MATHEMATICS OF COMPUTATION, VOLUME 32, NUMBER 141
JANUARY 1978, PAGES 111–126

Computational Experiments and Techniques for the Penalty Method With Extrapolation

By J. Thomas King and Steven M. Serbin

Abstract. In this note we present results of a computational investigation of the extrapolated penalty method for approximate solution of elliptic boundary value problems. We investigate the effects of extrapolation and present an iterative technique for solving the extra linear algebraic systems necessary to perform the process. We indicate how convergence of the iterative procedure may be accelerated when boundary weights are appropriately selected. We consider the Euclidean relative error in the iterative procedure and the effect of conditioning. We develop a bound for the difference between an extrapolate obtained assuming exact solution of all linear systems and the corresponding quantity computed by a terminated iterative procedure.

1. Introduction. Several recent papers have expounded the use of an extrapolation procedure to improve the performance of an underlying variational technique so as to produce quasioptimal error behavior. One of the authors [10] originally applied the idea to the penalty method approximate solution of elliptic boundary value problems and also treated elliptic interface problems [11]. The authors [12] used a similar technique to obtain improved estimates of boundary flux in elliptic problems. A common feature of these extrapolation procedures is that in order to accomplish the technique, it is required to solve several systems of linear equations of the form

$$(A + \gamma B)x = F + \gamma G, \tag{1.1}$$

where γ is a real, positive parameter which varies from one system to the next, while A and B are symmetric $N \times N$ matrices and F and G are $N \times 1$ column vectors which remain fixed for all systems.

An important question which must be addressed is whether or not the expense of solving additional systems (1.1) can warrant the improvements gained in performing the extrapolation technique. In this paper we contend that by employing a simple iterative procedure, after one system has been solved by a direct technique, the solution of subsequent systems with appropriately chosen parameters is asymptotically virtually costless. Moreover, the numerical results we present for a model problem show that the use of the iterative method costs us nothing in the accuracy we obtain when compared to direct solution of each system followed by extrapolation; in doing so, we also exhibit heretofore unpublished computational results which correspond to error estimates in [10].

In Section 2, we briefly describe the penalty method with extrapolation for a simple model problem and present some error estimates which have motivated our study. We describe a model problem and exhibit several sets of experimental results

Received June 14, 1976; revised March 28, 1977.
AMS (MOS) subject classsifications (1970). Primary 65N30, 65F10.
*This work was supported by NSF Grant MPS 76-24453.

obtained via direct solution of each required system, comparing these with theoretical estimates.

In Section 3, we consider some alternatives for the solution of additional systems and then fix on a particularly simple iterative scheme and a method of accelerating its convergence; proofs of convergence and estimation of rate are furnished. We estimate the relative error of the iterative scheme in the Euclidean norm. We discuss the work involved per iteration and compare it to a standard direct approach.

Finally, in Section 4, we present supporting numerical evidence for our iterative scheme via comparison with direct solution and discuss a concurrent implementation of the iterative solutions with the extrapolation procedure, with relevant error estimates.

2. The Penalty Method with Extrapolation; Computational Study. The penalty method of Aubin [1] and Babuška [2] is a finite-element technique for the approximate solution of elliptic boundary value problems. Since the aim here is to discuss the computational aspect of the problem and not the general theory, we restrict our attention to the model problem

$$\Delta u = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega, \tag{2.1}$$

where Δ is the Laplace operator, Ω is some bounded open subset of $\mathbf{R}^2$ with boundary $\partial\Omega$ (it is often assumed that $\partial\Omega \in C^\infty$ for theoretical investigations, but we shall deal computationally with only a piecewise smooth boundary.)

Following King [10], we denote the L_2 inner products on Ω and $\partial\Omega$, respectively, by

$$(u, v) = \int_\Omega uv\,dx \quad \text{and} \quad \langle u, v\rangle = \int_{\partial\Omega} uv\,ds$$

and let $H^s(\Omega)$ ($s \geqslant 0$) be the usual Sobolev space of order s on Ω with norm $\|\cdot\|_s$ (cf. [13]).

The penalty method approximate solution to (2.1) is an element of a certain finite-dimensional subspace $V_h^r \subset H^1(\Omega)$ which satisfies approximability assumptions discussed by King [10] and others. For example, V_h^r may be tensor products of spline functions of order $r-1$ constructed on a mesh of width h imposed on Ω. An important consideration is that the elements of V_h^r need not satisfy the boundary conditions of the problem.

The approximation is obtained by selecting $v = v(\gamma) \in V_h^r$ such that

$$D(v, \phi) + \gamma h^{-\sigma}\langle v, \phi\rangle = -(f, \phi) + \gamma h^{-\sigma}\langle g, \phi\rangle \quad \text{for all } \phi \in V_h^r. \tag{2.2}$$

Here $D(\cdot, \cdot)$ denotes the Dirichlet integral, $\sigma \geqslant 1$, and γ is a positive constant.

If $\{\phi_i\}_{i=1}^N$ is a basis for V_h^r, then $v = \sum_{j=1}^N x_j\phi_j$ is determined by requiring that (2.2) hold for $\phi = \phi_i$, $i = 1, \ldots, N$:

$$\begin{aligned} &D\left(\sum_{j=1}^N x_j\phi_j, \phi_i\right) + \gamma h^{-\sigma}\left\langle \sum_{j=1}^N x_j\phi_j, \phi_i\right\rangle \\ &\quad = \sum_{j=1}^N \{D(\phi_j, \phi_i) + \gamma h^{-\sigma}\langle \phi_j, \phi_i\rangle\}\, x_j \\ &\quad = -(f, \phi_i) + \gamma h^{-\sigma}\langle g, \phi_i\rangle, \qquad i = 1, \ldots, N. \end{aligned} \tag{2.3}$$

Define matrices $A = [a_{ij}]$, $B = [b_{ij}]$, and column vectors $F = [f_i]$, $G = [g_i]$ by

$$a_{ij} = D(\phi_j, \phi_i), \quad b_{ij} = h^{-\sigma}\langle \phi_j, \phi_i \rangle,$$
$$f_i = -(f, \phi_i), \quad g_i = h^{-\sigma}\langle g, \phi_i \rangle. \tag{2.4}$$

Then (2.3) can be written

$$(A + \gamma B)x = F + \gamma G. \tag{2.5}$$

$A + \gamma B$ is symmetric and positive definite (thus nonsingular) for any $\gamma > 0$.

