

Efficient Higher Order Single Step Methods for Parabolic Problems: Part I

By James H. Bramble and Peter H. Sammon*

Abstract. Some efficient, high order methods are discussed for approximating the solution of an initial boundary value problem for a homogeneous parabolic equation with time dependent coefficients. The methods are based on Galerkin-type approximations in the spacial variables and single step methods in the time variable. The equations defining the time-stepping procedure are solved only approximately, however. A preconditioned iterative technique is used for this purpose. The resulting algorithm is shown to produce optimal order approximations using only the order of work required by the single step method applied to the parabolic problem with time independent coefficients.

I. Introduction. In this paper we study efficient ways to calculate approximations to the solution of a parabolic equation that are of third or fourth order in time and of high order in space. The approximations are generated by rational function based schemes (cf. Nassif and Descloux [7] or Baker, Bramble, and Thomée [2] if the operator in question is time independent), but these schemes are modified in a manner suggested by Douglas, Dupont, and Ewing [4] in their work on the Crank-Nicolson method. We study the schemes in the context of a linear parabolic equation with time dependent coefficients in this part of the work, and we will generalize these schemes to nonlinear equations in Part II of this work.

The schemes suggested by Nassif and Descloux in [7] are single step methods based on a certain class of rational function approximations to the function e^{-x} , $x \in \mathbf{R}$, and a given family of discrete spacial operators. Nassif and Descloux give estimates in [7] that show that the resulting approximations make errors that are of optimal order. However, the schemes are not really suitable for practical computation since each step of the time-stepping procedure involves the solution of a new linear system that is related to the family of spacial operators.

Douglas, Dupont, and Ewing address this problem in [4] and suggest the remedy of using a preconditioned iterative technique to approximately solve the linear system. This approach only requires the solution of linear systems involving a fixed discrete spacial operator if sufficiently many iterations are done at each time step. Conditions are discussed in [4] that also allow one to iterate a fixed number of times at each time

Received July 12, 1979; revised November 15, 1979.

1980 *Mathematics Subject Classification.* Primary 65M15; Secondary 65N30.

Key words and phrases. Parabolic equations, Galerkin methods, higher order methods.

* Sponsored by the United States Army under Contract No. DAAG29-75-C-0024. This material is based upon work supported by the National Science Foundation under Grants No. MCS76-07236 A02 and MCS78-09525.

step (a number that is independent of the discretization parameters) and still observe the optimal order errors. The overall work required by this strategy is of the magnitude of the work required by the usual Crank-Nicolson scheme applied to a linear problem with time-independent coefficients. Thus, under these conditions, they obtain a scheme that is efficient as well as effective.

The results of this paper are similar to those of [4], with regard to our higher order schemes, and are in fact in some ways stronger. In particular, because the schemes which we consider are inherently more dissipative than the Crank-Nicolson scheme, we are able to obtain the best results unconditionally. In addition, our analysis shows that the special closeness requirements of the initial values to the “elliptic projection” demanded in [4] are unnecessary and that a more natural and more easily computed initial projection may be used.

An alternate approach to the problem of finding efficient time-stepping algorithms can be found in work by Douglas and Dupont in [3]. They analyze two efficient schemes for parabolic problems. The first is a method which is of first order in time (the Laplace-modified procedure) and the second is a three-level method of second order in time.

We now introduce the parabolic problem and some convenient notation. Let $\Omega \subset \mathbb{R}^d$, $d \geq 1$, be a compact domain with a sufficiently smooth boundary $\partial\Omega$ and an outward pointing unit normal $\mathbf{n}(x) = (n_1(x), \dots, n_d(x))$. Let $\tau > 0$. The following parabolic problem will be studied under certain smoothness assumptions:

$$(1.1) \quad \begin{cases} -u_t = L(t)u = - \sum_{i,j=1}^d D_i(a_{ij}(x, t)D_j u) + a_0(x, t)u & \text{on } \Omega \times (0, \tau), \\ u|_{\partial\Omega} = 0 \quad \text{or} \quad \mathbf{n}A\nabla u|_{\partial\Omega} = \sum_{i,j=1}^d n_i a_{ij} D_j u|_{\partial\Omega} = 0 & \text{on } \partial\Omega \times (0, \tau), \\ u|_{t=0} = v & \text{on } \Omega. \end{cases}$$

Here $A = [a_{ij}(x, t)]_{i,j=1}^d$ is a symmetric, uniformly positive definite family of matrices of sufficiently smooth coefficients on $\bar{\Omega} \times [0, \tau]$, $a_0(x, t)$ is a nonnegative, sufficiently smooth function on $\bar{\Omega} \times [0, \tau]$, and $v(x)$ is a given initial data function on Ω . If the Neumann boundary conditions are under consideration, we will further require that $a_0(x, t)$ not vanish on $\bar{\Omega} \times [0, \tau]$ and that the coefficients $\{a_{ij}\}$ have the following special form:

$$a_{ij}(x, t) = a(x, t)\tilde{a}_{ij}(x), \quad 1 \leq i, j \leq d,$$

for sufficiently smooth functions a and $\{\tilde{a}_{ij}\}$. (This extra requirement ensures time independent boundary conditions). We will refer to the Dirichlet boundary conditions as BC_D and to the Neumann conditions as BC_N .

We let H^l denote the usual $L^2(\Omega)$ -based Sobolev space with norm $\|\cdot\|_l$, l a non-negative integer. We also let H_0^1 denote the subspace of H^1 that consists of functions that vanish (in the sense of trace) on $\partial\Omega$. We will use (\cdot, \cdot) to denote the usual $L^2(\Omega)$ -inner product on Ω .

The operators $\{L(t)\}_{0 \leq t \leq \tau}$ form a family of $L^2(\Omega)$ -selfadjoint elliptic operators on the following domain:

$$D_L = \begin{cases} H^2 \cap H_0^1 & \text{if we have BC}_D, \\ H^2 \cap \{w \in H^2: nA \nabla w|_{\partial\Omega} = 0\} & \text{if we have BC}_N. \end{cases}$$

Moreover, the form

$$D(t) (\cdot, \cdot) = \sum_{i,j=1}^d (a_{ij} D_j(\cdot), D_i(\cdot)) + (a_0(\cdot), (\cdot))$$

is (strongly) coercive over $H_0^1 \times H_0^1$ if we have BC_D or $H^1 \times H^1$ if we have BC_N . Thus we can apply the standard parabolic equation theory to (1.1) (cf. Friedman [5] or Lions and Magenes [6]) and obtain the existence and uniqueness of solutions u for various classes of initial data. We will always assume that $v \in D_L$ and further smoothness and compatibility conditions will be added later.

We let $T(t): L^2(\Omega) \rightarrow D_L$ denote the solution operator for $L(t)$; that is, $L(t)[T(t)f] = f$ for all $f \in L^2(\Omega)$. We note that $\{T(t)\}_{0 \leq t \leq \tau}$ is a smooth family of bounded operators from H^l to $H^{l+2} \cap D_L$, for $l \geq 0$ and that $\{L(t)\}_{0 \leq t \leq \tau}$ is a smooth family of bounded operators from $H^{l+2} \cap D_L$ to H^l , for $l \geq 0$. In fact, $L^{(j)}(t) = (d/dt)^j L(t)$, $j \geq 0$, can be calculated by differentiating the coefficients of $L(t)$ with respect to time, and if we let $T^{(j)}(t) = (d/dt)^j T(t)$, we have that

$$T^{(1)}(t) = -T(t)L^{(1)}(t)T(t).$$

We shall use the symbols C, C_1 and C_2 to denote generic, positive constants throughout this paper. The symbol C will also be used to denote a generic, increasing, positive function on \mathbf{R} . We will define $\Sigma_{j=m_1}^{m_2}(\cdot) \equiv 0$ if $m_2 < m_1$.

II. Spatial Discretization Operators. We will assume that we have a finite dimensional subspace $S_h \subset L^2(\Omega)$ (associated with parameters $0 < h < 1$ and $r \geq 2$) and a sufficiently smooth family of selfadjoint, positive semidefinite operators $\{T_h(t)\}_{0 \leq t \leq \tau}$ on $L^2(\Omega)$ that have range in S_h and that are positive definite on S_h . We define $L_h(t)$ on S_h as the inverse of $T_h(t)|_{S_h}$ for each $0 \leq t \leq \tau$. Given $f \in L^2(\Omega)$, we will regard $T_h(t)f$ as an approximation to $T(t)f$. In fact, we will require that the following be true:

$$(2.1) \quad \|(T^{(j)}(t) - T_h^{(j)}(t))f\| \leq C(j)h^{l+2} \|f\|_l \quad \text{for } j \geq 0 \text{ and } 0 \leq l \leq r - 2;$$

here $T_h^{(j)}(t) = (d/dt)^j T_h(t): L^2(\Omega) \rightarrow L^2(\Omega)$. Finally, we will assume that there is a norm $\|\cdot\|_I$ on S_h that satisfies the following:

$$(2.2) \quad \|\varphi\|^2 \leq C \|\varphi\|_I^2 \leq C(L_h(t)\varphi, \varphi) \quad \text{for } 0 \leq t \leq \tau \text{ and } \varphi \in S_h,$$

$$(2.3) \quad |(L_h^{(j)}(t)\varphi_1, \varphi_2)| \leq C(j) \|\varphi_1\|_I \|\varphi_2\|_I \quad \text{for } 0 \leq t \leq \tau, \varphi_1, \varphi_2 \in S_h \text{ and } j \geq 0;$$

here $L_h^{(j)} = (d/dt)^j L_h(t): S_h \rightarrow S_h$.

We note that many of the well-known Galerkin-type methods satisfy these assumptions. For a discussion, see Sammon [8].

We let $P_I(t) \equiv T_h(t)L(t): D_L \rightarrow S_h$, $0 \leq t \leq \tau$, define a family of "elliptic projection" operators. We also let $P: L^2(\Omega) \rightarrow S_h$ denote the $L^2(\Omega)$ -orthogonal projection onto S_h . (Note that $T_h = PT_hP$.) If $w \in H^{l+2} \cap D_L$ for some $0 \leq l \leq r-2$, then

$$(2.4) \quad \|w - Pw\| \leq \|w - P_I w\| \leq Ch^{l+2} \|w\|_{l+2} \quad \text{for } 0 \leq t \leq \tau.$$

We let $P_I^{(j)}(t) = (d/dt)^j P_I(t)$ for $j \geq 0$.

