MATHEMATICS OF COMPUTATION
VOLUME 52, NUMBER 186
APRIL 1989, PAGES 339-367

Incomplete Iterations in Multistep Backward Difference Methods for Parabolic Problems with Smooth and Nonsmooth Data*

By James H. Bramble, Joseph E. Pasciak, Peter H. Sammon, and Vidar Thomée

Dedicated to Professor Eugene Isaacson on the occasion of his seventieth birthday

Abstract. Backward difference methods for the discretization of parabolic boundary value problems are considered in this paper. In particular, we analyze the case when the backward difference equations are only solved 'approximately' by a preconditioned iteration. We provide an analysis which shows that these methods remain stable and accurate if a suitable number of iterations (often independent of the spatial discretization and time step size) are used. Results are provided for the smooth as well as nonsmooth initial data cases. Finally, the results of numerical experiments illustrating the algorithms' performance on model problems are given.

1. Introduction. In this paper, we shall study implicit multistep backward difference methods for linear parabolic equations. In particular, we shall focus on generalizations which involve the application of iterative procedures to the resulting algebraic systems. We will provide results which justify the use of incomplete iteration in a way that will not affect the error estimates for the discretization. Thus, the time stepping scheme with incomplete iteration will achieve the same order of convergence as the original scheme (solving the implicit equations exactly at each time step). These results will be given for problems with smooth as well as nonsmooth solutions.

In general, we only assume that the iterative procedure leads to a reduction in an appropriate energy norm. Typical examples can be developed by applying the preconditioned conjugate gradient method with an appropriate preconditioner. Some examples of preconditioners and their analyses can be found in [3]–[9], [13], [17], [25] and the references cited in [13].

Our results show that the error in the numerical methods has the same asymptotic behavior even when an appropriate number of iterations (often only a fixed number independent of the spatial mesh and time step size) are used at each time

Received May 31, 1988.

1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 65N30; Secondary 65F10.

*This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS84-05352 and under the Air Force Office of Scientific Research, Contract No. ISSA86-0026 and by the U.S. Army Research Office through the Mathematical Science Institute, Cornell University.

step. The procedure studied thus reduces the work involved in the actual computation in a significant way. This is illustrated both in our theoretical and computational results.

In Section 2 we review the error estimates for stable backward difference methods in the case that the solution is smooth, and under the assumption that the difference equations are solved exactly at each time step. Since our emphasis is on the discretization in time, we shall start by considering an abstract evolution equation in a Hilbert space setting, where a selfadjoint positive definite operator plays the role of the elliptic operator in the parabolic equation. We shall then apply the analysis to the case of a partial differential equation in space and time.

Sections 3 and 4 provide an analysis for the backward difference algorithms using incomplete iteration. In Section 3 we consider algorithms applied to a partial differential equation in space and time with a smooth solution. In Section 4 we study the case of a homogeneous equation with nonsmooth initial data.

The proofs of our results given in Sections 3 and 4 are based on certain *a priori* estimates for the solution of the backward difference equation (without iteration). These estimates are proved in Section 5.

In Section 6 we provide the results of numerical experiments illustrating the theory developed in this paper. We also provide additional details concerning preconditioning and the starting procedure.

Error estimates for semidiscrete in space and completely discrete single step methods applied to parabolic problems for both smooth and nonsmooth solutions have been derived by many authors, cf. Thomée [24] and references therein. Multistep methods have been studied similarly by Zlámal [26] and Crouzeix and Raviart [14], for smooth, and by Le Roux [21] for nonsmooth solutions.

The idea of incomplete iterations was first analyzed for parabolic problems in Douglas, Dupont and Ewing [16] and Bramble and Sammon [10] (cf. also Bramble [2], Keeling [20]), in the context of single step schemes and under the assumption that the exact solution is smooth.

2. The Basic Backward Difference Method. In this preliminary section, we shall give error and stability estimates for the basic backward difference approximation to parabolic problems with smooth solutions. We first study the abstract parabolic equation in Hilbert space and then turn to the concrete situation of a parabolic partial differential equation in space and time. In this section, we assume that the equations resulting from the backward difference time discretization will be solved exactly at each time level.

We start with the abstract parabolic equation on a separable Hilbert space H given by

$$\begin{aligned} u_t + \mathscr{A}u &= f \quad \text{for } 0 \le t \le T, \\ u(0) &= v, \end{aligned} \tag{2.1}$$

where $\mathscr{A}$ is a selfadjoint, positive definite, not necessarily bounded operator on H with dense domain of definition $\mathscr{D}(\mathscr{A})$, and f is a function of t with values in H.