The impetus for the introduction of an extrapolation procedure comes from the asymptotic error estimate [10]:

If $u \in H^s(\Omega)$, $2 \leqslant s \leqslant r$, and $k \leqslant [s-2]$, then there exist functions $w_1, \ldots, w_k$ (independent of h and γ) so that

$$\left\| u - v(\gamma) - \sum_{j=1}^{k} (\gamma^{-1}h^{\sigma})^j w_j \right\|_1 \leqslant C\gamma^{1/2}h^{\mu}\|u\|_s, \tag{2.6}$$

where C is independent of γ and h, and $\mu = \min\{s-1, k+1\}$.

Thus if $\sigma = 1$, $s = r$, and $e(\gamma) = u - v(\gamma)$, then in $H^1(\Omega)$

$$e(\gamma) = \sum_{j=1}^{k} (\gamma^{-1}h)^j w_j + O(\gamma^{1/2}h^{k+1}). \tag{2.7}$$

If $\gamma_1, \ldots, \gamma_{k+1}$ are distinct values of the parameter γ, we may determine constants $a_i^{(k)}$, $1 \leqslant i \leqslant k+1$, from the Vandermonde system

$$\sum_{i=1}^{k+1} a_i^{(k)} = 1, \quad \sum_{i=1}^{k+1} a_i^{(k)}\gamma_i^{-j} = 0, \qquad 1 \leqslant j \leqslant k, \tag{2.8}$$

define

$$u_h^{(k)} \equiv u_h^{(k)}(\gamma_1, \ldots, \gamma_{k+1}) = \sum_{i=1}^{k+1} a_i^{(k)} v(\gamma_i),$$

and it follows from (2.6) that in $H^1(\Omega)$

$$u - u_h^{(k)} = O(h^{k+1}).$$

Thus, for fixed h, we determine a collection of solutions $v(\gamma_1), \ldots, v(\gamma_{k+1})$ and the kth extrapolate, $u_h^{(k)}$, is an appropriately chosen linear combination of the $v(\gamma_j)$.

From a practical point of view, in the determination of $u_h^{(1)}$, we require that the term $\gamma^{-1}hw_1$ be the dominant term in the error expansion (2.7). This requires that γ be sufficiently large. However, we must not choose γ too large as the term $O(\gamma^{1/2}h^{k+1})$ would then dominate the error and the extrapolation process would yield no improvement in accuracy. Clearly an appropriate choice of γ depends on the value of h.

We note that for typical choices of V_h^r we have $r \leqslant 5$, and thus we could only expect improvement in accuracy in the extrapolates $u_h^{(k)}$, $k \leqslant r - 2 \leqslant 3$.

TABLE I: ($\sigma = 1$) *Extrapolation via direct solution*

A. No Extrapolation

γ	8		16		32		64	
h_j	$e_0(h_j)$	ρ_{j0}	$e_0(h_j)$	ρ_{j0}	$e_0(h_j)$	ρ_{j0}	$e_0(h_j)$	ρ_{j0}
1/4	.801E-1		.409E-1		.206E-1		.104E-1	
1/6	.541E-1	.97	.274E-1	.99	.138E-1	.99	.693E-1	1.00
1/8	.408E-1	.98	.206E-1	.99	.104E-1	.98	.520E-2	1.00
1/10	.328E-1	.98	.165E-1	.99	.831E-2	1.00	.416E-2	1.00

B. One Extrapolation

C. Two Extrapolations

	γ_1=8, γ_2=16		γ_1=32, γ_2=64		γ_1=8,γ_2=16,γ_3=32		γ_1=16,γ_2=32,γ_3=64	
h_j	$e_1(h_j)$	ρ_{j1}	$e_1(h_j)$	ρ_{j1}	$e_2(h_j)$	ρ_{j2}	$e_2(h_j)$	ρ_{j2}
1/4	.239E-2		.190E-3		.161E-3		.318E-4	
1/6	.116E-2	1.78	.900E-4	1.84	.714E-4	2.00	.130E-4	2.21
1/8	.695E-3	1.78	.527E-4	1.86	.402E-4	2.00	.723E-5	2.04
1/10	.464E-3	1.81	.348E-4	1.86	.257E-4	2.00	.461E-5	2.01

TABLE II: ($\sigma = 1$) *Extrapolation via direct solution*

A. No Extrapolation

γ	10		100		1000	
h_j	$e_0(h_j)$	ρ_{j0}	$e_0(h_j)$	ρ_{j0}	$e_0(h_j)$	ρ_{j0}
1/4	.646E-1		.666E-2		.668E-3	
1/6	.435E-1	.97	.444E-2	1.00	.445E-3	1.00
1/8	.328E-1	.98	.333E-2	1.00	.334E-3	1.00
1/10	.263E-1	.99	.267E-2	1.00	.267E-3	1.00

B. One Extrapolation

C. Two Extrapolations

	γ_1 = 10 γ_2 = 100		γ_1 = 10, γ_2 = 100, γ_3 = 1000	
h_j	$e_1(h_j)$	ρ_{j1}	$e_2(h_j)$	ρ_{j2}
1/4	.347E-3		.136E-4	
1/6	.166E-3	1.81	.289E-5	3.82
1/8	.983-4	1.83	.960E-6	3.83
1/10	.652-4	1.84	.416E-6	3.75

TABLE III: ($\sigma = 5/4$) *Extrapolation via direct solution*

Two Extrapolations: $\gamma_1 = 50$, $\gamma_2 = 100$, $\gamma_3 = 200$

h_j	$e_2(h_j)$	ρ_{j2}
1/4	.136E-4	
1/6	.291E-5	3.81
1/8	.981E-6	3.78
1/10	.435E-6	3.63