Suppose we choose $v \in H^r$ so that $u(t) \in H^r$ and $\|u(t)\|_r \leq C\|v\|_r$ for $0 \leq t \leq \tau$. Then, setting $W(t) = P_I(t)u(t)$, we have the following:

$$(2.5) \quad \|u(t) - W(t)\| \leq Ch^r \|v\|_r \quad \text{for } 0 \leq t \leq \tau.$$

We also wish to see how well the time derivatives of W approximate those of u . To this end we study the following:

PROPOSITION 2.1. *If $w \in H^{l+2} \cap D_L$ for some $0 \leq l \leq r-2$, then*

$$\|P_I^{(m)}(t)w\| \leq C(m)h^{l+2} \|w\|_{l+2} \quad \text{for } 0 \leq t \leq \tau \text{ and } m > 0.$$

Proof. We see that (2.1) implies that

$$\begin{aligned} \|P_I^{(m)}(t)w\| &\leq \left\| \sum_{j=0}^m \binom{m}{j} T_h^{(j)} L^{(m-j)} w \right\| \\ &\leq \left\| \sum_{j=0}^m \binom{m}{j} T^{(j)} L^{(m-j)} w \right\| + C(m)h^{l+2} \|w\|_{l+2} \\ &= \left\| \left(\frac{d}{dt} \right)^m TLw \right\| + C(m)h^{l+2} \|w\|_{l+2} = C(m)h^{l+2} \|w\|_{l+2}. \end{aligned}$$

We will now apply the above result. If $0 \leq l \leq r-2$, $m \geq 0$ and $v \in H^{l+2+2m}$ satisfies certain (boundary) compatibility conditions (see [5], [6] or [8]), it follows that $u^{(j)}(t) \in H^{l+2}$ and that $\|u^{(j)}(t)\|_{l+2} \leq C(m)\|v\|_{l+2+2m}$, provided that $0 \leq j \leq m$ and $0 \leq t \leq \tau$. Thus, if we set $W^{(m)}(t) = (d/dt)^m W(t)$, Proposition 2.1 and (2.1) show that

$$(2.6) \quad \begin{aligned} \|u^{(m)}(t) - W^{(m)}(t)\| &\leq \|u^{(m)}(t) - P_I u^{(m)}(t)\| + C(m) \sum_{l=1}^m \|P_I^{(l)} u^{(m-l)}(t)\| \\ &\leq C(m)h^{l+2} \|v\|_{l+2+2m} \quad \text{for } 0 \leq t \leq \tau. \end{aligned}$$

Thus, under the aforementioned conditions on v , $W^{(m)}(t)$ is uniformly bounded in $L^2(\Omega)$ for $0 \leq t \leq \tau$.

At times, we will assume that the following condition holds:

$$B_h: \quad \|L_h^{(j)}(t)T_h(s)\|, \|T_h(s)L_h^{(j)}(t)\| \leq C(j) \quad \text{for } 0 \leq s, t \leq \tau \text{ and } j \geq 0.$$

Estimates in Sammon [8] or Nassif and Descloux [7] show that this condition holds for various Galerkin-type methods if inverse assumptions are valid. The following estimate

is a consequence of Condition B_h :

$$\begin{aligned} \|L_h(t)W^{(m)}(t)\| &\leq \left\| \left(\frac{d}{dt} \right)^m L_h W(t) \right\| + C(m) \sum_{l=1}^m \|L_h^{(l)} T_h L_h W^{(m-l)}(t)\| \\ &\leq \|u^{(m+1)}(t)\| + C(m) \sum_{l=1}^m \|L_h W^{(m-l)}(t)\| \quad \text{for } 0 \leq t \leq \tau \text{ and } m \geq 0; \end{aligned}$$

here we have used the fact that $L_h W = PLu = -Pu^{(1)}$. Thus, an induction argument shows that

$$(2.7) \quad \|L_h(t)W^{(m)}(t)\| \leq C(m) \sum_{j=1}^{m+1} \|u^{(j)}(t)\| \quad \text{for } 0 \leq t \leq \tau \text{ and } m \geq 0,$$

provided that Condition B_h holds.

III. Time Discretizations. We now consider a method of computing approximations to the solution u of (1.1).

We begin by studying rational function approximations to the exponential e^{-x} on \mathbf{R}^+ . It is well known that there are rational functions $r(x) = P(x)/Q(x)$ (P and Q are relatively prime polynomials) that satisfy the following conditions:

$$(3.1) \quad \begin{cases} \text{(i)} & Q(x) > 0 \quad \text{for } x > 0, \quad Q(0) = 1, \\ \text{(ii)} & -1 + \delta < Q^{-1}(x)P(x) < 1 \quad \text{for some } \delta \geq 0 \text{ and all } x > 0, \\ \text{(iii)} & |r(x) - e^{-x}| \leq Cx^{\nu+1} \quad \text{for some } \nu \geq 1 \text{ and all } x \geq 0. \end{cases}$$

We will use P 's and Q 's that are no more than quadratics in our later work, but, for now, we assume that $P(x) = \sum_{i=0}^{\nu} p_i x^i$ and $Q(x) = \sum_{i=0}^{\nu} q_i x^i$, where $p_0 = q_0 = 1$. We have the following examples:

- (i) $P(x)/Q(x) = 1/(1+x)$ with $\delta = 1, \nu = 1$ (Backwards Euler).
- (ii) $P(x)/Q(x) = (1-x/2)/(1+x/2)$ with $\delta = 0, \nu = 2$ (Crank-Nicolson).
- (iii) The family parametrized by $\lambda \geq 1/4, \lambda \neq 1/2$:

$$P(x)/Q(x) = (1 + (2\lambda - 1)x + (\lambda^2 - 2\lambda + 1/2)x^2)/(1 + \lambda x)^2 \quad \text{with } \delta > 0, \nu = 2.$$

If $\lambda = 1/2(1 + 1/\sqrt{3})$, then $\nu = 3$ (Calahan).

(iv) $P(x)/Q(x) = 1/(1+x+x^2/2)$ with $\delta = 1, \nu = 2$ (Padé).

(v) $P(x)/Q(x) = (1-x/3)/(1+2x/3+x^2/6)$ with $\delta > 0, \nu = 3$ (Padé).

(vi) $P(x)/Q(x) = (1-x/2+x^2/12)/(1+x/2+x^2/12)$ with $\delta = 1, \nu = 4$ (Padé).

We are particularly interested in the cases where $\delta > 0$.

We will now show how property (3.1) (iii) can lead to a two-point Taylor expansion used by Nassif and Descloux in [7].

PROPOSITION 3.1. *Suppose that g is a smooth function on $[0, t_0]$. Then for each $0 \leq t \leq t_0$, we have that*

$$(3.2) \quad (Q(-tD)g)(t) = (P(-tD)g)(0) + \int_0^t K(t, s) D^{\nu+1} g(s) ds,$$

where D denotes the differentiation operator on $[0, t_0]$ and where

$$(3.3) \quad K(t, s) = \sum_{j=0}^{\nu} \frac{q_j}{(\nu-j)!} (-t)^j (t-s)^{\nu-j}.$$

Proof. From (3.1) (iii) we see that $Q(x)e^{-x} - P(x) = O(x^{\nu+1})$. Now take m derivatives, where $0 \leq m \leq \nu$, of each side of this equation and evaluate them at $x = 0$. We see that

$$m!p_m = \sum_{j=0}^m \binom{m}{j} (-1)^{m-j} j! q_j \quad \text{or} \quad p_m(-t)^m m! = \sum_{j=0}^m q_j (-t)^j \binom{m}{j} j! t^{m-j}.$$

This gives (3.2) with $g(s) = s^m$ for $0 \leq s \leq t_0$. Thus we have (3.2) if g is a polynomial of degree no more than ν .

We now expand the sufficiently smooth function g in a Taylor series of degree ν and apply our above work. This shows that (3.2) holds with the kernel given and completes the proof.

If we let $\{u_h(t)\}_{0 \leq t \leq \tau} \subset S_h$ be defined by the equations

$$(3.4) \quad u_{h,t}(t) + L_h(t)u_h(t) = 0 \quad \text{for } 0 \leq t \leq \tau \text{ and } u_h(0) = v_h,$$

where $v_h \in S_h$ is some function "close" to v (for instance, $v_h = Pv$), then work by Sammon [8] shows that

$$(3.5) \quad \|u(t) - u_h(t)\| = O(h^r) \quad \text{for } 0 \leq t \leq \tau,$$

under certain conditions. Thus, if we could approximate the solution u_h of (3.4) with a known small error, we could use (3.5) to show that our approximation is actually close to u . We will use our two-point Taylor polynomial to construct an approximation to u_h .

Let $0 < k < 1$ so that $Mk = \tau$ for some integer $M \geq 1$. We will study a method of approximating $u_h(k)$. Choose a rational function $r(x) = P(x)/Q(x)$ that satisfies (3.1) (i) through (3.1) (iii) and where the degrees of P and Q are two or less. (This implies that $\nu \leq 4$.) If we note that

$$\begin{aligned} u_h^{(1)}(t) &= \frac{d}{dt} u_h(t) = -L_h(t)u_h(t) \quad \text{for } 0 \leq t \leq \tau, \\ u_h^{(2)}(t) &= \left(\frac{d}{dt}\right)^2 u_h(t) = (L_h^2(t) - L_h^{(1)}(t))u_h(t) \quad \text{for } 0 \leq t \leq \tau, \end{aligned}$$

then setting $g(s) = u_h(s)$, $0 \leq s \leq k$, and using (3.2) gives us the following:

$$(3.6) \quad \begin{aligned} &\{I + q_1 k L_h(k) + q_2 k^2 (L_h^2(k) - L_h^{(1)}(k))\} u_h(k) \\ &= \{I + p_1 k L_h(0) + p_2 k^2 (L_h^2(0) - L_h^{(1)}(0))\} u_h(0) + O(k^{\nu+1} u_h^{(\nu+1)}). \end{aligned}$$

Thus, if the quantity in the first set of braces is invertible, we might expect the following to be "close" to $u_h(k)$:

$$\{I + q_1 k L_h + q_2 k^2 (L_h^2 - L_h^{(1)})\}^{-1}(k) \{I + p_1 k L_h + p_2 k^2 (L_h^2 - L_h^{(1)})\}(0) v_h.$$

This scheme for approximating $u_h(k)$ will be the basis of a scheme for approximating $u_h(Nk)$ for any $1 \leq N \leq M$. It will be seen that the approximations can continue to be defined using operators that are constructed from P, Q and the family $\{L_h^{(j)}(t)\}_{j=0,1}$. We note that these schemes can be defined if the degrees of P and Q are higher than two, as was done in [8].

The solution u_h of (3.4) only plays a motivational role and will not enter in any way in the rest of this work. For purposes of our estimates, we will need some function in S_h that is uniformly close to u in the sense of (3.5), and u_h would be a possible candidate. We will use W for this purpose, however, mainly because it is easy to estimate its time derivatives.