We shall study the numerical approximation to (2.1) by a q-step backward difference method. Let k denote the time step size and $t_n = nk$. For given starting

values $\{U^i \in H, i = 0, \dots, q-1\}$, define the sequence of functions $\{U^i \in H\}$ by the difference equations

$$\sum_{j=1}^{q} \frac{k^{j-1}}{j} \bar{\partial}_t^j U^n + \mathscr{A} U^n = f^n \quad \text{for } n \geq q, \tag{2.2}$$

where $\bar{\partial}_t U^n = (U^n - U^{n-1})/k$ and $f^n = f(t_n)$. It is a straightforward consequence of the Spectral Theorem that the sequence $\{U^n\}$ is well defined and can be computed by a marching algorithm. U^n is an approximation to $u^n = u(t_n)$ when appropriate starting values are used. Equation (2.2) can be rearranged in the form

$$L_k U^n \equiv k^{-1} \sum_{j=0}^{q} \alpha_j U^{n-j} + \mathscr{A} U^n = f^n \quad \text{for } n \geq q. \tag{2.3}$$

In order for U^n to be an accurate approximation to u^n, appropriate values of U^j, $j = 0, \dots, q-1$, must be defined by a separate starting procedure. For accuracy, these values of U^i should approximate u^i for $i = 0, \dots, q-1$ to order q.

Note that method (2.2) is accurate of order q. This can be seen as follows. The Newton backward difference formula (cf. [19]) is given by

$$\begin{aligned} u(t) = u^n &+ (t - t_n)\bar{\partial}_t u^n + \frac{(t-t_n)(t-t_{n-1})}{2!}\bar{\partial}_t^2 u^n \\ &+ \dots + \frac{(t-t_n)\dots(t-t_{n-q+1})}{q!}\bar{\partial}_t^q u^n \\ &+ \frac{(t-t_n)\dots(t-t_{n-q})}{(q+1)!} u^{(q+1)}(\zeta). \end{aligned} \tag{2.4}$$

Here, $u^{(q+1)}(\zeta)$ is the $q+1$'st derivative of u evaluated somewhere in the interval $[t_{n-q}, t_n]$. Applying (2.4) to a polynomial P of degree q and differentiating shows that

$$\sum_{j=1}^{q} \frac{k^{j-1}}{j} \bar{\partial}_t^j P^n = \frac{\partial P(t_n)}{\partial t},$$

and hence it follows immediately from Taylor's Theorem that (2.2) is accurate of order q.

For $q = 1$, (2.2) reduces to the backward Euler method

$$\frac{U^n - U^{n-1}}{k} + \mathscr{A} U^n = f^n \quad \text{for } n \geq 1, \tag{2.5}$$

and only the starting value $U^0 = v$ is needed. For $q = 2$, (2.2) takes the form

$$\left(\frac{3}{2}U^n - 2U^{n-1} + \frac{1}{2}U^{n-2}\right) \Big/ k + \mathscr{A} U^n = f^n \quad \text{for } n \geq 2.$$

In this case, natural choices for U^0 and U^1 are

$$U^0 = v$$

and

$$(U^1 - U^0)/k + \mathscr{A} U^1 = f^1.$$

Note that U^1 is determined by taking one step of the backward Euler method (2.5).

For $q > 2$, starting values can be generated by using the partial sums of the Taylor expansion of $u(t_j)$, i.e.,

$$U^j = \sum_{l=0}^{q-1} \frac{(jk)^l}{l!} D_t^l u(0) \quad \text{for } j = 0, \dots, q-1. \tag{2.6}$$

Here the function $D_t^l u(0)$ can be computed from the differential equation in terms of data, i.e.,

$$\begin{aligned} D_t u(0) &= f(0) - \mathscr{A} u(0), \\ D_t^2 u(0) &= f_t(0) - \mathscr{A}(f(0) - \mathscr{A} u(0)), \quad \text{etc.} \end{aligned} \tag{2.7}$$

This choice is only appropriate for the smooth data case. Starting values for the nonsmooth data case will be discussed later.

It is well known from the theory for the numerical solution of ordinary differential equations (cf., e.g., Gear [18, p. 214], Cryer [15]) that the backward difference method employed in (2.2) is $A(0)$-stable for $q \leq 6$. As a result of this stability, we can prove the following theorem.

THEOREM 2.1. *For $q \leq 6$, let U^n and u be the solution of* (2.2) *and* (2.1) *respectively. Assume that the starting values $\{U^j\}$ satisfy*

$$\left\| U^j - u^j \right\| \leq C(u) k^q \quad \text{for } j = 0, \dots, q-1.$$

Then, provided that u is sufficiently smooth,

$$\|U^n - u^n\| \leq C(u) k^q \quad \text{for } n \geq q \text{ and } t_n \leq T.$$

To prove the theorem, we shall use scales of spaces induced by the operator $\mathscr{A}$. Note that the powers of the operator $\mathscr{A}$ are well defined in terms of its spectral decomposition. We define $\dot{H}^s$ to be the domain of $\mathscr{A}^{s/2}$. Then $\dot{H}^s$ is a Hilbert space with norm given by

$$\|v\|_s = (\mathscr{A}^s v, v)^{1/2}.$$

A major ingredient in the proof of the above theorem, as well as those to be stated later, is the following fundamental *a priori* inequality.