TABLE IV: ($\sigma = 1$) *Large parameters*

A. No Extrapolation

γ	1.0E + 5	1.0E + 7	1.0E + 8	1.0E + 10	1.0E + 12
h_j	$e_0(h_j)$	$e_0(h_j)$	$e_0(h_j)$	$e_0(h_j)$	$e_0(h_j)$
1/4	.150E-4	.136E-4	.136E-4	.156E-4	.518E-3
1/6	.527E-5	.286E-5	.286E-5	.800E-5	.431E-3
1/8	.346E-5	.931E-6	.933E-6	.803E-5	.414E-3

B. Extrapolation - L_2 Error Behavior

	Extrapolation Level	γ_1=1.0E+8	γ_2=9.0E+7	γ_3=8.0E+7	γ_4=7.0E+7
$h=\frac{1}{4}$	0	.136E-4	.136E-4	.136E-4	.136E-4
	1	.136E-4	.136E-4	.136E-4	
	2	.136E-4	.136E-4		
	3	.143E-4			
$h=\frac{1}{6}$	0	.286E-5	.286E-5	.286E-5	.286E-5
	1	.286E-5	.286E-5	.286E-5	
	2	.300E-5	.300E-5		
	3	.549E-5			
$h=\frac{1}{8}$	0	.934E-6	.934E-6	.934E-6	.934E-6
	1	.936E-6	.941E-6	.937E-6	
	2	.125E-5	.111E-5		
	3	.400E-5			

In [10] the first author considers the case $\gamma_j = 2^{j-1}\gamma$ with γ fixed. Then we may give explicit formulae for the extrapolates

$$u_h^{(1)}(\gamma_1, \gamma_2) = \frac{2v(\gamma_2) - v(\gamma_1)}{2 - 1},$$

$$u_h^{(j)}(\gamma_1, \ldots, \gamma_{j+1}) = \frac{2^j u_h^{(j-1)}(\gamma_2, \ldots, \gamma_{j+1}) - u_h^{(j-1)}(\gamma_1, \ldots, \gamma_j)}{2^j - 1}. \tag{2.9}$$

The following L_2 error estimates may be obtained from Corollary 3.1 of [10]:

$$\|u - u_h^{(k)}\|_0 \leqslant C\gamma h^{k+1}\|u\|_{r+1}, \qquad 0 \leqslant k \leqslant r-1,\ \sigma = 1, \tag{2.10}$$

$$\|u - u_h^{(2)}\|_0 \leqslant C\gamma h^{15/4}\|u\|_4, \qquad r = 4,\ \sigma = 5/4. \tag{2.11}$$

We note that the estimate (2.10) for $k = r - 1$ is the optimal order in h provided $u \in H^{r+1}(\Omega)$. From (2.11) we nearly obtain the optimal order, 4, with only two extrapolations. For brevity sake, we shall present experimental results for one model problem

$$\begin{aligned} \Delta u &= 2e^{x+y} \quad \text{in } \Omega = [0, 1] \times [0, 1], \\ u &= e^{x+y} \quad \text{on } \partial\Omega. \end{aligned} \tag{2.12}$$

We choose a subspace V_h^4 of bicubic splines on a uniform mesh of size h.

In [15], the second author has treated the same model problem with several other projection techniques; however, results for the penalty method, with or without extrapolation, were not presented therein. We present in Tables I–IV at the end of this section, evidence that the extrapolation procedure, with appropriate γ's, provides significant improvement in the performance of the penalty method.

In Tables I–IV all required systems (2.5) are solved by separate application of band Cholesky decomposition; we are interested only in the performance of extrapolation.

In these and subsequent tables, we present mesh sizes, corresponding L_2 error $e_k(h) = \|u - u_h^{(k)}\|_0$, and the rate of error reduction,

$$\rho_{jk} = \log(e_k(h_j)/e_k(h_{j-1}))/\log(h_j/h_{j-1}).$$

We attempt to estimate the asymptotic rate of convergence ρ_k, assuming $e_k(h) = Ch^{\rho_k}$ as $h \to 0$ via ρ_{jk}, to compare with the estimates (2.10) and (2.11), but caution that the mesh size with which we have computed are far from the asymptotic range.

Tables I and II for $\sigma = 1$ produce some notable trends and yet some anomalies. Clearly, without extrapolation the convergence is first order for any given choice of γ. The error decreases in direct proportion to increasing γ for fixed h. Moreover, each successive application of extrapolation significantly improves the error when compared to behavior shown for fewer extrapolations involving the same weights. The first extrapolate is seen to produce nearly second order accuracy. However, Tables I and II show contrasting behavior for the second extrapolate. When the weights are small, i.e.

Table I, the reduction is only second order, while in Table II we witness reductions which are greater than third order. In neither case do we see predicted third order convergence; again we emphasize that the error bounds for extrapolation are only asymptotically correct.

Results are presented in Table III for $\sigma = 5/4$ which nearly concur with the accuracy predicted by (2.11).

Lest our experimental results lead to any unwarranted general conclusions we include the following remarks:

(i) Our extrapolation procedure may be thought of, in some sense, as extrapolation to the limit as $\gamma \to \infty$. Our experiments (Tables IA, IIA) indicate that the error decreases with increasing γ. This suggests that one could just use large boundary weights rather than use extrapolation. However, the effect of this approach is to force the boundary conditions to be more nearly satisfied at the expense of satisfying the equation in the interior. In our model problem, the use of a tensor product basis whose elements coincide with the boundary allows a freedom which is not present in more general situations. Approximation of the boundary conditions becomes a one-dimensional problem; each of the boundary elements still has the freedom in the other dimension to admit approximation of the equation in the interior near the boundary. For other elements in our model problem or for other regions using elements which do not coincide with the boundary we would lose approximability near the boundary.