For $0 \leq n \leq M$, let $t_n = nk$, $L_n^{(j)} = L_h^{(j)}(t_n)$ ($j \geq 0$), $T_n^{(j)} = T_h^{(j)}(t_n)$ ($j \geq 0$), $P_n = P(kL_n)$, $Q_n = Q(kL_n)$, $\tilde{P}_n = P_n - p_2 k^2 L_n^{(1)}$, $\tilde{Q}_n = Q_n - q_2 k^2 L_n^{(1)}$. We now settle the question of \tilde{Q}_n 's invertibility on S_h .

PROPOSITION 3.2. *If k is sufficiently small, we have that*

$$(3.7) \quad C_1(Q_n \varphi, \varphi) \leq (\tilde{Q}_n \varphi, \varphi) \leq C_2(Q_n \varphi, \varphi) \quad \text{for } 0 \leq n \leq M \text{ and } \varphi \in S_h.$$

Proof. We first see that if $0 \leq n \leq M$ and $\varphi \in S_h$, then

$$|(Q_n \varphi, \varphi) - (\tilde{Q}_n \varphi, \varphi)| = k^2 q_2 |(L_n^{(1)} \varphi, \varphi)| \leq Ck^2 q_2 \|\varphi\|_I^2 \leq Ckq_2 (kL_n \varphi, \varphi),$$

where we have used (3.1) (i) to show that $q_2 \geq 0$. Then, if $q_2 > 0$, $x \leq (\sup_{0 \leq y < \infty} yQ^{-1}(y))Q(x)$ for any $0 \leq x < \infty$ and it follows that $(kL_n \varphi, \varphi) \leq C(Q_n \varphi, \varphi)$. The proof is now easily completed.

We shall assume throughout this work that k is small enough to allow the conclusion of Proposition 3.2 to hold. Thus, since Q_n is invertible due to $Q(x)$'s positivity on \mathbf{R}^+ , it follows that the selfadjoint operator \tilde{Q}_n is also invertible, for $0 \leq n \leq M$.

We now return to our description of an approximation to the solution of (1.1). Given v , we choose a $V^0 \in S_h$ that is close to v (for instance $V^0 = Pv$, although we will describe perhaps better possibilities later) and recursively define V^{n+1} ($0 \leq n < M$), given V^n , by the following formula:

$$(3.8) \quad \tilde{Q}_{n+1} V^{n+1} = \tilde{P}_n V^n, \quad 0 \leq n < M.$$

We expect V^n to be close to $u_h(t_n)$ which is in turn close to $u(t_n)$, where $0 \leq n \leq M$, and we will derive corresponding estimates.

As noted before, this approximation scheme has been studied in an $L^2(\Omega)$ setting by Nassif and Descloux in [7]. We now see how computation of this scheme involves solving a new linear problem at each time step and why a more efficient variant would be desirable. We shall later study the variant suggested by Douglas, Dupont, and Ewing in [4]. However, since the analysis of this variant requires estimates of the original scheme (the one defined by (3.8)) in new norms, we shall first present another analysis for the latter. We shall also define a natural choice for V^0 .

IV. Preliminary Error Estimates. We are primarily interested in how close V^n is to $W^n \equiv W(t_n)$ ($0 \leq n \leq M$) since we already know (recall (2.6)) how close W^n is to $u^n \equiv u(t_n)$ ($0 \leq n \leq M$). As noted before, these estimates are already known in the $L^2(\Omega)$ -norm, but we wish to study them in the (possibly) stronger norms given by $(Q_n(\cdot), \cdot)^{1/2}$. This will allow us to study a variant of the scheme where the $(Q_n(\cdot), \cdot)^{1/2}$ -norm is in some sense a natural norm of the problem. We note that most of our work will go on in S_h in the next two sections.

Letting $E^n = V^n - W^n$ for $0 \leq n \leq M$, we see that

$$(4.1) \quad \begin{aligned} Q_{n+1}E^{n+1} &= P_{n+1}E^n + (P_n - P_{n+1})E^n + (\tilde{P}_n - P_n)E^n + (Q_{n+1} - \tilde{Q}_{n+1})E^{n+1} \\ &\quad - (\tilde{Q}_{n+1}W^{n+1} - \tilde{P}_nW^n) \quad \text{for } 1 \leq n + 1 \leq M. \end{aligned}$$

This will be an important error equation.

We note that (4.1) is of the form $QX = PV + F$, where $X, V, F \in S_h$ and Q and P are selfadjoint operators on S_h that satisfy the following:

$$(4.2) \quad \left\{ \begin{aligned} (i) \quad &(Q\varphi, \varphi) > 0 \quad \text{for } 0 \neq \varphi \in S_h, \\ (ii) \quad &((Q - P)\varphi, \varphi) > 0 \quad \text{for } 0 \neq \varphi \in S_h, \\ (iii) \quad &((Q + P)\varphi, \varphi) > \delta(Q\varphi, \varphi) \quad \text{for } 0 \neq \varphi \in S_h, \text{ where } \delta \geq 0. \end{aligned} \right.$$

(Of course (4.2) (i) through (4.2) (iii) follow from (3.1) (i) through (3.1) (iii).) This situation leads to the following

PROPOSITION 4.1. *Let $QX = PV + F$, where Q and P are selfadjoint operators that satisfy (4.2). Then we have that*

$$(4.3) \quad (QX, X) \leq (QV, V) - \delta((Q - P)V, V) + 2(F, X).$$

Proof. Let $Y = Q^{-1}PV$. Then $QX = QY + F$ and

$$\begin{aligned} (QX, X) &= (QY, X) + (F, X) = (QY, Y) + (QY, Q^{-1}F) + (F, X) \\ &= (QY, Y) + (QX, Q^{-1}F) - (F, Q^{-1}F) + (F, X) \leq (QY, Y) + 2(F, X). \end{aligned}$$

Set $\hat{Y} = Q^{1/2}Y$, $\hat{V} = Q^{1/2}V$ and $A = Q^{-1/2}PQ^{-1/2}$. Thus $\hat{Y} = A\hat{V}$. Now (4.2) implies that $(I - A)$ and $((1 - \delta)I + A)$ are positive definite operators on S_h , and since they commute, we have that

$$\|\varphi\|^2 - \delta((I - A)\varphi, \varphi) - \|A\varphi\|^2 = ((I - A)((1 - \delta)I + A)\varphi, \varphi) \geq 0$$

for all $\varphi \in S_h$. We now complete the proof by observing that

$$(QY, Y) = \|\hat{Y}\|^2 = \|A\hat{V}\|^2 \leq \|\hat{V}\|^2 - \delta((I - A)\hat{V}, \hat{V}) = (QV, V) - \delta((Q - P)V, V).$$

We will want to apply Proposition 4.1 to (4.1) and obtain an estimate for $(Q_{n+1}E^{n+1}, E^{n+1})$. In anticipation of this, we prove the following estimates.

LEMMA 4.2. *Suppose that Condition B_h holds if $Q(x)$ is quadratic. Then, if $0 \leq n \leq n + m \leq M$ and $\varphi_1, \varphi_2 \in S_h$, we have that*

$$(4.4) \quad \left. \begin{aligned} |((P_{n+m} - P_n)\varphi_1, \varphi_2)| \\ |((Q_{n+m} - Q_n)\varphi_1, \varphi_2)| \end{aligned} \right\} \leq Ct_m(Q_n\varphi_1, \varphi_1)^{1/2}(Q_n\varphi_2, \varphi_2)^{1/2}.$$

Proof. We first note that (3.1) (ii) implies that the degree of $P(x)$ is no greater than the degree of $Q(x)$ and (3.1) (ii) and (3.1) (iii) imply that $Q(x)$ cannot be the constant 1. Also, if we let $R(x) = 1 + x$ if $q_2 = 0$ and $R(x) = 1 + x + x^2$ if $q_2 \neq 0$, we see that $R^{-1}(x)Q(x), R(x)Q^{-1}(x) \leq C$ for $x \geq 0$. Thus, letting $R_j = R(kL_j)$ ($0 \leq j \leq M$), we have that

$$C_1(R_j\varphi, \varphi) \leq (Q_j\varphi, \varphi) \leq C_2(R_j\varphi, \varphi) \quad \text{for } \varphi \in S_h \text{ and } 0 \leq j \leq M.$$

Thus it suffices to prove (4.4) with R_n -inner products on the right-hand side.

We have the following estimates:

$$\begin{aligned}
 k|((L_{n+m} - L_n)\varphi_1, \varphi_2)| &\leq Ct_m k \sup_{t_n \leq s \leq t_{n+m}} |(L_h^{(1)}(s)\varphi_1, \varphi_2)| \\
 (4.5) \quad &\leq Ct_m k \|\varphi_1\|_I \|\varphi_2\|_I \leq Ct_m k (L_n \varphi_1, \varphi_1)^{1/2} (L_n \varphi_2, \varphi_2)^{1/2} \\
 &\leq Ct_m (R_n \varphi_1, \varphi_1)^{1/2} (R_n \varphi_2, \varphi_2)^{1/2},
 \end{aligned}$$

$$\begin{aligned}
 k^2|((L_{n+m}^2 - L_n^2)\varphi_1, \varphi_2)| &\leq k^2|((L_{n+m} - L_n)\varphi_1, L_{n+m}\varphi_2)| \\
 &\quad + k^2|(L_n\varphi_1, (L_{n+m} - L_n)\varphi_2)| \\
 (4.6) \quad &\leq Ct_m k^2 \left(\sup_{t_n \leq s \leq t_{n+m}} \|L_h^{(1)}(s)T_n\| \right) (1 + \|L_{n+m}T_n\|) \|L_n\varphi_1\| \|L_n\varphi_2\| \\
 &\leq Ct_m (R_n \varphi_1, \varphi_1)^{1/2} (R_n \varphi_2, \varphi_2)^{1/2} \text{ if } q_2 \neq 0;
 \end{aligned}$$

note that Condition B_h was used in (4.6).

We can now use (4.5) and (4.6) to complete the proof.

LEMMA 4.3. *If $0 \leq n \leq M$ and $\varphi_1, \varphi_2 \in S_h$, we have that*

$$(4.7) \quad \left. \begin{aligned} &|((\tilde{P}_n - P_n)\varphi_1, \varphi_2)| \\ &|((\tilde{Q}_n - Q_n)\varphi_1, \varphi_2)| \end{aligned} \right\} \leq Ck(Q_n \varphi_1, \varphi_1)^{1/2} (Q_n \varphi_2, \varphi_2)^{1/2}.$$

Proof. The result essentially follows from the proof of Proposition 3.2.

