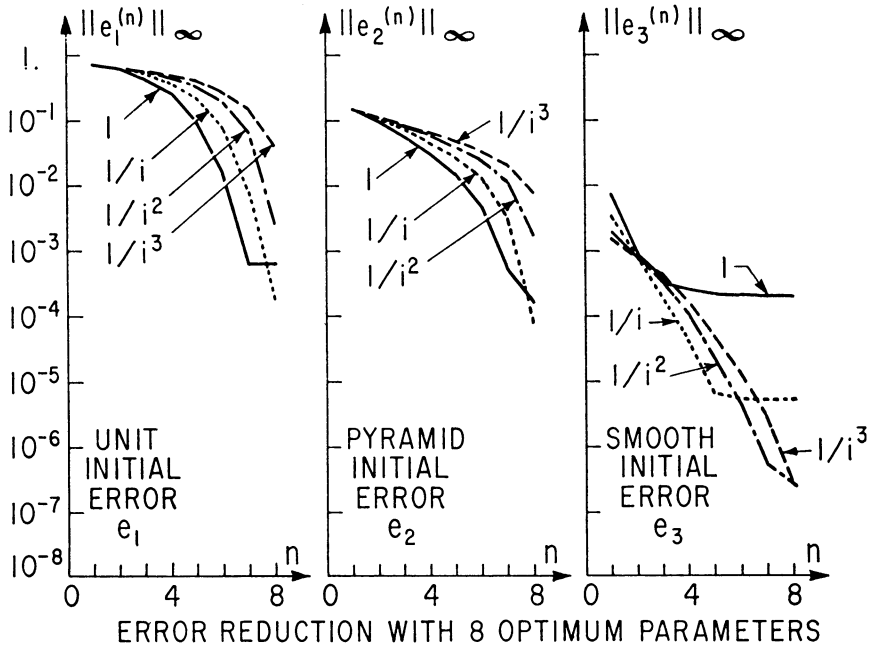
were used so that the components of $u^{(m)}$ were the errors $e^{(m)}$. We report on experiments in which the maximum error of the n iteration, $||e^{(n)}||_\infty = \max_{ij}|e_{ij}|$, was determined for different parameter sets for the initial errors (2.12):

$$u_i(x, y) = e_i(x, y), \qquad i = 1, 2, 3. \tag{7.2}$$

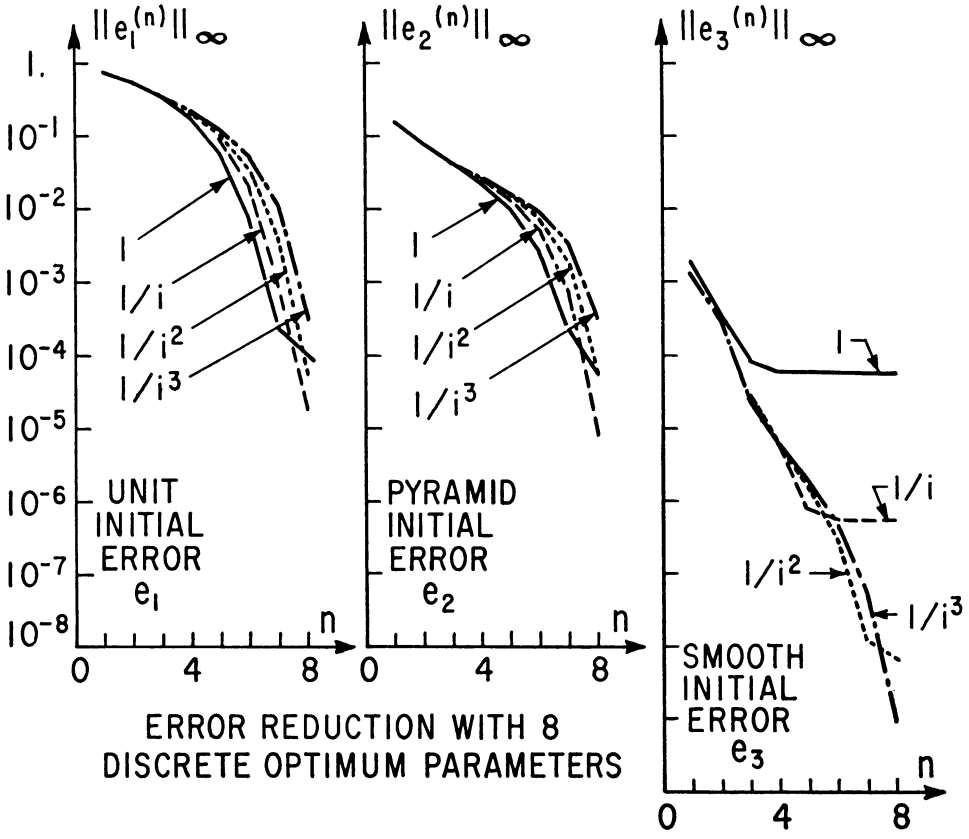
The Fourier expansions of these initial functions are given by (2.13).