(ii) In fact, by taking very large boundary weights the error begins to increase (see Table IVA). The asymptotic error formula (2.10) contains a term of order γh^4 for cubic splines. For fixed h and very large γ this term becomes the dominant one in the error expansion. Not surprisingly, this is manifested in rates of error reduction near 4, as h decreases, until γ gets so large that the error is being controlled by this factor of γ. The error increases as γ is further increased, extrapolation is no longer warranted and in fact results in loss of accuracy (see Table IV).

(iii) We thus have a dilemma: (a) how large should the boundary weight, γ, be? (b) should we extrapolate at all? We are concerned here with the implementation and performance of extrapolation; our only concern is that the leading terms in the error expansion, at each stage of the extrapolation process, are in fact dominant. We contend that, rather than having to make a decision on how large to force γ before the term γh^4 begins to dominate, extrapolation with moderately large boundary weights will produce the same net effect. The determination of suitable values of γ may easily be tested computationally by a standard device given in [17, p. 313].

3. Iterative Solution. Let us first mention that several direct approaches for solving the systems (2.5) have been investigated. These include the method of modification [7], modification of the LDL^t decomposition [6], partitioning [3], and a Lanczos-type procedure [8]. None of these seems to compete with the reaccumulation of $A + \gamma B$ and subsequent triangular band factorization. We have compared these techniques, as well as the later-described iterative method, with the band Cholesky decomposition of $A + \gamma B$ obtained after the usual row-by-row ordering of elements. We must remark that we do not account for any sparsity within the band; for solution

of a single problem on a rectangle, a technique like that of George's nested dissection [5] may be appropriate.

Our aim here is not a general comparison of iterative vs direct techniques in the finite-element procedure. Since $A + \gamma B$ is positive definite, we know that SOR will converge. Our approach is to assume that (2.5) has been solved for $\gamma = \gamma_1$ by a direct method involving Cholesky decomposition of $A + \gamma_1 B$ and to use this in solving (2.5) for $j = 2, \ldots, k$. To this end, we write (2.5) in the form

$$(A + \gamma_1 B)x = F + \gamma_1 G, \tag{3.1}$$

$$(A + \gamma_j B)y = F + \gamma_j G. \tag{3.2}$$

Put

$$M = A + \gamma_1 B, \quad u = y - x, \quad \delta = \gamma_j - \gamma_1, \quad K = -\delta B, \tag{3.3}$$

then (3.2) becomes

$$(M - K)(u + x) = (F + \gamma_1 G) + \delta G.$$

Expanding and using (3.1),

$$Mu = Ku + Kx + \delta G. \tag{3.4}$$

This suggests the iterative scheme

$$Mu^{(n+1)} = Ku^{(n)} + Kx + \delta G, \quad u^{(0)} \text{ arbitrary}. \tag{3.5}$$

We could proceed to analyze the spectral radius $\rho(C)$ of the iteration matrix $C = M^{-1}K$. However, let us go on to suggest a technique for accelerating the convergence and handle the above as a special case. Using a device suggested in Isaacson and Keller [9], we let $0 \leqslant \theta < 1$ be a parameter, and rewrite (3.4) as

$$(1 - \theta)Mu = (K - \theta M)u + (Kx + \delta G), \tag{3.6}$$

which induces the iterative scheme

$$(1 - \theta)Mu^{(n+1)} = (K - \theta M)u^{(n)} + Kx + \delta G, \tag{3.7}$$

so

$$Mu^{(n+1)} = \frac{1}{1 - \theta}\left[(K - \theta M)u^{(n)} + Kx + \delta G\right]. \tag{3.8}$$

This defines an iteration scheme with matrix

$$C_\theta = \frac{1}{1 - \theta} M^{-1}(K - \theta M) = \frac{1}{1 - \theta}(C - \theta I).$$

Observe that $C_0 = C$. Now, from (3.4), and (3.8) we have, with $e^{(n)} = u - u^{(n)}$

$$Me^{(n+1)} = \frac{1}{1 - \theta}[k - \theta M]e^{(n)} \quad \text{or} \quad e^{(n+1)} = C_\theta e^{(n)}. \tag{3.9}$$

Then, in order to establish convergence, we must determine conditions under which $\rho(C_\theta) < 1$ (cf. [16]). Moreover, the rate of convergence is essentially determined by $\rho(C_\theta)$, so we seek to minimize bounds on this quantity in terms of $\nu = |\delta|/\gamma_1$ by appropriate choice of θ.

THEOREM 3.1. *Let A, B be real, symmetric positive semidefinite matrices and $A + \gamma B$ be positive definite for any $\gamma > 0$. Let δ, C_θ, and ν be defined as above and $0 \leqslant \theta < 1/2$.*

(i) *If $\delta < 0$ and $\nu < 1$, then $\rho(C_\theta) < 1$ and the bound on $\rho(C_\theta)$ is minimized by choosing $\theta = \hat{\theta} = \nu/2$, with $\rho(C_{\hat{\theta}}) \leqslant \nu/(2 - \nu)$.*

(ii) *If $\delta > 0$ and $\nu < 1$ and $\theta < (1 - \nu)/2$, then $\rho(C_\theta) < 1$ and the optimal choice is $\theta = \tilde{\theta} = 0$ with $\rho(C) \leqslant \nu$.*

Proof. Let λ be any eigenvalue of C_θ. Then, for some $v \neq 0$, $C_\theta v = \lambda v$, so

$$\frac{1}{1-\theta}[C - \theta I]v = \lambda v \Rightarrow [C - \theta I]v = \lambda(1 - \theta)v,$$

or

$$Cv = [\theta + \lambda(1 - \theta)]v,$$

so by (3.3)

$$-\delta Bv = [\theta + \lambda(1 - \theta)](A + \gamma_1 B)v. \tag{3.10}$$

Denote the conjugate transpose of v by v^H. Then

$$-\delta v^H B v = [\theta + \lambda(1 - \theta)]\,[v^H A v + \gamma_1 v^H B v]. \tag{3.11}$$

Define $z = v^H B v$, $w = v^H A v$; z and w are nonnegative since A and B are positive semidefinite. (3.10) becomes

$$-\delta z = [\theta + \lambda(1 - \theta)]\,[w + \gamma_1 z], \tag{3.12}$$

or

$$\theta + \lambda(1 - \theta) = -\delta z/(w + \gamma_1 z). \tag{3.13}$$

The denominator is nonzero since $A + \gamma_1 B$ is positive definite. Now, consider two cases.