We now study the truncation error term in (4.1) by comparing it to the true solution u of (1.1).

PROPOSITION 4.4. *Suppose that $v \in H^\mu$, $\mu = \max(2(v + 1), r + 2)$, is such that $\|u(t)\|_{r+2} \leq C\|v\|_{r+2}$, $\|u^{(v+1)}(t)\| \leq C\|v\|_{2(v+1)}$ for $0 \leq t \leq \tau$ and we have Condition B_h if $Q(x)$ is quadratic. Then, if $0 \leq n < M$ and $\varphi \in S_h$, we have that*

$$(4.8) \quad |(\tilde{Q}_{n+1}W^{n+1} - \tilde{P}_nW^n, \varphi)| \leq Ck(h^r\|v\|_{r+2} + k^\nu\|v\|_{2(v+1)})(Q_n\varphi, \varphi)^{1/2}.$$

Proof. We note that $u_t = -Lu$, $u_{tt} = (L^2 - L^{(1)})u = -L(I + T^{(1)})u_t$ for $0 \leq t \leq \tau$ and that Proposition 3.1 implies that

$$(4.9) \quad \|(u^{n+1} - q_1ku_t^{n+1} + q_2k^2u_{tt}^{n+1}) - (u^n - p_1ku_t^n + p_2k^2u_{tt}^n)\| \leq Ck^{\nu+1}\|v\|_{2(v+1)}.$$

We also have the following for $0 \leq j \leq M$: $W^j = P_I u^j = P u^j + (P_I - P)u^j$,

$$(4.10) \quad L_j W^j = L_h(t_j)T_h(t_j)P L(t_j)u(t_j) = -P u_t^j,$$

$$\begin{aligned}
 (L_j^2 - L_j^{(1)})W^j &= -L_j(P + T_j^{(1)})u_t^j \\
 (4.11) \quad &= -L_j(P_I u_t^j + P_I T^{(1)}(t_j)u_t^j) + L_j(P_I - P)u_t^j \\
 &\quad + L_j(P_I - P)T^{(1)}(t_j)u_t^j + L_j P(T^{(1)}(t_j) - T_j^{(1)})u_t^j,
 \end{aligned}$$

and $-L_j P_I(u_t^j + T^{(1)}(t_j)u_t^j) = P u_{tt}^j$, where we have used the fact that $T_j^{(1)} = -T_j L_j^{(1)} T_j$.

We now use these facts to see that

$$\begin{aligned}
 & |(\tilde{Q}_{n+1}W^{n+1} - \tilde{P}_nW^n, \varphi)| \\
 & \leq Ck^{\nu+1}\|v\|_{2(\nu+1)}\|\varphi\| + C\left(\int_{t_n}^{t_{n+1}} \|W^{(1)}(s) - Pu^{(1)}(s)\| ds\right)\|\varphi\| \\
 & \quad + C\{\|(P_I - P)u_t^{n+1}\| + \|(P_I - P)u_t^n\| + \|(P_I - P)T^{(1)}(t_{n+1})u_t^{n+1}\| \\
 & \quad \quad + \|(P_I - P)T^{(1)}(t_n)u_t^n\| + \|(T^{(1)}(t_{n+1}) - T_{n+1}^{(1)})u_t^{n+1}\| \\
 & \quad \quad \quad + \|(T^{(1)}(t_n) - T_n^{(1)})u_t^n\|\} \\
 & \quad \cdot k^2(\|L_{n+1}\varphi\| + \|L_n\varphi\|) \\
 & \leq Ck(h^r\|v\|_{r+2} + k^\nu\|v\|_{2(\nu+1)})\{\|\varphi\|^2 + k^2\|L_n\varphi\|^2\}^{1/2},
 \end{aligned}$$

which gives our result.

We can now put these results together and demonstrate a bound in the $Q_n^{1/2}$ -norm for the difference between V^n and W^n for $0 \leq n \leq M$.

THEOREM 4.5. *Suppose that v is sufficiently smooth and compatible (the hypotheses of Proposition 4.4 would suffice) and that we have Condition B_h if $Q(x)$ is quadratic. Then, if k is sufficiently small, we have the following:*

$$(4.12) \quad \|Q_N^{1/2}(V^N - W^N)\| \leq C(h^r + k^\nu)\|v\|_\mu + C\|Q_0^{1/2}(V^0 - P_I v)\| \quad \text{for } 0 \leq N \leq M.$$

Proof. Let $0 \leq n < M$. An application of the results of this section to the terms on the right of (4.1) shows that

$$\begin{aligned}
 (Q_{n+1}E^{n+1}, E^{n+1}) & \leq (1 + Ck)(Q_nE^n, E^n) + Ck(k^\nu + h^r)^2\|v\|_\mu^2 \\
 & \quad + Ck(Q_{n+1}E^{n+1}, E^{n+1}),
 \end{aligned}$$

which gives the result.

We will now examine possibilities for the starting function V^0 . We require that $V^0 - P_I v$ be bounded by $C(k^\nu + h^r)$ in the $Q_0^{1/2}$ -norm if we wish a comparable error in (4.12). We can always let $V^0 = P_I v$, but we note that the approximation scheme defined by (3.8) never requires that we determine T_h applied to any function. In applications, this would amount to a special, expensive calculation required only at the beginning. There is another choice for V^0 that involves solving a system with \tilde{Q}_0 (or even Q_0). Since such systems have to be solved anyway to take the first step, this would seem to be a better approach.

We have

PROPOSITION 4.6. *If $v \in H^{r+2} \cap D_L$ and $L(0)v \in D_L$, define $V^{0,1}$ and $V^{0,2}$ in S_h by the following:*

$$(4.13) \quad \begin{cases} \tilde{Q}_0 V^{0,1} = P(v + q_1 kL(0)v + q_2 k^2(L^2(0) - L^{(1)}(0))v), \\ Q_0 V^{0,2} = P(v + q_1 kL(0)v + q_2 k^2 L^2(0)v) = PQ(kL(0))v. \end{cases}$$

Then $\|Q_0^{1/2}V^{0,1}\| + \|Q_0^{1/2}V^{0,2}\| + \|Q_0^{1/2}P_I v\| \leq C\|v\|_4$ and

$$(4.14) \quad \left. \begin{aligned} \|Q_0^{1/2}(V^{0,1} - P_I v)\| \\ \|Q_0^{1/2}(V^{0,2} - P_I v)\| \end{aligned} \right\} \leq Ch^r \|v\|_{r+2}.$$

Proof. We observe that the following hold for $\varphi \in S_h$:

$$\begin{aligned} (L_0 P_I v - L(0)v, \varphi) &= ((P - I)L(0)v, \varphi) = 0, \\ (L_0^2 P_I v - L^2(0)v, \varphi) &= (L_0 PL(0)v - PL^2(0)v, \varphi) = (L_0(PL(0)v - P_I L(0)v), \varphi) \\ &= ((P - P_I)L(0)v, L_0 \varphi) = ((I - P_I)L(0)v, L_0 \varphi), \\ (L_0^{(1)} P_I v - L^{(1)}(0)v, \varphi) &= -(L_0 T_0^{(1)} L(0)v - L(0)T^{(1)}(0)L(0)v, \varphi) \\ &= (L_0(P_I T^{(1)}(0) - T_0^{(1)})L(0)v, \varphi) = ((T^{(1)}(0) - T_0^{(1)})L(0)v, L_0 \varphi) \\ &\quad + ((P_I - I)T^{(1)}(0)L(0)v, L_0 \varphi). \end{aligned}$$

Thus, if we let $\varphi = V^{0,1} - P_I v$, then

$$\begin{aligned} \|Q_0^{1/2}(V^{0,1} - P_I v)\|^2 &\leq C \|\tilde{Q}_0^{1/2}(V^{0,1} - P_I v)\|^2 = C(\tilde{Q}_0(V^{0,1} - P_I v), \varphi) \\ &\leq |(v - P_I v, \varphi)| + q_2 k^2 Ch^r \|v\|_{r+2} \|L_0 \varphi\| \\ &\leq Ch^r \|v\|_{r+2} \|Q_0^{1/2} \varphi\|. \end{aligned}$$

This completes the proof of the first part of (4.14) and the remainder of the estimates follow even more simply from the above observations.

This result completes our error analysis for the approximation scheme defined by (3.8) with V^0 defined by either equation of (4.13). We will call this the *base scheme* in the sequel.

We note that if $V^{0,2}$ is defined by

$$(4.15) \quad \bar{Q}(kL_0)V^{0,2} = P\bar{Q}(kL(0))v,$$

where $\bar{Q}(x)$ is a polynomial that satisfies $C_1 < \bar{Q}(x)/Q(x) < C_2$ for $0 \leq x < \infty$, an estimate like (4.14) will still hold. This modification might prove useful if $\bar{Q}(kL_0)$ is a preconditioning operator for the kind of linear system solving techniques we will study in the next section.

V. A Variant of the Base Scheme. As we noted before, the calculation of the base scheme involves the solution of a new linear problem at each time step. We wish to study a variant of this scheme where we only approximately solve the linear system at each time step. We propose to use an iterative technique for this purpose which, as we will see, can be provided with a good initial guess for the true solution.

If we are at a point in our calculations where we have several accurate approximations to the function u at previous time steps, it can be seen that there is an extrapolation of these values that yields just as good an approximation at the next time step. The smoothness of u makes this possible. This extrapolation could be used as an initial guess for an iterative procedure. But this observation raises a question. Since even the exact solution of the system which we are approximately solving is no closer to u (in

the sense of order) than the extrapolated guess, why iterate at all? If we did no iterations and used this procedure as an algorithm to generate further approximations, errors would grow and the approximations would deteriorate. Such an algorithm is not stable. Of course, the base algorithm (solve the system exactly and forget about iterations) can easily be shown to be stable although we will not formally state this result. Also, it will not be too hard to see that if we make an error in approximately solving the system that is of the order of the local truncation error and that is in some sense independent of the initial guess, then the algorithm is stable and gives accurate approximations. For the iterative schemes that we will consider, this strategy requires a quantity of iterations that is on the order of the logarithm of the total number of time steps, per time step. However, there is a more efficient strategy available if the polynomials $P(x)$ and $Q(x)$ have the correct properties. If one does in fact give a good initial guess to the iterative scheme and then iterates only a certain number of times per time step (a number that is independent of the total number of steps), then, even though accuracy is not improved, the resulting algorithm is stable and generates accurate approximations for u . This phenomenon was first observed in [4] in relation to the Crank-Nicolson scheme. We will give arguments in this section that show that these results hold for schemes that have the right kind of dissipation; that is, P and Q are such that $\delta > 0$. Similar results can be proven for polynomial pairs that are just stable ($\delta = 0$) but the condition $k \leq Ch^2$, for some constant C , is required. This condition introduces dissipation and was used in [4].