Figs. 5A and 5B show plots of $||e^{(n)}||_\infty$ versus n for the three initial errors of (2.12) for both optimum parameters and discrete optimum parameters for $\omega(i) = 1, 1/i, 1/i^2$, and $1/i^3$, and for $m = 8$.

Fig. 6 shows the effect of the ordering of the parameters.* In one case an increasing order was used: $\rho_1 < \rho_2 < \cdots < \rho_m$, and in the second case the reverse order was used.



* With the exception of the experiments to obtain the data for Fig. 6, the parameters were always used in increasing order: $\rho_1 < \rho_2 < \cdots < \rho_m$, in each cycle.


$\|e_1^{(n)}\|_\infty$
$\|e_2^{(n)}\|_\infty$
$\|e_3^{(n)}\|_\infty$
1.
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
1
1/i
$1/i^2$
$1/i^3$
UNIT INITIAL ERROR e_1
PYRAMID INITIAL ERROR e_2
SMOOTH INITIAL ERROR e_3
n
0
4
8
ERROR REDUCTION WITH 8 DISCRETE OPTIMUM PARAMETERS

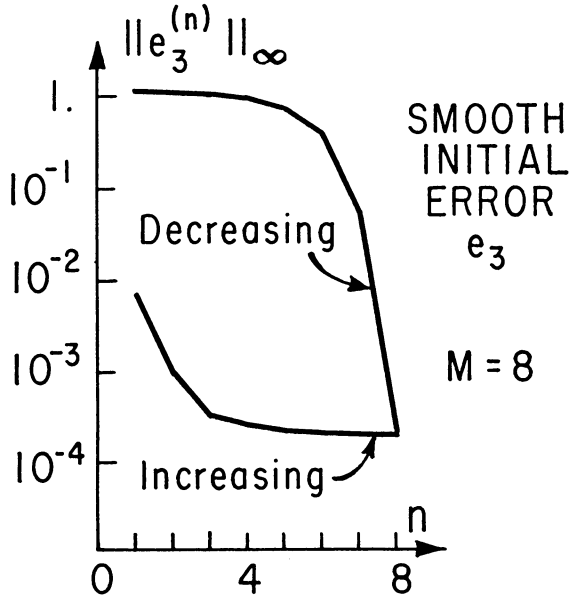

$\|e_3^{(n)}\|_\infty$
1.
10^{-1}
10^{-2}
10^{-3}
10^{-4}
Decreasing
Increasing
SMOOTH INITIAL ERROR e_3
M = 8
n
0
4
8