Case 1. If $\delta < 0$, (3.13) implies $0 \leqslant \theta + \lambda(1 - \theta) \leqslant -\delta/\gamma_1 = \nu$, so

$$-\frac{\theta}{1-\theta} \leqslant \lambda \leqslant \frac{\nu - \theta}{1 - \theta}. \tag{3.14}$$

Since $\theta < 1/2$, $-\theta/(1 - \theta) > -1$ and since $\nu < 1$, $(\nu - \theta)/(1 - \theta) < 1$, hence $|\lambda| < 1$. Moreover, if $\theta \leqslant \nu$, we minimize the bound in (3.14) by setting $(\nu - \theta)/(1 - \theta) = \theta/(1 - \theta)$, or $\theta = \nu/2$, in which case

$$|\lambda| \leqslant \nu/(2 - \nu). \tag{3.15}$$

On the other hand, if we would allow $\theta \geqslant \nu$, then $|\lambda| \leqslant \theta/(1 - \theta)$, and since $q(\theta) = \theta/(1 - \theta)$ is monotone increasing, the best bound would occur at $\theta = \nu$, giving $|\lambda| \leqslant \nu/(1 - \nu)$, which compares unfavorably with (3.15).

Case 2. If $\delta > 0$, (3.13) yields

$$-\nu \leqslant \theta + \lambda(1 - \theta) \leqslant 0, \tag{3.16}$$

or

$$\frac{-\nu - \theta}{1-\theta} \leqslant \lambda \leqslant \frac{-\theta}{1-\theta} \leqslant 0. \tag{3.17}$$

If $\theta < (1 - \nu)/2$ and $\nu < 1$, $(-\nu - \theta)/(1 - \theta) > 1$, so $|\lambda| \leqslant (\nu + \theta)/(1 - \theta) < 1$. But, since $r(\theta) = (\nu + \theta)/(1 - \theta)$ is monotonically increasing on $[0, 1]$, the best bound is $|\lambda| \leqslant \nu$ obtained at $\theta = 0$.

From this theorem, we may conclude that the iterative method (3.5) ($\theta = 0$) will converge if $\nu < 1$ and that $\rho(C) \leqslant \nu$. Moreover, if we wish to accelerate the convergence, our best bet is to select $\delta < 0$, $\theta = \nu/2$. This has the advantage of making γ_1 the largest parameter, which allows ν to be made smaller. In fact if we wish to solve a sequence of problems of the form (2.5), then we should order the parameters as $\gamma_1 > \gamma_2 > \ldots > \gamma_k$ with $\gamma_1 - \gamma_k$ small so that the spectral radii of the iteration matrices can be made as small as desired. We note that choice of $\theta < 0$ or $\theta > 1$ has no advantage; moreover for $\theta = \nu/2$ we obtain $\rho(C_\theta) = (\gamma_1 - \gamma_j)/(\gamma_1 + \gamma_j)$.

We can estimate the relative error in the nth iterate $u^{(n)}$ in (3.8) in the Euclidean vector norm $\|\cdot\|_E$ and in doing so investigate the role of the condition number in the iterative process. We recall the following result on simultaneous diagonalization (cf. [4]).

LEMMA 3.1. *Let M and K be symmetric, with M positive definite. Let $0 \leqslant \theta < 1$, so that $(1 - \theta)M$ is positive definite and $K - \theta M$ is symmetric. Then there exists P such that*

$$(1 - \theta)P^H M P = I, \tag{3.18}$$

$$P^H(K - \theta M)P = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_N), \tag{3.19}$$

$$P = VDU, \tag{3.20}$$

where V and U are unitary,

$$D = \operatorname{diag}\left(\frac{1}{\sqrt{\mu_1}}, \ldots, \frac{1}{\sqrt{\mu_N}}\right),$$

$\{\mu_j\}_{j=1}^N$ are the eigenvalues of $(1 - \theta)M$, and $\mu_1 \geqslant \mu_2 \geqslant \ldots \geqslant \mu_N > 0$.

With these results we may give an explicit characterization of the spectral radius of C_θ as a norm. Define a norm on $\mathbf{R}^N$ by $N(\phi) = \|P^{-1}\phi\|_E$ and the subordinate matrix norm by

$$N(C_\theta) = \max_{\phi \neq 0} \frac{N(C_\theta \phi)}{N(\phi)} = \max_{\phi \neq 0} \frac{\|P^{-1} C_\theta \phi\|_E}{\|P^{-1}\phi\|_E}.$$

But from (3.18) and (3.19), we find

$$P^{-1} C_\theta P = \Lambda. \tag{3.21}$$

Thus, $\{\lambda_j\}_{j=1}^N$ are the eigenvalues of C_θ. Letting $\psi = P^{-1}\phi$, it follows that

$$N(C_\theta) = \max_{\psi \neq 0} \frac{\|P^{-1} C_\theta P \psi\|_E}{\|\psi\|_E} = \max_{1 \leqslant j \leqslant N} \{|\lambda_j|\} = \rho(C_\theta).$$

Hence, letting $u^{(0)} = -x$, (3.9) yields

$$\frac{N(e^{(n)})}{N(y)} \leqslant [\rho(C_\theta)]^n. \tag{3.22}$$

Now $\|e^{(n)}\|_E = \|PP^{-1}e^{(n)}\|_E \leqslant \|P\|_E N(e^{(n)})$ so that

$$\frac{\|e^{(n)}\|_E}{\|P\|_E\|P^{-1}\|_E\|y\|_E} \leqslant \frac{N(e^{(n)})}{\|P^{-1}\|_E\|y\|_E}. \tag{3.23}$$

But $N(y) \leqslant \|P^{-1}\|_E\|y\|_E$ so we compose (3.22) and (3.23) to yield

$$\frac{\|e^{(n)}\|_E}{\|y\|_E} \leqslant \|P\|_E\|P^{-1}\|_E\frac{N(e^{(n)})}{N(y)} \leqslant \|P\|_E\|P^{-1}\|_E\rho(C_\theta)^n.$$