We begin by discussing the properties of a particular type of preconditioned iterative technique for solving linear systems. We will assume that we are working in a finite dimensional space H with an inner product $(\cdot, \cdot)_H$ and a norm $\|\cdot\|_H = (\cdot, \cdot)_H^{1/2}$. Suppose A is a positive definite selfadjoint operator on H , and we wish to find an approximation to the vector x that satisfies $Ax = y$, where y is known. We will also assume that the situation is such that we have another positive definite selfadjoint operator A_0 at our disposal for which $A_0^{-1}z, z \in H$, is easy to find and for which we know the following spectral estimate:

$$(5.1) \quad \lambda_0(A_0z, z)_H \leq (Az, z)_H \leq \lambda_1(A_0z, z)_H \quad \text{for } z \in H,$$

where $0 < \lambda_0 \leq \lambda_1$ are known constants. Then there are methods which, when given an initial guess $x^{(0)}$ for the solution x , generate a sequence of approximations $x^{(\alpha)}$, $\alpha \geq 1$, to x that have the following properties:

- (i) The calculation of $x^{(\alpha+1)}$, given $\{x^{(j)}\}_{j=0}^\alpha$, only requires evaluating A and A_0 , solving systems involving A_0 , and Hilbert space operations.
- (ii) The sequence $x^{(\alpha)} \rightarrow x$ as $\alpha \rightarrow \infty$ geometrically in the following way. There is a smooth decreasing function $0 \leq \gamma(\xi) < 1$ ($0 < \xi \leq 1$) that satisfies $\gamma(1) = 0$ and which gives the rate of convergence of the iterative scheme in the $A_0^{1/2}$ -norm:

$$(5.2) \quad \|A_0^{1/2}(x - x^{(\alpha)})\|_H \leq C \left(\frac{\lambda_1}{\lambda_0}\right) \gamma^\alpha \left(\frac{\lambda_0}{\lambda_1}\right) \|A_0^{1/2}(x - x^{(0)})\|_H \quad \text{for } \alpha \geq 0.$$

A given method may or may not actually use the spectrum estimation constants λ_0 and

λ_1 in its calculations. We also note that (5.2) implies that the following estimate holds if $C(\lambda_1/\lambda_0)\gamma^\alpha(\lambda_0/\lambda_1) < 1$ for some $\alpha > 0$:

$$(5.3) \quad \begin{aligned} & \|A_0^{1/2}(x - x^{(\alpha)})\|_H \\ & \leq C\left(\frac{\lambda_1}{\lambda_0}\right) \left(1 - C\left(\frac{\lambda_1}{\lambda_0}\right)\gamma^\alpha\left(\frac{\lambda_0}{\lambda_1}\right)\right)^{-1} \gamma^\alpha\left(\frac{\lambda_0}{\lambda_1}\right) \|A_0^{1/2}(x^{(\alpha)} - x^{(0)})\|_H. \end{aligned}$$

The preconditioned conjugate gradient algorithm fits into this framework with $\gamma = (1 - \xi^{1/2})/(1 + \xi^{1/2})$; see [1]. We also note that another example of such an algorithm is given by the following splitting method. Let $\mu > 0$ and, given $x^{(\alpha)}$ for some $\alpha \geq 0$, define $x^{(\alpha+1)}$ by the following:

$$A_0x^{(\alpha+1)} = \mu y + (A_0 - \mu A)x^{(\alpha)}.$$

This method converges for certain values of μ , and if we choose $\mu = 2/(\lambda_1 + \lambda_0)$, then we have a method that satisfies (i), (ii) above with $\gamma = (1 - \xi)/(1 + \xi)$; see [1] or [9].

We intend to use an iterative scheme with the properties described above to approximately solve the system (3.8) which defines our base scheme. The Hilbert space H will be S_h with the $L^2(\Omega)$ -inner product and the above discussions outline possible error results. We will keep the conjugate gradient algorithm in mind since it offers a good convergence rate and it does not require the values of λ_0 and λ_1 in its calculations; they only enter into its error analysis.

We must now decide what to use as a preconditioning operator. We note that the contribution of the $L_{n+1}^{(1)}$ term in \tilde{Q}_{n+1} is small, so we can ignore this term when constructing a preconditioner. We now discuss two possibilities:

A. Q_0 as a preconditioning operator. This is a good choice if linear systems involving Q_0 are easy to solve. For instance, if $Q(x) = 1 + q_1x$ is linear, then $Q_0 = I + q_1kL_0$ has essentially the same structure as the L_0 operator and solving this type of problem is well studied. If $Q(x) = (1 + \lambda x)^2$, as in the Calahan method, then solving systems with $Q_0 = (I + \lambda kL_0)^2$ only involves solving two successive systems with the $(I + \lambda kL_0)$ operator. Thus, Q_0 is also a good preconditioner for this method. If $Q(x)$ is not a perfect square, the fourth order diagonal Padé scheme being a notable example, there are other methods for solving systems involving Q_0 that use complex arithmetic. Thus, using Q_0 as a preconditioning operator is possible for these methods. We will offer an alternative in B however.

We observe that the following result contains (5.1) for Q_0 as a preconditioning operator.

PROPOSITION 5.1. *Let $0 \leq m, n \leq M$, and assume Condition B_h if $Q(x)$ is quadratic. Then, if $\varphi \in S_h$,*

$$(5.4) \quad (1 + C(|t_n - t_m| + k))^{-1}(Q_n\varphi, \varphi) \leq (\tilde{Q}_m\varphi, \varphi) \leq (1 + C(|t_n - t_m| + k))(Q_n\varphi, \varphi).$$

Proof. We see that

$$(\tilde{Q}_m\varphi, \varphi) = (Q_n\varphi, \varphi) + ((Q_m - Q_n)\varphi, \varphi) + ((\tilde{Q}_m - Q_m)\varphi, \varphi),$$

and Lemmas 4.2 and 4.3 show that

$$|((Q_m - Q_n)\varphi, \varphi) + ((\tilde{Q}_m - Q_m)\varphi, \varphi)| \leq C(|t_n - t_m| + k)(Q_n\varphi, \varphi).$$

This gives the second inequality and the first is done in a similar fashion using Proposition 3.2.

Thus, if we are using an iterative technique as described above, with the preconditioning operator Q_0 , to approximately solve (3.8) at some time step $1 \leq n + 1 \leq M$, we can expect an error reduction of at least $\gamma^\alpha((1 + Ct_{n+1})^{-2})$ (in the sense of (5.3)) after α iterations. Since $\gamma(1) = 0$, $\gamma((1 + Ct_{n+1})^{-2}) \leq Ct_{n+1}$ and hence the error reduction is of order $O(k^\alpha)$ for the first few steps. In anticipation of later results, we note that if $0 < \epsilon < 1$ is given, there is an $\alpha = \alpha(\epsilon)$ so that $\gamma^\alpha((1 + Ct_{n+1})^{-2}) \leq \min(\epsilon, t_{n+1}^{1/2})$. Thus, only a fixed number of iterations $\alpha = \alpha(\epsilon)$ would be required to achieve this error reduction at each step.

Now we consider another possibility for a preconditioning operator that is useful if $Q(x)$ is quadratic but not a perfect square. Let $\lambda > 0$ and set $S_n = I + \lambda kL_n$ for $0 \leq n \leq M$.

B. $S_0 = (I + \lambda kL_0)$ or S_0^2 as a preconditioning operator. We first note that it is easy to solve systems using these operators; that is, we only need to solve (perhaps successive) systems with the $(I + \lambda kL_0)$ operator. We can also prove the following result by the methods used in the proofs of Proposition 3.2 and Lemmas 4.2 and 4.3:

PROPOSITION 5.2. *Let $0 \leq m, n \leq M$ and l be the degree of $Q(x)$. Suppose that Condition B_h is satisfied if $Q(x)$ is quadratic. Then, if $\varphi \in S_n$, we have that*

$$(5.5) \quad C_1(S_n^l\varphi, \varphi) \leq (\tilde{Q}_m\varphi, \varphi) \leq C_2(S_n^l\varphi, \varphi).$$

Thus, if we are using an iterative technique with S_0^l as a preconditioning operator (where l is the degree of $Q(x)$) to approximately solve (3.8) at some time step $1 \leq n + 1 \leq M$, we can expect an error reduction of at least $\gamma^\alpha(C_1/C_2)$ (in the sense of (5.3)) after α iterations. Also, for each $0 < \epsilon < 1$, there is an $\alpha_{n+1} = \alpha_{n+1}(\epsilon)$ so that $\gamma^{\alpha_{n+1}}(C_1/C_2) \leq \min(\epsilon, t_{n+1}^{1/2})$ and so that $\alpha_{n+1} \leq C(|\log t_{n+1}| + |\log \epsilon|)$. Thus, the number of iterations $\alpha_{n+1} = \alpha_{n+1}(\epsilon)$ required to achieve the given error reduction of $\min(\epsilon, t_{n+1}^{1/2})$ at the $(n + 1)$ st step is a function of n and becomes large if $t_{n+1} = O(k)$. However, since

$$\frac{1}{M} \sum_{l=1}^M (|\log t_l| + |\log \epsilon|) \leq C(1/\epsilon),$$

the average number of iterations per time step required, to attain the given error reduction at each time step, can be assumed to be bounded independently of k .

We now gather these ideas. We will assume that we have chosen a preconditioning operator, which we will call PQ , and we have Condition B_h if $Q(x)$ is quadratic. Thus, we can assume that

$$(5.6) \quad C_1({}^PQ\varphi, \varphi) \leq (\tilde{Q}_n\varphi, \varphi) \leq C_2({}^PQ\varphi, \varphi) \quad \text{for } 0 \leq n \leq M \text{ and } \varphi \in S_n.$$

We also assume that we have an iterative linear system solving process \mathbf{P} which uses this preconditioning operator. We now wish to use \mathbf{P} to calculate approximations to the solutions of systems like $\tilde{Q}_n\chi = F$, where $0 \leq n \leq M$, assuming we have been

given the right-hand side F , an initial guess χ_0 for χ , and a tolerance $\beta_n > 0$. We will assume that there is an $\alpha_n = \alpha_n(\beta_n) \geq 1$, so that if $\chi^{(\alpha_n)}$ is the α_n th iterate of the process \mathbf{P} applied to this system, then

$$(5.7) \quad \|{}^P Q^{1/2}(\chi - \chi^{(\alpha_n)})\| \leq \beta_n \|{}^P Q^{1/2}(\chi - \chi_0)\|.$$

Finally, we will make an assumption about the total number of iterations needed to achieve certain tolerances. If $\beta_n = \min(\epsilon, t_n^{1/2})$ for $1 \leq n \leq M$, where $0 < \epsilon < 1$, then we will require that $(1/M)\sum_{n=1}^M \alpha_n \leq C(1/\epsilon)$; that is, we only need finitely many iterations per time step, on the average, to achieve these tolerances.