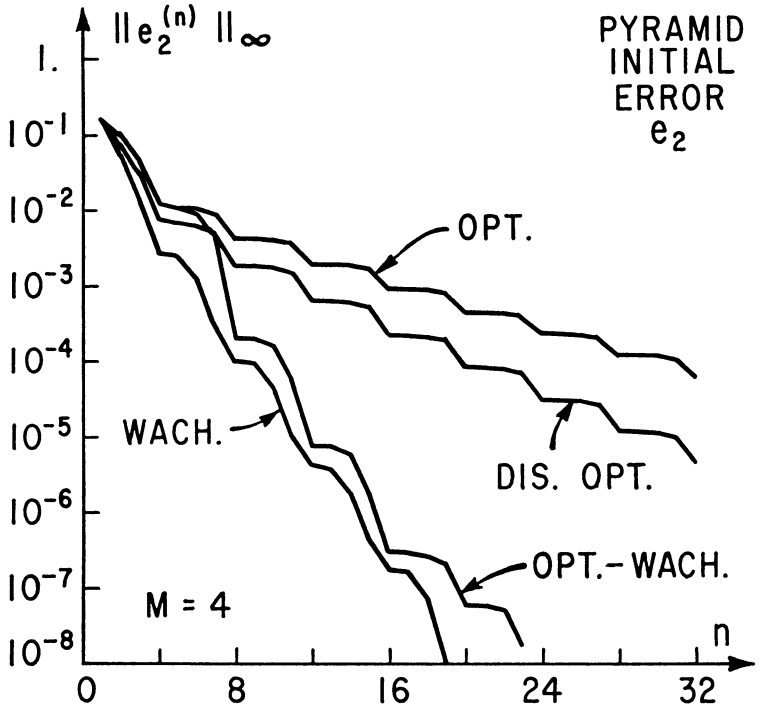
Fig. 7 shows the effect of applying:

i. Successive cycles of optimum parameters with $\omega(i) = 1/i^2$: "OPT."

ii. One cycle of optimum parameters followed by success cycles of Wachspress parameters (4.4): "OPT.-WACH."

iii. Successive cycles of discrete optimum parameters: DIS. OPT.

iv. Successive cycles of Wachspress parameters: WACH.



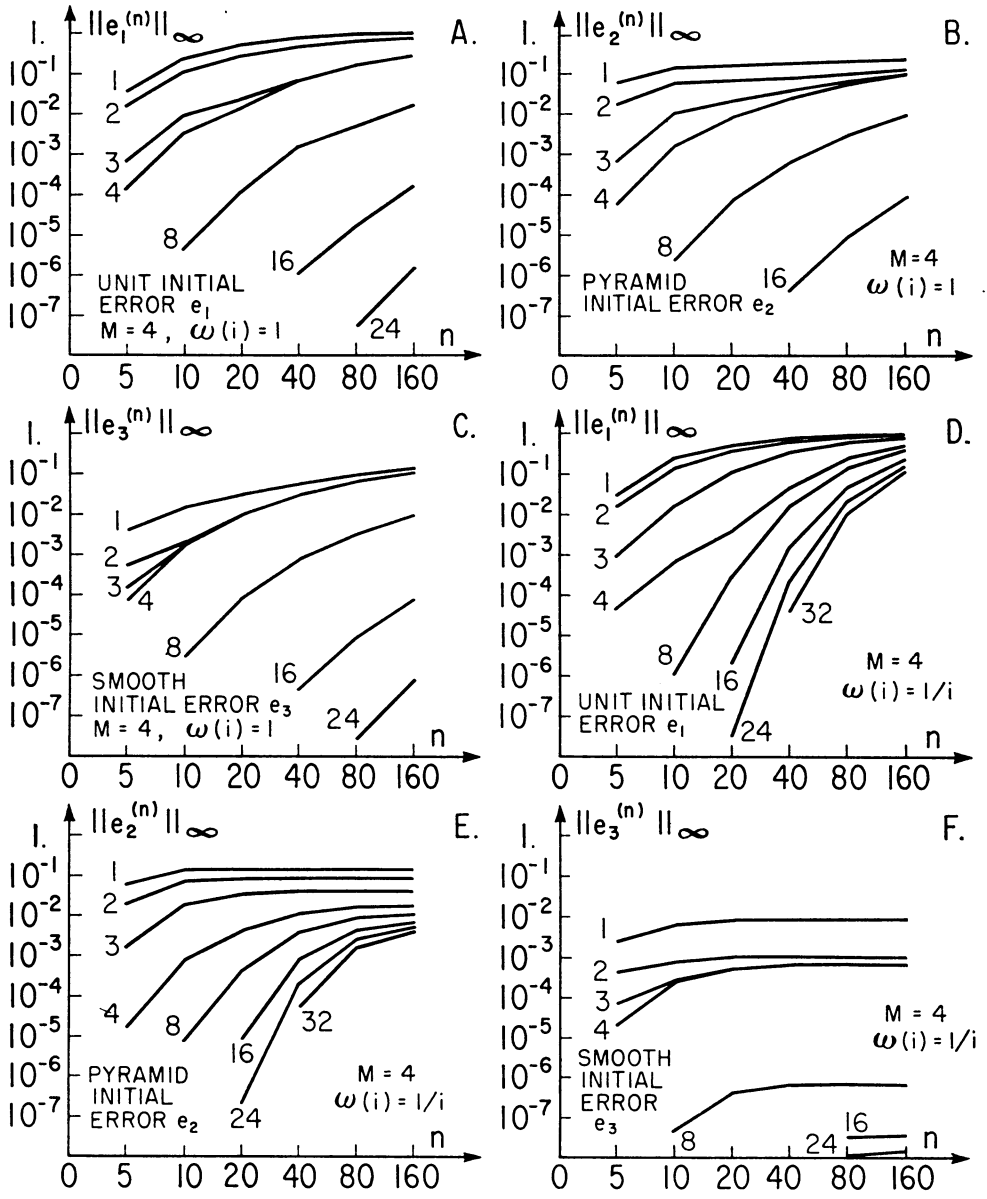
In each case the initial error was the pyramid function $e_2 = u_2$ (2.12) and $m = 4$.