LEMMA 1. *Let $q \leq 6$ and $p \geq 0$. Let $\{U^n\}$ be the solution of* (2.2) *and n be greater than or equal to q. Then*

$$\begin{aligned} t_n^p \|U^n\|^2 + k \sum_{j=q}^{n} t_j^p \left\|U^j\right\|_1^2 \leq{}& Ck \sum_{j=q}^{n} (t_j^p \left\|f^j\right\|_{-1}^2 + \left\|f^j\right\|_{-p-1}^2) \\ &+ C \sum_{j=0}^{q-1} (\left\|U^j\right\|_{-p}^2 + k^p \left\|U^j\right\|^2). \end{aligned}$$

This lemma will be proved in Section 5. Its generality was introduced for later use. We next prove Theorem 2.1, assuming the lemma.

Proof of Theorem 2.1. Let $e^n = u^n - U^n$. Then

$$L_k e^n = \tau^n, \tag{2.8}$$

where τ^n is the truncation error in the discretization of the time derivative, i.e.,

$$\tau^n = k^{-1} \sum_{j=0}^{q} \alpha_j u^{n-j} - u_t^n. \tag{2.9}$$

Applying Lemma 1 with $p = 0$ to $\{e^j\}$ gives

$$\|e^n\|^2 \le Ck \sum_{j=q}^{n} \|\tau^j\|_{-1}^2 + C \sum_{j=0}^{q-1} \|e^j\|^2. \tag{2.10}$$

By (2.4),

$$k^{-1} \sum_{j=0}^{q} \alpha_j P(t_{n-j}) - \frac{\partial P(t_n)}{\partial t} = 0$$

holds for polynomials P of degree less than or equal to q. Hence, Taylor's formula implies that

$$\tau^n = k^{-1} \sum_{j=0}^{q} \alpha_j \varsigma(t_{n-j}) - \frac{\partial \varsigma(t_n)}{\partial t} \tag{2.11}$$

where

$$\varsigma(t) = \frac{1}{q!} \int_{t_{n-q}}^{t} (t-s)^q D^{q+1} u(s)\, ds.$$

Hence, if u is smooth enough,

$$\|\tau^j\|_{-1} \le C(u) k^q,$$

and (2.10) implies the theorem.

We shall now show how Lemma 1 can be used to appraise the error in the numerical solution of a parabolic partial differential equation in space and time. The time stepping procedure will be applied to an equation which has first been discretized in the space variables.

We consider the initial boundary value problem

$$\begin{aligned} u_t + Au &= f \quad \text{in } \Omega,\ 0 \le t \le T, \\ u &= 0 \quad \text{on } \partial\Omega,\ 0 < t \le T, \\ u(\cdot, 0) &= v(\cdot) \quad \text{in } \Omega, \end{aligned} \tag{2.12}$$

where Ω is a bounded domain in R^d with smooth boundary and A is the second-order selfadjoint elliptic operator given by

$$Au = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i} \left(a_{ij}(x) \frac{\partial u}{\partial x_j} \right) + a_0(x) u.$$

Here we assume that the coefficients defining A are smooth, $a_0(x) \ge 0$ and $\{a_{ij}(x)\}$ is uniformly positive definite.

Let $H^s(\Omega)$ denote the usual Sobolev space of order s defined on Ω (cf. [22]). Let the scales of spaces $\{\dot{H}^s\}$ be defined as previously discussed with $H = L^2(\Omega)$ and $\mathscr{A}$ replaced by A. It was shown in [12] that for nonnegative integers s,

$$\dot{H}^s = \{\phi \in H^s(\Omega) | \Delta^j \phi = 0 \text{ on } \partial\Omega \text{ for } j < s/2\}.$$

In particular, $\dot{H}^1 = H_0^1(\Omega)$, the space of functions in $H^1(\Omega)$ whose trace vanishes in the appropriate sense on $\partial\Omega$.