With the process \mathbf{P} at hand, we now formally state a variant of the algorithm stated in (3.8). First of all, given v , we choose a $U_0 \in S_h$ that is close to v , and, given a set $\{\beta_n\}_{n=1}^M$ of positive tolerances, we define U^{n+1} in terms of $\{U^j\}_{j=0}^n$, $0 \leq n \leq M - 1$, in the following way. We use enough iterations of the process \mathbf{P} (which uses the preconditioning operator ${}^P Q$) to generate an approximation U^{n+1} to the (true) solution \bar{U}^{n+1} of the following system:

$$(5.8) \quad \tilde{Q}_{n+1} \bar{U}^{n+1} = \tilde{P}_n U_n,$$

where the error made is to be less than the tolerance β_{n+1} , in the sense of (5.7). We use

$$(5.9) \quad Z_{n+1}(U) = \sum_{j=0}^n \gamma_{n+1,j} U^j$$

for certain coefficients $\{\gamma_{n+1,j}\}_{j=0}^n$, as an initial guess. (We will fix values for these coefficients later. Of course, letting $Z_{n+1}(U) = U^n$ is a possibility and, in general, we will never use more than the past few values.)

If we redefine $E^n = U^n - W^n$ for $0 \leq n \leq M$, we note that we have the following important identity, an analogue of (4.1):

$$(5.10) \quad \begin{aligned} Q_{n+1} E^{n+1} &= P_{n+1} E^n + (P_n - P_{n+1}) E^n + (\tilde{P}_n - P_n) E^n \\ &+ (Q_{n+1} - \tilde{Q}_{n+1}) E^{n+1} - (\tilde{Q}_{n+1} W^{n+1} - \tilde{P}_n W^n) \\ &+ \tilde{Q}_{n+1} (U^{n+1} - \bar{U}^{n+1}) \quad \text{for } 0 \leq n \leq M - 1. \end{aligned}$$

We now analyze the error made by this kind of approximation algorithm. We will begin by studying a result that is easy to obtain but is not the best possible for our situation. We will briefly assume that we solve (5.8) to an error of $\beta_{n+1} = k^\nu$ for $0 \leq n \leq \nu - 2$ (if $\nu \geq 2$) and to an error of $\beta_{n+1} = k$ for $n \geq \nu - 1$. We note that these latter tolerances imply that, for our types of processes \mathbf{P} , we must do on the order of $\log(M) = \log(\tau/k)$ iterations per time step in general. One must expect these tolerances to lead to good error estimates (if we use the appropriate initial guesses), and we will show that this is indeed the case. However, we will later show that we can get the same type of estimates for a modified algorithm that only requires finitely many iterations per time step, on the average.

We will assume that v is sufficiently smooth and compatible for this discussion. In particular, this implies that we can take U^0 to be an approximation generated by \mathbf{P} to either $V^{0,1}$ or $V^{0,2}$ (recall that these were defined in Proposition 4.6 or (4.15)) with an initial guess of zero and an error tolerance of $\beta_0 = h^r + k^\nu$.

Our algorithm is of course not well defined and in fact will not obtain the accuracy claimed unless we make some special choices for the starting guesses required by the process P. To be able to do this for the various schemes, we introduce some specific examples of the operators discussed in (5.9) as follows:

$$Z_{n+1}^{(0)}(U) = 0 \quad \text{for } 0 \leq n < M,$$

$$Z_{n+1}^{(1)}(U) = U^n \quad \text{for } 0 \leq n < M,$$

$$Z_{n+1}^{(2)}(U) = 2U^n - U^{n-1} \quad \text{for } 1 \leq n < M,$$

$$Z_{n+1}^{(3)}(U) = 3U^n - 3U^{n-1} + U^{n-2} \quad \text{for } 2 \leq n < M,$$

$$Z_{n+1}^{(4)}(U) = 4U^n - 6U^{n-1} + 4U^{n-2} - U^{n-3} \quad \text{for } 3 \leq n < M,$$

$$Z_{n+1}^{(5)}(U) = 5U^n - 10U^{n-1} + 10U^{n-2} - 5U^{n-3} + U^{n-4} \quad \text{for } 4 \leq n < M;$$

here, for the moment, U represents an arbitrary sequence $U = \{U^j\}_{j=0}^n \subset L^2(\Omega)$, where $0 \leq n < M$. We note that $U^n - Z_{n+1}^{(i)}(U)$, where $0 \leq n < M$ and $2 \leq i \leq \min(n+1, 5)$, is expressible as a sum of terms of the form $U^j - U^{j-1}$, for $n-i+2 \leq j \leq n$.

We can now state this not quite best possible algorithm in its entirety:

Algorithm (1). Use P with the preconditioning operator PQ to

(1) generate an approximation U^0 to either $V^{0,1}$ or $V^{0,2}$ using zero as an initial guess and a tolerance $\beta_0 = h^r + k^v$;

(2) generate an approximation U^{n+1} to the (true) solution of (5.8) using $Z_{n+1}^{(1)}(U) = U^n$ as an initial guess and a tolerance $\beta_{n+1} = k^v$ in the range $1 \leq n+1 \leq v-1$;

(3) generate an approximation U^{n+1} to the (true) solution of (5.8) using $Z_{n+1}^{(v)}(U)$ as an initial guess and a tolerance $\beta_{n+1} = k$, in the range $v \leq n+1 \leq M$.

Again we note that, since we are using a tolerance of $\beta_{n+1} = k$ for $v \leq n+1 \leq M$, we need on the order of $\log(M) = \log(\tau/k)$ iterations per time step, in general.

We use the techniques of Section IV to study this process via Eq. (5.10). We first observe the following, where $1 \leq n+1 \leq M$:

$$\begin{aligned} (5.11) \quad & |(\tilde{Q}_{n+1}(U^{n+1} - \bar{U}^{n+1}), E^{n+1})| \leq C \|Q_{n+1}^{1/2} E^{n+1}\| \|PQ^{1/2}(U^{n+1} - \bar{U}^{n+1})\| \\ & \leq C \|Q_{n+1}^{1/2} E^{n+1}\| \beta_{n+1} \|Q_{n+1}^{1/2}(U^{n+1} - Z_{n+1}^{(i)}(U))\| \\ & \leq C \beta_{n+1} \|Q_{n+1}^{1/2} E^{n+1}\| (\|Q_{n+1}^{1/2}(E^{n+1} - Z_{n+1}^{(i)}(E))\| \\ & \quad + \|Q_{n+1}^{1/2}(W^{n+1} - Z_{n+1}^{(i)}(W))\|), \end{aligned}$$

$$\begin{aligned} (5.12) \quad & \|Q_{n+1}^{1/2}(E^{n+1} - Z_{n+1}^{(i)}(E))\| \leq C \sum_{j=n-i+2}^{n+1} \|Q_j^{1/2}(E^j - E^{j-1})\| \\ & \leq C \sum_{j=n-i+1}^{n+1} \|Q_j^{1/2} E^j\|, \end{aligned}$$

where $i = 1$ or ν depending on n . If we have Condition B_h and a suitable ν , then

$$\begin{aligned}
 & \|Q_{n+1}^{1/2}(W^{n+1} - Z_{n+1}^{(i)}(W))\| \\
 & \leq C\|W^{n+1} - Z_{n+1}^{(i)}(W)\| + Ck\|L_{n+1}(W^{n+1} - Z_{n+1}^{(i)}(W))\| \\
 (5.13) \quad & \leq Ck^i \left(\sup_{\substack{0 \leq s \leq t_{n+1} \\ 0 \leq j < i}} \|W^{(j)}(s)\| + \sup_{\substack{0 \leq s \leq t_{n+1} \\ 0 \leq j < i-1}} \|L_n(s)W^{(j)}(s)\| \right) \\
 & \leq Ck^i \|v\|_{2(i+1)},
 \end{aligned}$$

where again $i = 1$ or ν depending on n .

Thus, we can show the following:

THEOREM 5.3. *Suppose that Condition B_h holds, $v \in H^\mu$ ($\mu = \max(r + 2, 2\nu + 2)$) is sufficiently compatible on $\partial\Omega$, and k is sufficiently small. Then the sequence of approximations $\{U^n\}_{n=0}^M$ generated by Algorithm (1) satisfies the following:*

$$(5.14) \quad \|Q_N^{1/2}(U^N - W^N)\| \leq C(h^r + k^\nu)\|v\|_\mu \quad \text{for } 0 \leq N \leq M.$$

Proof. Let $0 \leq n \leq M$. We have via the propositions and lemmas of Section IV and (5.10) through (5.13), that

$$\begin{aligned}
 (5.15) \quad (Q_{n+1}E^{n+1}, E^{n+1}) & \leq (1 + Ck)(Q_nE^n, E^n) + Ck\|v\|_\mu^2(h^r + k^\nu)^2 \\
 & + Ck \sum_{j=\max(n-\nu+1, 0)}^{n+1} (Q_jE^j, E^j).
 \end{aligned}$$

Also, Proposition 4.6 shows that $\|Q_0^{1/2}E^0\| \leq C(k^\nu + h^r)\|v\|_\mu$. This inequality and (5.15) give our result.

Thus, we have optimal order errors for Algorithm (1).

As a preliminary to analyzing an algorithm that requires only finitely many iterations per time step on the average, we prove some results for the following situation.

Suppose that a set of approximations $\{U^j\}_{j=n-i+1}^n$ to the functions $\{W^j\}_{j=n-i+1}^n$ are given, where $1 \leq i \leq \nu + 1$ and $i - 1 \leq n < M$. Use the process P (with the preconditioning operator PQ) to generate an approximation U^{n+1} to the (true) solution \bar{U}^{n+1} of (5.8) using $Z_{n+1}^{(i)}(U)$ as an initial guess and a tolerance $0 < \beta_{n+1} < 1$.

By the analysis done so far, we already know the following.