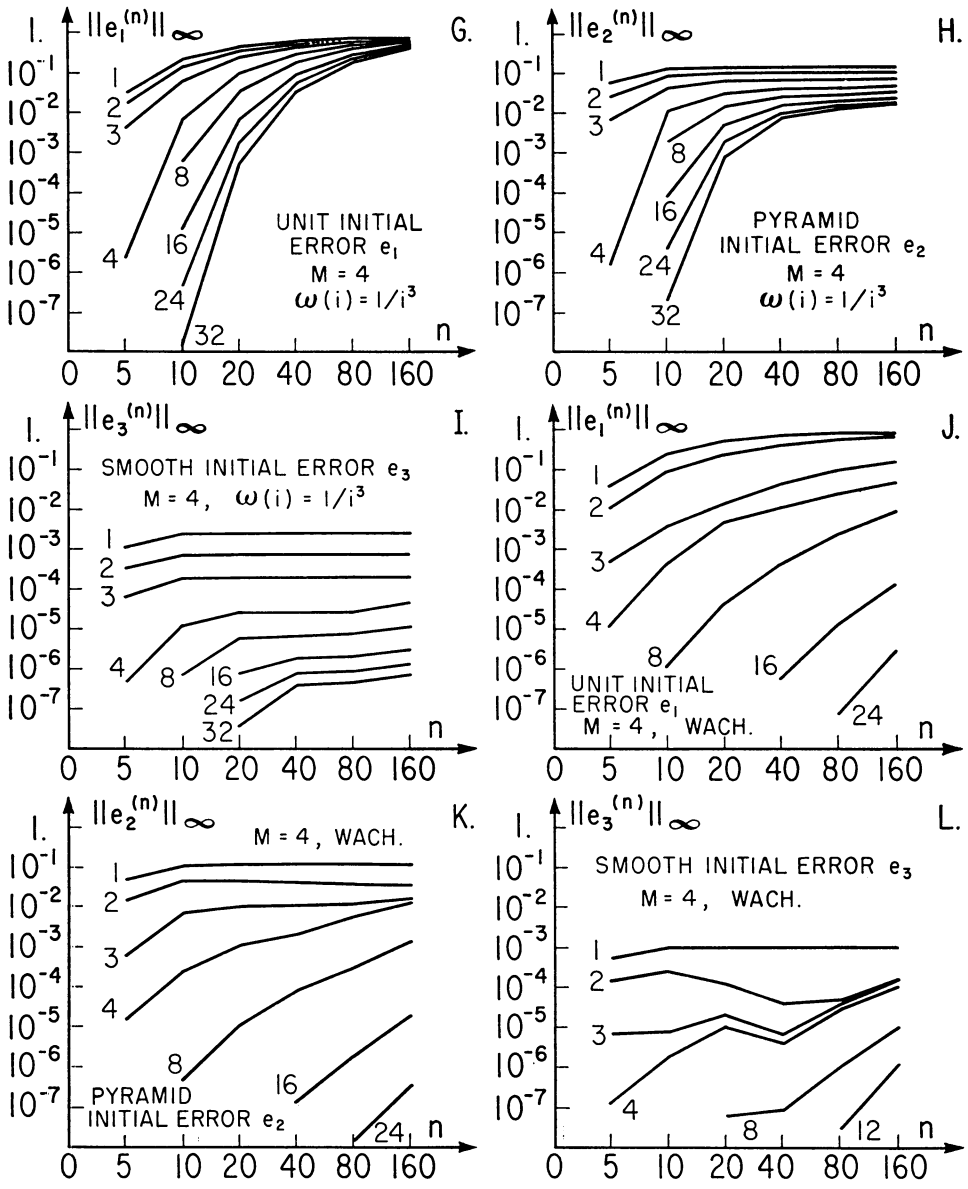
Figs. 8A through 8L show the effect of the mesh size on the number of iteration K_ϵ to reduce the error below a prescribed value ϵ. Graphs of $||e^{(n)}||_\infty$ versus $N + 1 = 1/h$ for several values of n are shown for various initial errors and optimum parameter sets; in each case $m = 4$, and the parameters were cycled; also, the results for Wachspress parameters are shown.