Referring back to (3.20),

$$\|P\|_E \leqslant \|D\|_E = \sqrt{(1-\theta)^{-1}\|M^{-1}\|_E} \quad \text{and} \quad \|P^{-1}\|_E \leqslant \|D^{-1}\|_E = \sqrt{(1-\theta)\|M\|_E}.$$

Recalling that the spectral condition number of M is $\kappa(M) = \|M\|_E\|M^{-1}\|_E$ we obtain finally the Euclidean relative error bound

$$\|e^{(n)}\|_E/\|y\|_E \leqslant \sqrt{\kappa(M)}\rho(C_\theta)^n. \tag{3.24}$$

Now, King [10] has shown that $\kappa(M) = O(h^{-1-\sigma})$; thus from (3.24) we note in the particular case $\sigma = 1$, the effect of conditioning in the iterative scheme appears only as $O(h^{-1})$.

As mentioned above, we have compared the work involved per iterative step (3.8) with the computational effort expended in a band Cholesky decomposition of the matrices in (3.2), since this would be the dominant part of the cost in solving the additional problems. Let us suppose, for example, that Ω is the unit square with mesh size $h = 1/H$ and V_h^r consists of tensor products of splines of order r, then the matrix K has $(H + r)^2 = N^2$ rows and columns, but only $4Hr$ of these are nonzero. Then, assuming we have kept $Mu^{(n)}$, determination of the right side of (3.8) will cost about $(4Hr)^2$ multiplications, which is negligible in comparison to the back substitution work. With the usual ordering, the band width is $r(H + 1)$, so using the estimates of Martin and Wilkinson [14], each back solve costs $2N^2r(H + 1) \sim 2rN^3$ multiplications. In comparison, to perform Cholesky band factorization requires about $1/2N^2r^2(H + 1)^2 \sim 1/2N^4r^2$ multiplications. Hence it would take roughly $N/4$ iterations before the work would be comparable, and we shall see that we achieve convergence to workable accuracy in far fewer iterations. In fact the results of Theorem 3.1 do not depend upon N. We should note that a similar comparison should be carried out if other techniques of direct solution were to be contemplated.

4. Studies of Iterative Technique. In order to support our contention that the iterative method (3.8) yields an economical, effective method for producing the penalty extrapolates, we examine here the effect of the parameter θ for acceleration and some experimental results for the model problem obtained via these iterations. In Table V, we examine the relative l_1 error in the solution vector $u^{(n+1)}$ of (3.8) for $\theta = 0$ and $\theta = \hat{\theta} = \nu/2$. In this example, $\gamma_1 = 100$, $\gamma_2 = 80$, so $\nu = .2$, $\hat{\theta} = .1$. We see that Theorem 3.1 is substantiated; when $\theta = 0$, the error is reduced at each step by about $.195 \approx \nu$, while when $\theta = .1$, the rate is about .106, whereas the bound is $\nu/(2 - \nu) = 1/9$.

TABLE V. *Comparison of iteration parameters*

Iteration Number	$\theta = .1$ ℓ_1 Relative Error	$\theta = 0$ ℓ_1 Relative Error
1	.101E-0	.190E-0
2	.102E-1	.363E-1
3	.107E-2	.695E-2
4	.113E-3	.134E-2
5	.120E-4	.257E-3
6	.127E-5	.501E-4
7	.135E-6	.977E-5
8	.145E-7	.191E-5

Now, we present in Table VI a representative sample of extrapolation with the accelerated iterative solution of linear systems. Having iterated until the l_1 relative error is less than 10^{-7}, these results are identical to those produced via direct solution. We note that even if we only iterate until the error is less than 10^{-2}, the results do not deviate in the most significant digits from those produced via direct solution. The L_2 error behaves much as expected for $k = 0$, 1, and 3 extrapolates, and as in Section 2, Table II, the 2nd extrapolate shows reductions at a rate greater than the asymptotically predicted rate of 3. For this parameter range, each extrapolation produces improved results. Starting with $\gamma_1 = 100$, we required 6 iterations to solve for $\gamma_2 = 90$, 8 iterations for $\gamma_3 = 80$, and 10 iterations for $\gamma_4 = 70$. We have computed with weights even closer together than these, thus requiring even fewer iterations, and noted similar behavior.

Suppose we want to compute the kth extrapolate based on $\gamma_1, \gamma_2, \ldots, \gamma_{k+1}$, i.e. $u_h^{(k)} \equiv u_h^{(k)}(\gamma_1, \ldots, \gamma_{k+1})$. We order the parameters so that $\gamma_1 > \gamma_2 > \ldots > \gamma_{k+1} > 1$ with $\gamma_1 - \gamma_{k+1}$ small. If we use the iteration (3.5) for the computations, then in place of $u_h^{(k)}$, we find

$$\tilde{u}_h^{(k)} = \sum_{i=1}^{k+1} a_i v_{n_i}(\gamma_i).$$

Here $v_{n_1}(\gamma_1) \equiv v(\gamma_1)$ is obtained by solving (2.2) and for $i = 2, \ldots, k+1$, $v_{n_i}(\gamma_i)$ is the solution of (2.2) with $\gamma = \gamma_i$ which is obtained from the iterative scheme (3.5) upon termination with iteration n_i.

Our aim here is to obtain estimates for $\|u_h^{(k)} - \tilde{u}_h^{(k)}\|$ in some norm $\|\cdot\|$. Toward this end we note that (3.2) may be written as

$$\begin{aligned} &D(v(\gamma_i) - v(\gamma_1), \phi) + \gamma_1 h^{-1}\langle v(\gamma_i) - v(\gamma_1), \phi\rangle \\ &\qquad = (\gamma_1 - \gamma_i)h^{-1}\{\langle v(\gamma_i) - v(\gamma_1), \phi\rangle + \langle v(\gamma_1) - g, \phi\rangle\} \end{aligned} \tag{4.1}$$

for all $\phi \in V_h^r$. Here $v(\gamma_i)$ is the solution of (2.3) with $\gamma = \gamma_i$ and $\sigma = 1$.