PROPOSITION 5.4. *Suppose we have Condition B_h and $v \in \bar{H}^\mu$ ($\bar{\mu} = \max(r + 2, 2\nu + 4)$) is sufficiently compatible on $\partial\Omega$. Then, for any $0 < \epsilon < 1$, we have that*

$$\begin{aligned}
 (5.16) \quad (Q_{n+1}E^{n+1}, E^{n+1}) & \leq (1 + Ck)(Q_nE^n, E^n) - \delta((Q_n - P_n)E^n, E^n) \\
 & + Ck((k^\nu + h^r)^2 + \epsilon\beta_{n+1}^2k^{2i-2})\|v\|_\mu^2 \\
 & + \frac{C}{\epsilon}k(Q_{n+1}E^{n+1}, E^{n+1}) \\
 & + \epsilon \frac{\beta_{n+1}^2}{2k} \sum_{j=n-i+2}^{n+1} \|Q_j^{1/2}(E^j - E^{j-1})\|^2.
 \end{aligned}$$

The above result motivates an investigation of the *differences* of the errors. Recalling the form of the equation satisfied by the errors (that is, (5.10)), we study the following

PROPOSITION 5.5. *Suppose that $QX = PV + F$, where Q and P are selfadjoint operators on S_h . Then we have that*

$$(5.17) \quad \begin{aligned} & ((Q + P)(X - V), X - V) + ((Q - P)X, X) \\ & = ((Q - P)V, V) + 2(F, X - V). \end{aligned}$$

Suppose further that Q and P satisfy (4.2) (iii). Then

$$\delta(Q(X - V), X - V) + ((Q - P)X, X) \leq ((Q - P)V, V) + 2(F, X - V).$$

We now apply Proposition 5.5 to (5.10). If we can enforce a certain important condition, namely that $\delta > 0$ for our polynomials $P(x)$ and $Q(x)$ (recall (3.1)), we can obtain an estimate that will allow us to analyze the last term in (5.16).

PROPOSITION 5.6. *Suppose we have Condition B_h , $v \in H^{\bar{\mu}}$ ($\bar{\mu} = \max(r + 2, 2v + 4)$) is sufficiently compatible on $\partial\Omega$ and $\delta > 0$. Then there are constants $\beta^* > 0$ and $C_\delta > 0$ so that if $\beta_{n+1} \leq \beta^*$, we have that*

$$(5.18) \quad \begin{aligned} & C_\delta \|Q_{n+1}^{1/2}(E^{n+1} - E^n)\|^2 + ((Q_{n+1} - P_{n+1})E^{n+1}, E^{n+1}) \\ & \leq (1 + Ck)((Q_n - P_n)E^n, E^n) + Ck^2(Q_{n+1}E^{n+1}, E^{n+1}) \\ & \quad + Ck^2(Q_nE^n, E^n) + Ck^2\{(h^r + k^v)^2 + \beta_{n+1}^2 k^{2i-2}\} \|v\|_\mu^2 \\ & \quad + C\beta_{n+1}^2 \sum_{j=n-i+2}^n \|Q_j^{1/2}(E^j - E^{j-1})\|^2. \end{aligned}$$

Proof. Proposition 5.5, applied to (5.10) in conjunction with the propositions and lemmas of Section IV, suitable modifications of (5.10) through (5.13) and different uses of the arithmetic-geometric mean inequality, yields (5.18) with $(1 + Ck)$ replaced by 1 and with the following extra term on the right-hand side:

$$(5.19) \quad |(((Q_{n+1} - P_{n+1}) - (Q_n - P_n))E_n, E_n)|.$$

Note that $Q(x) - P(x) = x + O(x^2)$ by the accuracy condition (3.1) (iii), and $Q(x) - P(x) > 0$ for $x > 0$. If we redefine $R(x) = x + |q_2 - p_2|x^2$, then $C_1R(x) \leq Q(x) - P(x) \leq C_2R(x)$ for $0 < x < \infty$. Thus, under Condition B_h , the techniques of Section IV show that

$$(5.20) \quad \begin{aligned} & |(((Q_{n+1} - P_{n+1}) - (Q_n - P_n))E^n, E^n)| \\ & \leq Ck^2(L_nE^n, E^n) + Ck^3|q_2 - p_2| \|L_nE^n\|^2 \\ & \leq Ck(R(kL_n)E^n, E^n) \leq Ck((Q_n - P_n)E^n, E^n). \end{aligned}$$

We could now combine (5.16) and (5.18) with suitable choices for the parameter i . We would then essentially find that $\|Q_N^{1/2}E^N\|$ is bounded by terms that are

$O(h^r + k^\nu)$, terms that measure the initial error in the $Q_0^{1/2}$ -norm, and terms that measure the initial error in the $k^{-1/2}(Q_0 - P_0)^{1/2}$ -norm. The projection we have chosen for the initial data (as given in Proposition 4.6) is defined so that it is computable by the **P** process and so that it leads to an initial error that is good in the $Q_0^{1/2}$ -norm. Unfortunately, it does not necessarily lead to one that is good in the $k^{-1/2}(Q_0 - P_0)^{1/2}$ -norm. We could now proceed in two ways. One approach is to let U^0 be $P_1 v$ or look for another special approximation which is good in all the required norms. But, since the process **P** would probably not be useful in generating such an approximation (the spectrum of ${}^P Q$ does not bear the correct relationship to the spectrum of $L_n(0)$), a special process would be needed to generate only U^0 . Since we would prefer to avoid this situation, we are led to using (5.18) in some other way. After all, it was only the direct use of (5.18) that gave this apparent problem.

We have the following result which combines (5.16) and a variant of (5.18); the latter uses multiplication by the time variable to avoid potential problems at time zero.

PROPOSITION 5.7. *Suppose that Condition B_n holds, $v \in H^{\bar{\mu}}$ ($\bar{\mu} = \max(r + 2, 2\nu + 4)$) is sufficiently compatible on $\partial\Omega$ and $\delta > 0$. Then we have the following for $0 < \epsilon < 1$, provided that $\beta_{n+1} \leq \min(\epsilon^{1/2}, t_{n+1}^{1/2})$:*

$$\begin{aligned}
 & (Q_{n+1} E^{n+1}, E^{n+1}) + C \frac{t_{n+1}}{k} \|Q_{n+1}^{1/2} (E^{n+1} - E^n)\|^2 \\
 & + C \left[\frac{t_{n+1}}{k} ((Q_{n+1} - P_{n+1}) E^{n+1}, E^{n+1}) - \frac{t_n}{k} ((Q_n - P_n) E^n, E^n) \right] \\
 (5.21) \quad & \leq (1 + Ck)(Q^n E^n, E^n) - (\delta/2)(Q_n - P_n) E^n, E^n \\
 & + Ck(h^{2r} + k^{2\nu} + \beta_{n+1}^2 k^{2i-2}) \|v\|_{\bar{\mu}}^2 + \frac{C_1}{\epsilon} k(Q_{n+1} E^{n+1}, E^{n+1}) \\
 & + \epsilon \frac{t_{n+1}}{k} \sum_{j=n-i+2}^n \|Q_j^{1/2} (E^j - E^{j-1})\|^2.
 \end{aligned}$$

Proof. It is a rather straightforward computation using (5.16) and (5.18) to obtain (5.21) with $\delta/2$ replaced by δ and a term

$$C\epsilon_1((Q_n - P_n)E^n, E^n)$$

on the right-hand side where $\epsilon_1 > 0$ is arbitrary. This gives the result.

We can now define and state results for our final algorithm:

Algorithm (2). Use **P** with the preconditioning operator ${}^P Q$ to

- (1) generate an approximation U^0 to either $V^{0,1}$ or $V^{0,2}$ using zero as an initial guess and a tolerance $\beta_0 = h^r + k^\nu$;
- (2) generate an approximation U^{n+1} to the (true) solution of (5.8) using $Z_{n+1}^{(n+1)}(U)$ as an initial guess and a tolerance $\beta_{n+1} \leq \min(k^{\nu-n-1}, \tilde{\beta})$ in the range $1 \leq n + 1 \leq \nu$, where $\tilde{\beta} > 0$ is small;
- (3) generate an approximation U^{n+1} to the (true) solution of (5.8) using $Z_{n+1}^{(\nu+1)}(U)$ as an initial guess and a tolerance $\beta_{n+1} \leq \min(\tilde{\beta}, t_{n+1}^{1/2})$ in the range $\nu + 1 \leq n + 1 \leq M$, where $\tilde{\beta} > 0$ is small.

We note that the important difference between this algorithm and Algorithm (1) is that, by our assumptions on the process P , we only have to iterate a fixed number of times at each time step, on the average. Thus, we are demanding fewer iterations, but as we will soon see, we get the same convergence rates.

THEOREM 5.8. *Suppose that Condition B_h holds, $v \in H^{\bar{\mu}}$ ($\bar{\mu} = \max(r + 2, 2\nu + 4)$) is sufficiently compatible on $\partial\Omega$, $\delta > 0$, and k sufficiently small. Then there is a (computable) $\tilde{\beta} > 0$ so that the approximations $\{U^n\}_{n=0}^M$ generated by Algorithm (2) satisfy the following for $0 \leq N \leq M$:*

$$\begin{aligned} \|Q_N^{1/2}(W^N - U^N)\| &\leq C(k^\nu + h^r)\|v\|_{\bar{\mu}}, \\ \|W^N - U^N\|_I &\leq Ct_N^{-1/2}(k^\nu + h^r)\|v\|_{\bar{\mu}} \quad \text{if } N > 0, \\ \|u(t_N) - U^N\| &\leq C(k^\nu + h^r)\|v\|_{\bar{\mu}}. \end{aligned}$$

Note that we get a superconvergence result for $W^N - U^N$ in the $\|\cdot\|_I$ -norm for times bounded away from zero.

Proof. The proof is almost immediate from (5.21). We first see that

$$(5.22) \quad \begin{aligned} (Q_{n+1}E^{n+1}, E^{n+1}) + \|Q_{n+1}^{1/2}(E^{n+1} - E^n)\|^2 + C((Q_{n+1} - P_{n+1})E^{n+1}, E^{n+1}) \\ \leq C(h^{2r} + k^{2\nu})\|v\|_{\bar{\mu}}^2 \end{aligned}$$

for $1 \leq n + 1 \leq \nu$, where (5.16) and (5.18) have been used.

If $\nu + 1 \leq N \leq M$, we set $i = \nu + 1$ in (5.21), multiply the inequality by $e^{-Ct_{n+1}/\epsilon}(1 - C_1k/\epsilon)^{-1}$, and sum over the range $\nu \leq n \leq N - 1$. Then, by choosing $\epsilon > 0$ sufficiently small and using (5.22), we obtain the first result. The third result now easily follows.