Eigenvalue uncertainty. In many practical problems the value of $\alpha = \lambda_1/\lambda_N$ is uncertain, and it must be estimated to obtain any of the various sets of parameters. Experiments were made to determine the effect of this uncertainty on the error reduction.

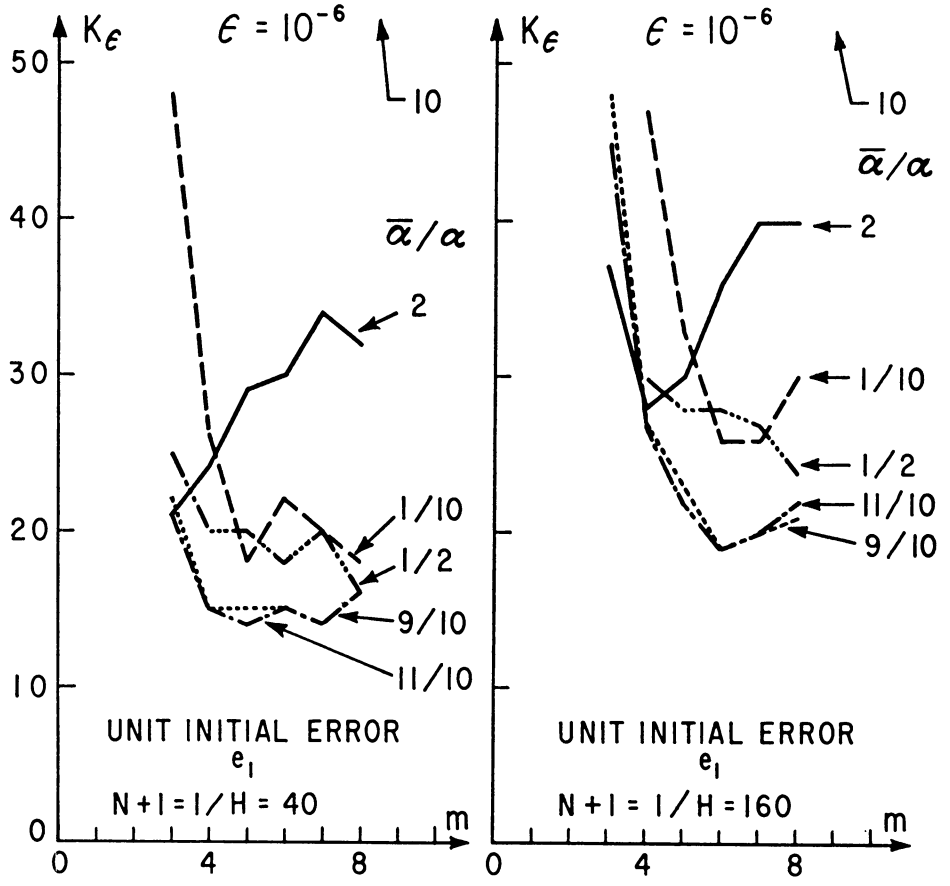
The problem described above was solved with the Wachspress parameters (6.4) on two meshes, $N = 39$ and $N = 159$, with uniform initial error e_1. The parameters were computed with $\bar{\alpha}$ in place of $\alpha = \lambda_1/\lambda_N$ such that $\bar{\alpha}/\alpha$ took on the values 1/10, 1/2, 9/10, 11/10, 2, and 10. The results for $N = 39$ are illustrated in Fig. 9, which shows plots of the number K_ϵ of iterations required to reduce the error to $\epsilon = 10^{-6}$ versus the number m of parameters. The number of iterations exceeded 50 for $\bar{\alpha}/\alpha = 10$, thus these results are not included in Fig. 9.