TABLE VI: ($\sigma = 1$) *Three extrapolates, iterative solution*

k^{th} Extrapolate	h_j	$e_k(h_j)$	ρ_{jk}
$k = 0$	1/4	.665E-2	
$\gamma_1 = 100$	1/6	.444E-2	1.00
	1/8	.333E-2	1.00
$k = 1$	1/4	.477E-4	
$\gamma_1 = 100$	1/6	.218E-4	1.93
$\gamma_2 = 90$	1/8	.127E-4	1.89
$k = 2$	1/4	.137E-4	
$\gamma_1 = 100$	1/6	.296E-5	3.77
$\gamma_2 = 90$			
$\gamma_3 = 80$	1/8	.103E-5	3.68
$k = 3$			
$\gamma_1 = 100$	1/4	.136E-4	
$\gamma_2 = 90$	1/6	.286E-5	3.84
$\gamma_3 = 80$			
$\gamma_4 = 70$	1/8	.931E-6	3.91

The iteration (3.5) may be written as

$$\begin{aligned} &D(u_n(\gamma_i), \phi) + \gamma_1 h^{-1}\langle u_n(\gamma_i), \phi\rangle \\ &\qquad = (\gamma_1 - \gamma_i)h^{-1}\{\langle u_{n-1}(\gamma_i), \phi\rangle + \langle g - v(\gamma_1), \phi\rangle\} \end{aligned} \tag{4.2}$$

for all $\phi \in V_h^r$. Here $u_n(\gamma_i) = v_n(\gamma_i) - v(\gamma_1)$ where the iteration is terminated after n_i iterations.

We subtract (4.2), with $n = n_i$, from (4.1) to yield

$$\begin{aligned} &D(v(\gamma_i) - v_{n_i}(\gamma_i), \phi) + \gamma_1 h^{-1}\langle v(\gamma_i) - v_{n_i}(\gamma_i), \phi\rangle \\ &\qquad = (\gamma_1 - \gamma_i)h^{-1}\langle v(\gamma_i) - v_{n_i-1}(\gamma_i), \phi\rangle. \end{aligned} \tag{4.3}$$

Choose $\phi = v(\gamma_i) - v_{n_i}(\gamma_i)$ in (4.3) to yield

$$H_{\gamma_1}^2(v(\gamma_i) - v_{n_i}(\gamma_i)) = (\gamma_1 - \gamma_i)h^{-1}\langle v(\gamma_i) - v_{n_i-1}(\gamma_i), v(\gamma_i) - v_{n_i}(\gamma_i)\rangle, \tag{4.4}$$

where $H_\gamma^2(\phi) = D(\phi, \phi) + \gamma h^{-1}|\phi|_0^2$ and $|\phi|_0^2 = \langle\phi, \phi\rangle$.

From (4.4) and the Schwarz inequality it follows easily that

$$|v(\gamma_i) - v_{n_i}(\gamma_i)|_0 \leqslant \left|\frac{\gamma_1 - \gamma_i}{\gamma_1}\right| |v(\gamma_i) - v_{n_i-1}(\gamma_i)|_0;$$

and hence,

$$H_{\gamma_1}^2(v(\gamma_i) - v_{n_i}(\gamma_i)) \leq \frac{(\gamma_1 - \gamma_i)^2}{\gamma_1} h^{-1} |v(\gamma_i) - v_{n_i-1}(\gamma_i)|_0^2. \tag{4.5}$$

By the triangle inequality we have

$$H_{\gamma_1}(v(\gamma_i) - v_{n_i-1}(\gamma_i)) \leq H_{\gamma_1}(v(\gamma_i) - v_{n_i}(\gamma_i)) + H_{\gamma_1}(v_{n_i}(\gamma_i) - v_{n_i-1}(\gamma_i))$$

so that, by virtue of (4.5), we obtain

$$H_{\gamma_1}(v(\gamma_i) - v_{n_i-1}(\gamma_i)) \leq \frac{(\gamma_1 - \gamma_i)h^{-1}}{\sqrt{\gamma_1}} |v(\gamma_i) - v_{n_i-1}(\gamma_i)|_0 + H_{\gamma_1}(v_{n_i}(\gamma_i) - v_{n_i-1}(\gamma_i)).$$

It follows that

$$|v(\gamma_i) - v_{n_i-1}(\gamma_i)|_0 \leq \frac{\sqrt{\gamma_1}}{\gamma_i} h^{1/2} H_{\gamma_1}(v_{n_i}(\gamma_i) - v_{n_i-1}(\gamma_i)). \tag{4.6}$$

Combining (4.5) and (4.6) yields

$$H_{\gamma_1}(v(\gamma_i) - v_{n_i}(\gamma_i)) \leq \left(\frac{\gamma_1 - \gamma_i}{\gamma_i}\right) H_{\gamma_1}(v_{n_i}(\gamma_i) - v_{n_i-1}(\gamma_i)).$$

It is well known that for all $\phi \in H^1(\Omega)$, $\|\phi\|_1 \leq C\{D(\phi, \phi) + |\phi|_0^2\}$ so that

$$\|v(\gamma_i) - v_{n_i}(\gamma_i)\|_1 \leq C\left(\frac{\gamma_1 - \gamma_i}{\gamma_i}\right) H_{\gamma_1}(v_{n_i}(\gamma_i) - v_{n_i-1}(\gamma_i)).$$

Furthermore, it is known (cf. [10]) that for many subspaces V_h^r, $H_{\gamma_1}(\phi) \leq K\sqrt{\gamma_1}\,\|x\|_E$, where $\phi = \Sigma_{j=1}^N x_j\phi_j$, $x = (x_1, \ldots, x_N)^T$ and K is independent of h and γ_1. Thus

$$\|v(\gamma_i) - v_{n_i}(\gamma_i)\|_1 \leq C\sqrt{\gamma_1}\left(\frac{\gamma_1 - \gamma_i}{\gamma_i}\right)\|x_{n_i} - x_{n_i-1}\|_E,$$

where x_{n_i} is the column vector corresponding to $v_{n_i}(\gamma_i)$.