The remaining result is obtained by noting that if $\varphi \in S_h$, $0 \leq n \leq M$, and $R(x) = x + |q_2 - p_2|x^2$, then

$$\|\varphi\|_I^2 \leq C(L_n\varphi, \varphi) \leq \frac{C}{k}(R(kL_n)\varphi, \varphi) \leq \frac{C}{k}((Q_n - P_n)\varphi, \varphi).$$

Thus, Algorithm (2) generates accurate approximations to the solution u of (1.1) if the polynomial pair is such that $\delta > 0$, implying that dissipation effects are present. As was noted before, similar results can be proven for stable polynomial pairs (for which $\delta = 0$) if dissipation is introduced by requiring that $k \leq Ch^2$.

VI. Computational Considerations. We conclude this paper with a few remarks concerning the computational aspects of Algorithm (2) for quadratic $P(x)$ and $Q(x)$. Suppose that the functions $\{\varphi_i\}_{i=1}^J$ form a basis for S_h . Moreover, suppose that these functions have been chosen so that linear systems on \mathbf{R}^J involving the matrices

$$\begin{aligned} A_m &= [(L_m\varphi_i, \varphi_j)]_{i,j=1}^J, \quad A_m^{(1)} = [(L_m^{(1)}\varphi_i, \varphi_j)]_{i,j=1}^J, \quad 0 \leq m \leq M, \\ G &= [(\varphi_i, \varphi_j)]_{i,j=1}^J, \end{aligned}$$

have acceptable computational properties. We now wish to examine Algorithm (2) in this context.

We begin this discussion by making some definitions and identifications. Let $0 \leq m \leq M$. If $\varphi = \sum \xi_i \varphi_i \in S_h$, then

$$(6.1) \quad G\xi = [(\varphi, \varphi_j)]_{j=1}^J, \quad L_m \varphi = \sum_{j=1}^J [G^{-1} A_m \xi]_j \varphi_j.$$

Thus if $\bar{U} = \sum \zeta_j \varphi_j$ satisfies $\tilde{Q}_m \bar{U} = \varphi$, then

$$(6.2) \quad (I + q_1 k G^{-1} A_m + q_2 k^2 (G^{-1} A_m G^{-1} A_m - G^{-1} A_m^{(1)})) \xi = \xi,$$

or, equivalently,

$$(6.3) \quad B_m \xi \equiv (G + q_1 k A_m + q_2 k^2 A_m G^{-1} A_m - q_2 k^2 A_m^{(1)}) \xi = G \xi = [(\varphi, \varphi_j)]_{j=1}^J.$$

We note that (6.3) involves a symmetric, positive definite matrix B_m . We now let ${}^P Q$ be one of the polynomials in kL_0 discussed in Section V. Let ${}^P Q(x) = 1 + \tilde{q}_1 x + \tilde{q}_2 x^2$ define its coefficients, and let

$${}^P B = G + \tilde{q}_1 k A_0 + \tilde{q}_2 k^2 A_0 G^{-1} A_0,$$

which is also a symmetric, positive definite matrix on \mathbf{R}^J . Finally, rewriting (5.6) gives us the following:

$$(6.4) \quad C_1 \langle {}^P B \eta, \eta \rangle \leq \langle B_m \eta, \eta \rangle \leq C_2 \langle {}^P B \eta, \eta \rangle, \quad \eta \in \mathbf{R}^J,$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product on \mathbf{R}^J .

Thus, we see that S_h with the $L^2(\Omega)$ -inner product and \mathbf{R}^J with the $\langle G \cdot, \cdot \rangle$ -inner product are unitarily equivalent under the identification of φ_i with the i th canonical basis function in \mathbf{R}^J . This implies that any process \mathbf{P} on S_h that solves systems involving Q_m , using ${}^P Q$ as a preconditioner, is formally equivalent to a process \mathbf{P}^J on \mathbf{R}^J that solves systems involving $G^{-1} B_m$, using $G^{-1} {}^P B$ as a preconditioner. Since a process will usually be given on \mathbf{R}^J , in practice, this identification defines the corresponding process \mathbf{P} on S_h .

How do we compute the coefficients $\{\zeta_i^n\}_{1 \leq i \leq J, 0 \leq n \leq M}$ for the basis expansions of the functions $\{U^n\}_{0 \leq n \leq M} = \{\sum \zeta_i^n \varphi_i\}$ defined by Algorithm (2)? The linear system used to compute U^0 has $\varphi = Pf$ as a right-hand side where $f \in L^2(\Omega)$. If $\varphi = \sum \xi_i^f \varphi_i$, then

$$(6.5) \quad G \xi^f = [(Pf, \varphi_j)]_{j=1}^J = [(f, \varphi_j)]_{j=1}^J.$$

Thus, $G \xi^f$ can be calculated by taking inner products with f , and ξ^f can be found by solving a system involving G . The linear system used to compute U^{n+1} for some $n \geq 0$ has $\tilde{P}_n U^n = \sum \xi_i^n \varphi_i$ as a right-hand side. If we know ζ^n (the coefficients for U^n), then

$$(6.6) \quad G \xi^n = (G + p_1 k A_n + p_2 k^2 (A_n G^{-1} A_n - A_n^{(1)})) \zeta^n,$$

so that $G \xi^n$ can be calculated by solving one system involving G , and ζ^n can be calculated by solving two. Thus, we can compute the right-hand sides of the various systems. If we know $\{\zeta^j\}_{j=0}^n$, the initial guess for the iteration procedure is easy to calculate. Then, to find ζ^{n+1} via the process \mathbf{P}^J , we may have to evaluate $G^{-1} B_{n+1}$ and

$G^{-1}PB$, solve systems involving $G^{-1}PB$, and do various Hilbert space operations in \mathbf{R}^J with the $\langle \cdot, \cdot \rangle$ -inner product. Evaluating $G^{-1}B_{n+1}$ or $G^{-1}PB$ is straightforward. (However, note that calculating

$$\langle G(G^{-1}B_{n+1})\eta_1, \eta_2 \rangle = \langle B_{n+1}\eta_1, \eta_2 \rangle \quad \text{for } \eta_1, \eta_2 \in \mathbf{R}^J$$

only means solving one system involving G .) If we have to solve the system $G^{-1}PB\eta = \xi$ in \mathbf{R}^J , we observe that this is equivalent to solving $^PB\eta = G\xi$. (This becomes convenient in many situations as $G\xi$ may be easier to obtain than ξ . This was the case in (6.5) and (6.6).) We must now solve systems that involve PB . If $^PQ(x) = (1 + \lambda x)^2$ for some $\lambda > 0$ for instance, then

$$^PB = (G + \lambda kA_0)G^{-1}(G + \lambda kA_0).$$

Thus, solving $^PB\eta = G\xi$ in this case means solving two systems that involve the same matrix $(G + \lambda kA_0)$ and evaluating G once. If $^PQ(x)$ is not a perfect square, other techniques could be used to efficiently solve systems involving PB .

Regarding the attainment of certain tolerances by the process P^J , we follow a discussion in [4] and note that if $\zeta^{n+1,\alpha}$ gives the coefficients of the α th iterate $U^{n+1,\alpha} = \sum \zeta_i^{n+1,\alpha} \varphi_i$ of the preconditioned conjugate gradient algorithm (where $U^{n+1,0}$ is the initial guess), and $\hat{\zeta}^{n+1}$ gives the coefficients of the true solution $\bar{U}^{n+1} = \sum \hat{\zeta}_i^{n+1} \varphi_i$, the quantity $\langle ^PB^{-1}B_{n+1}(\zeta^{n+1,\alpha} - \hat{\zeta}^{n+1}), B_{n+1}(\zeta^{n+1,\alpha} - \hat{\zeta}^{n+1}) \rangle$ is calculated. Thus, since

$$\begin{aligned} & \| ^PQ^{1/2}(U^{n+1,\alpha} - \bar{U}^{n+1}) \|^2 / \| ^PQ^{1/2}(U^{n+1,0} - \bar{U}^{n+1}) \|^2 \\ &= \langle ^PB(\zeta^{n+1,\alpha} - \hat{\zeta}^{n+1}), \zeta^{n+1,\alpha} - \hat{\zeta}^{n+1} \rangle / \langle ^PB(\zeta^{n+1,0} - \hat{\zeta}^{n+1}), \zeta^{n+1,0} - \hat{\zeta}^{n+1} \rangle \\ &\leq \alpha \langle ^PB^{-1}B_{n+1}(\zeta^{n+1,\alpha} - \hat{\zeta}^{n+1}), B_{n+1}(\zeta^{n+1,\alpha} - \hat{\zeta}^{n+1}) \rangle / \langle ^PB^{-1}B_{n+1}(\zeta^{n+1,0} - \hat{\zeta}^{n+1}), B_{n+1}(\zeta^{n+1,0} - \hat{\zeta}^{n+1}) \rangle, \end{aligned}$$

where the latter quantity is known up to the constant, we have an estimate for the error reduction given by the preconditioned conjugate gradient process as the iterations are being carried out.

The particular choice of the iterative process will determine which of the above considerations is relevant in the implementation of Algorithm (2).

Department of Mathematics
Cornell University
Ithaca, New York 14853

Department of Mathematics
University of Chicago
Chicago, Illinois 60637

1. O. AXELSSON, *On Preconditioning and Convergence Acceleration in Sparse Matrix Problems*, CERN (European Organization for Nuclear Research), Geneva, 1974.
2. G. A. BAKER, J. H. BRAMBLE & V. THOMÉE, "Single step Galerkin approximations for parabolic problems," *Math. Comp.*, v. 31. 1977, pp. 818-847.
3. J. DOUGLAS, JR. & T. DUPONT, "Alternating direction methods on rectangles," *Numerical Solution of Partial Differential Equations—II* (B. Hubbard, Ed.), Academic Press, New York, 1971.

4. J. DOUGLAS, JR., T. DUPONT & R. EWING, "Incomplete iteration for time-stepping a Galerkin method for a quasilinear parabolic problem," *SIAM J. Numer. Anal.*, v. 16, 1979, pp. 503-522.
5. A. FRIEDMAN, *Partial Differential Equations*, Krieger, Huntington, New York, 1976.
6. J. L. LIONS & E. MAGENES, *Nonhomogeneous Boundary Value Problems and Applications*, Vol. II, Springer-Verlag, New York, 1973.
7. N. NASSIF & J. DESCLOUX, "Stability study for time-dependent linear parabolic equations and its application to Hermitian methods," *Topics in Numerical Analysis III* (J. Miller, Ed.), Academic Press, New York, 1977.
8. P. H. SAMMON, *Convergence Estimates for Semidiscrete Parabolic Equation Approximations*, Mathematics Research Center Technical Survey Report No. 2053, 1980.
9. RICHARD S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1962.