8. Conclusions. A simple explanation of Guilinger's result [4] follows from the relation between the error function $E_1(P^*, z)$ and the distribution of the Fourier components of the error; graphs of $E_1(P^*, z)$ for different α are similar to those in Fig. 1 but with only one zero. With Guilinger's normalization of $\lambda_1 = 1$, $\lambda_N = 1/\alpha$, the values of the first few eigenvalues $\lambda_1, \cdots, \lambda_k$ are fixed as $N \uparrow \infty$ $(h \downarrow 0)$. With smoothness assumptions such as $\omega(i) = i^q$, $q < -1$, the Fourier coefficients associated with these eigenvalues contribute most to the initial error,





and their values are, essentially, independent of N. Consequently, with a single parameter ρ_1, independent of N, most (i.e. as much as one pleases) of the initial error can be eliminated. With the normalization $\lambda_1 = \alpha$, $\lambda_N = 1$, the value of ρ_1 does depend on N, and it tends to α as $N \uparrow \infty$. Fig. 1, corresponding figures for $m > 2$, and the explanation for $m = 1$, indicate that the same type of result holds for the multiple-parameter case. For this case, with the usual normalization $\lambda_1 = \alpha$, $\lambda_N = 1$, one expects ρ_i/α, $i = 1, \cdots, m$ to be independent of N for sufficiently large N.



It is clear from the experimental results (see Figs. 5A and 5B) that the smoothness assumption does influence the actual error reduction. When *discrete* optimum parameters are used then, at the mth iteration, the smallest error occurs with the smoothness condition which most closely approximates the behavior of the Fourier components of the initial error. However, for (smooth) optimum parameters, this is *not* necessarily the case. For example, with $m = 8$ and initial error e_3 Fig. 5A shows that the smoothness assumption $\omega(i) = 1/i^2$ yields smaller error than $\omega(i) = 1/i^3$; this occurs because of the appreciable difference between the values of the (smooth) optimum parameters ρ_i and the discrete optimum parameters σ_i (see Tables 1 and 3 and Section 6).

Fig. 6 confirms that the effect of the smaller parameters is greater than the effect of the larger ones. The increasing order yields initially faster error reduction because the majority of the error is contained in the low-frequency Fourier components. Because of this, an increasing sequence of parameter is recommended since with it one is more likely to achieve a prescribed error reduction with fewer iterations.

It is clear from Fig. 7 that successive cycles of optimum parameters are much less effective than the first cycle. This is accounted for by the change in distribution of error components (see Section 6 and Fig. 3). Consequently, if m has been selected and if at the mth iteration the error is not sufficiently small, the computation should be continued with a different set of parameters. The effectiveness of the Wachspress parameters is probably due to the fact that the parameter $\rho_1^{(W)} = \alpha$, annihilates the largest error component.

Observe that our experimental results (see Fig. 7), as well as those of Birkhoff, Varga, and Young [2], indicate that the Wachspress parameters are just about as effective as the (smooth) optimum parameter to reduce the error by a prescribed factor. This can be explained by noting (again) that $\rho_1^{(W)} = \alpha$. In both our experiments and those reported in [2], the Fourier coefficient of the initial error associated with λ_1 was larger than any other. This also shows that it is to be expected that Wachspress parameters give larger error reduction than the Peaceman-Rachford parameters $\rho_i^{(P)}$ which have $\rho_i^{(P)} > \alpha$ and an error curve with extremal point at $z = \alpha$, just as do the optimum parameters ρ_i. This is confirmed by the data in [2]. Now, this suggests that the Wachspress parameters should be used in preference to the (smooth) optimum parameters. This is particularly true if there is some uncertainty in the nature of the behavior of the Fourier coefficients or the value of α.