It follows from the triangle inequality that

$$\|u_h^{(k)} - \tilde{u}_h^{(k)}\|_1 \leq C\sqrt{\gamma_1} \sum_{i=1}^{k+1} |a_i| \left(\frac{\gamma_1 - \gamma_i}{\gamma_i}\right)\|x_{n_i} - x_{n_i-1}\|_E.$$

We have proven the following:

THEOREM 4.1. *Suppose* $\gamma_1 > \gamma_2 > \ldots > \gamma_{k+1} > 1$, $u_h^{(k)} = \Sigma_{i=1}^{k+1} a_i v(\gamma_i)$ *where* $v(\gamma_i)$ *is the solution of* (2.2) *with* $\gamma = \gamma$ *and* $\sigma = 1$, $\tilde{u}_h^{(k)} = \Sigma_{i=1}^{k+1} a_i v_{n_i}(\gamma_i)$ *where* $v_{n_i}(\gamma_i) = u_{n_i}(\gamma_i) + v(\gamma_1)$ *and* $u_{n_i}(\gamma_i)$ *is the* n_i *th iterate of the scheme* (3.5); *then there exists* $C > 0$, *independent of* $\gamma_1, \ldots, \gamma_{k+1}$ *and* h, *such that for* $p = 0, 1$

$$\|u_h^{(k)} - \tilde{u}_h^{(k)}\|_p \leq C\sqrt{\gamma_1} \sum_{i=1}^{k+1} |a_i| \left(\frac{\gamma_1 - \gamma_i}{\gamma_i}\right) \|x_{n_i} - x_{n_i - 1}\|_E,$$

where x_{n_i} *is the vector representation of* $v_{n_i}(\gamma_i)$ *relative to the basis* $\{\phi_j\}_{j=1}^N$ *of* V_h^r.

Moreover, by a similar analysis, it follows that for $p = 0, 1$

$$\|u_h^{(k)} - \tilde{u}_h^{(k)}\|_p \leq C\gamma_1^{3/2} \sum_{i=1}^{k+1} |a_i| \left(\frac{\gamma_1 - \gamma_i}{\gamma_1}\right)^{n_i} \|x_1 - x_0\|_E,$$

where x_0 *is the vector representation of* $v(\gamma_1)$.

We remark that it is not difficult to obtain a bound for $\|u_h^{(k)} - \tilde{u}_h^{(k)}\|_p$ $(p = 0, 1)$ in terms of data. Specifically, it can be shown that

$$\|u_h^{(k)} - \tilde{u}_h^{(k)}\|_p \leq C \sum_{i=1}^{k+1} |a_i| \left(\frac{\gamma_1 - \gamma_i}{\gamma_1}\right)^{n_i} h^{-1/2} \{\gamma_i^{-1} h \|f\|_0^2 + |g|_0^2\}^{1/2}$$

with C independent of h and $\gamma_1, \ldots, \gamma_{k+1}$. Here we have taken $u_0(\gamma_i) = 0$ for $i = 2, \ldots, k + 1$.

Department of Mathematical Science
University of Cincinnati
Cincinnati, Ohio 45221

Department of Mathematics
University of Tennessee
Knoxville, Tennessee 37916

1. J. P. AUBIN, "Approximation des problèmes aux limites non homogènes et regularité de la convergence," *Calcolo*, v. 6, 1969, pp. 117–139.

2. I. BABUŠKA, "The finite element method with penalty," *Math. Comp.*, v. 27, 1973, pp. 221–228.

3. J. BUNCH & D. ROSE, "Partitioning, tearing, and modification of sparse linear systems," *J. Math. Anal. Appl.*, v. 48, 1974, pp. 574–593.

4. J. N. FRANKLIN, *Matrix Theory*, Prentice-Hall, Englewood Cliffs, N. J., 1968.

5. A. GEORGE, "Nested dissection of a regular finite element mesh," *SIAM J. Numer. Anal.*, v. 10, 1973, pp. 345–363.

6. P. E. GILL, G. H. GOLUB, W. MURRAY & M. A. SAUNDERS, "Methods for modifying matrix factorizations," *Math. Comp.*, v. 28, 1974, pp. 505–535.

7. D. GOLDFARB, "Modification methods for inverting matrices and solving systems of linear equations," *Math. Comp.*, v. 26, 1972, pp. 829–852.

8. G. H. GOLUB, R. UNDERWOOD & J. H. WILKINSON, *The Lanczos Algorithm for the Symmetric* $Ax = \lambda Bx$ *Problem*, Stanford Univ. Comput. Sci. Report SU 326P30-16, March 1972.

9. E. ISAACSON & H. B. KELLER, *Analysis of Numerical Methods*, Wiley, New York, 1966.

10. J. T. KING, "New error bounds for the penalty method and extrapolation," *Numer. Math.*, v. 23, 1974, pp. 153–165.

11. J. T. KING, "A quasioptimal finite element method for elliptic interface problems," *Computing*, v. 15, 1975, pp. 127–135.

12. J. T. KING & S. M. SERBIN, "Boundary flux estimates for elliptic problems by the perturbed variational method," *Computing*, v. 16, 1976, pp. 339–347.

13. J. L. LIONS & E. MAGENES, *Problèmes aux Limites non Homogènes et Applications*, Vol. 1, Dunod, Paris, 1968.

14. R. S. MARTIN & J. H. WILKINSON, "Symmetric decomposition of positive definite band matrices," *Numer. Math.*, v. 7, 1965, pp. 355–361.

15. S. M. SERBIN, "Computational investigations of least-squares type methods for the approximate solution of boundary value problems," *Math. Comp.*, v. 29, 1975, pp. 777–793.

16. R. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1962.

17. S. CONTE & C. de BOOR, *Elementary Numerical Analysis: An Algorithmic Approach*, McGraw-Hill, New York, 1972.