Figs. 8A through 8L confirm the conclusions of Theorem 4 and the Corollary of Section 4 that for sufficiently small h, the number of iterations K_ϵ necessary to reduce the error $||e||_\infty$ below a prescribed value ϵ is independent of h when either the pyramid error e_2 or the smooth error e_3 was used as initial error. Moreover, the Figs. 8A, 8D, 8G, and 8J indicate that this result is false when the uniform initial error e_1 is used which agrees with results in Birkhoff, Varga, and Young [2].

It seems to us, therefore, that except for those problems in which considerable information about the eigenvalues is known, such as the problem in the example of Section 2, there are four reasons for using the Wachspress parameters instead of the (smooth) optimum parameters. First, the error reductions are about the same. Second, they are easier to compute than the optimum parameters. Third, the exact value of α is, in general, not known and one cannot compute the exact optimum parameters even if the appropriate smoothness condition of the initial error were known. Fourth, the number of iterations required to reduce the error to a prescribed value is independent of N provided that smooth initial conditions are used.

Our experiments on the effect of the uncertainty in α indicate that one gets much better error reduction by underestimating α than overestimating it. This is readily explained by examining a typical error curve. For $z > \alpha$ (and hence $\lambda_j > \alpha\lambda_N$) the value of the error function is no larger than its deviation. However, for $z < \alpha$, the error function increases very quickly. Thus if α is overestimated, the error associated with λ_1 is reduced very slowly.

Acknowledgement. Some of the computation was done on the IBM 7094 in the General Motors Research Laboratories, and the rest was done on the CDC 1604 of The University of Texas Computation Center. We thank Mr. James A. Ayers of General Motors for help in the coding of the programs. We thank the referee for remarks which we feel have led to more clarity in the presentation.

Computation Center
The University of Texas
Austin, Texas 78712

Computer Science Department
Purdue University
West Lafayette, Indiana 47907

1. N. I. ACHIESER, *Theory of Approximation*, Ungar, New York, 1956, Section 33. MR **20** #1872.
2. G. BIRKHOFF, R. S. VARGA & D. YOUNG, "Alternating direction implicit methods," *Advances in Computers*, Vol. 3, Academic Press, New York, 1962, pp. 189–273. MR **29** #5395.
3. C. DE BOOR & J. R. RICE, "Chebyshev approximation by $\alpha\Pi((x - r_i)/(x + s_i))$ and application to ADI iteration," *J. Soc. Indust. Appl. Math.*, v. 11, 1963, pp. 159–169. MR **28** #4666.
4. W. H. GUILINGER, JR., "Peaceman-Rachford method for small mesh increments," *J. Math. Anal. Appl.*, v. 11, 1965, pp. 261–277. MR **32** #607.
5. R. E. LYNCH, J. R. RICE & D. H. THOMAS, "Tensor product analysis of partial difference equations," *Bull. Amer. Math. Soc.*, v. 70, 1964, pp. 378–384. MR **29** #6640.
6. R. E. LYNCH, J. R. RICE & D. H. THOMAS, "Direct solution of partial difference equations by tensor product methods," *Numer. Math.*, v. 6, 1964, pp. 185–199.
7. R. E. LYNCH, J. R. RICE & D. H. THOMAS, "Tensor product analysis of alternating direction implicit methods," *J. Soc. Indust. Appl. Math.*, v. 13, 1965, pp. 995–1006. MR **33** #6854.
8. D. W. PEACEMAN & H. H. RACHFORD, JR., "The numerical solution of parabolic and elliptic differential equations," *J. Soc. Indust. Appl. Math.*, v. 3, 1955, pp. 28–41. MR **17**, 196.
9. J. R. RICE, *The Approximation of Functions*, Vol. II, Addison-Wesley, Reading, Mass. (To appear.)
10. D. JACKSON, *The Theory of Approximation*, Amer. Math. Soc. Colloquium Publ., Vol. 11, Amer. Math. Soc., Providence, R. I., 1930.
11. ANTONI ZYGMUND, *Trigonometric Series*, Dover, New York, 1955. MR **17**, 361.
12. C. DE BOOR, "Bicubic spline interpolation," *J. Math. Phys.*, v. 41, 1962, pp. 212–218. MR **28** #1735.