Let r be an integer greater than one. Assume that we are given a family of finite-dimensional approximation spaces $S_h \subset H_0^1(\Omega)$ with the property

$$\inf_{\chi \in S_h} (\|v - \chi\|_{L^2(\Omega)} + h \|v - \chi\|_{H^1(\Omega)}) \le Ch^s \|v\|_{H^s(\Omega)} \quad \text{for } 1 \le s \le r.$$

Define the discrete operator $A_h\colon S_h \mapsto S_h$ by

$$(A_h v, \chi) = A(v, \chi) \quad \text{for all } \chi \in S_h,$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ and $A(\cdot,\cdot)$ is the bilinear form corresponding to A. Clearly, A_h is a symmetric positive definite operator on S_h.

Let P_h denote the $L^2(\Omega)$ orthogonal projection operator onto S_h and consider the semidiscrete problem

$$\begin{aligned} u_{h,t} + A_h u_h &= P_h f \quad \text{for } t \geq 0, \\ u_h(0) &= v_h. \end{aligned} \tag{2.13}$$

Equation (2.13) is of the same form as (2.1), and hence we can apply the time discretization method discussed earlier to define a fully discrete approximation to the solution u of (2.12). Thus, we define the sequence of functions $\{U^n \in S_h\}$ by replacing $\mathscr{A}$ by A_h in (2.2).

Remark 2.1. In terms of forms, (2.13) is equivalent to

$$\begin{aligned} (u_{h,t}, \chi) + A(u_h, \chi) &= (f, \chi) \quad \text{for all } \chi \in S_h \text{ and } t \geq 0, \\ u_h(0) &= v_h. \end{aligned}$$

Similarly, $\{U^n\}$ is the sequence of functions satisfying

$$\left(k^{-1}\sum_{j=0}^{q} \alpha_j U^{n-j}, \chi\right) + A(U^n, \chi) = (f, \chi) \quad \text{for all } \chi \in S_h \text{ and } n \geq q.$$

We will next prove the following theorem.

THEOREM 2.2. *For $q \leq 6$ let U^n be the solution of* (2.2) *with $\mathscr{A} = A_h$ and u be the solution of* (2.12). *Assume that the starting values $\{U^i\}$ satisfy*

$$\|U^i - u^i\| \leq C(u)(h^r + k^q) \quad \textit{for } i = 0, \ldots, q-1. \tag{2.14}$$

Then, provided that u is sufficiently smooth,

$$\|U^n - u^n\| \leq C(u)(h^r + k^q) \quad \textit{for } n \geq q \textit{ and } t_n \leq T. \tag{2.15}$$

Proof. For the purpose of proof, we introduce the Ritz projection $R_h\colon H_0^1(\Omega) \to S_h$ defined by

$$A(R_h V, \chi) = A(V, \chi) \quad \text{for all } \chi \in S_h$$

and write the error

$$U^i - u^i = (U^i - R_h u^i) + (R_h u^i - u^i) \equiv \theta^i + \rho^i \quad \text{for } i \geq 0. \tag{2.16}$$

By standard error estimates for the Ritz projection,

$$\|\rho^n\| \leq Ch^r \|u^n\|_{H^r(\Omega)} \leq C(u)h^r. \tag{2.17}$$

Thus it suffices to consider the remaining part θ^n, which is in S_h. From the definitions, it is easily checked that $A_h R_h = P_h A$, on the domain of A, hence θ^n satisfies

$$k^{-1}\sum_{j=0}^{q} \alpha_j \theta^{n-j} + A_h \theta^n = \sigma^n \equiv (P_h - R_h)u_t^n + R_h \tau^n \quad \text{for } n \geq q, \tag{2.18}$$

where τ^n is the truncation error given by (2.9).

By Lemma 1 applied (with $p = 0$) to θ^n,

$$\|\theta^n\|^2 \le Ck \sum_{j=q}^{n} \|\sigma^j\|_{-1,h}^2 + C \sum_{j=0}^{q-1} \|\theta^j\|^2,$$

where $\|\cdot\|_{-j,h}$ denotes the discrete norm defined by

$$\|\chi\|_{-j,h} = (A_h^{-j}\chi, \chi)^{1/2}.$$

Clearly, for sufficiently smooth u,

$$\|\sigma^j\|_{-1,h} \le C\|\sigma^j\| \le C(u)(h^r + k^q). \tag{2.19}$$

Note that by the triangle inequality, (2.14) and (2.17),

$$\|\theta^j\| \le C(u)(h^r + k^q) \quad \text{for } j = 0, \dots, q-1 \tag{2.20}$$

and hence

$$\|\theta^j\| \le C(u)(h^r + k^q). \tag{2.21}$$

The theorem follows combining (2.17) and (2.21).

An appropriate choice of starting values would be (cf. (2.6))

$$U^j = R_h \sum_{j=0}^{q-1} \frac{(jk)^l}{l!} D_t^l u(0) \quad \text{for } j = 0, \dots, q-1, \tag{2.22}$$

where the $D_t^l u(0)$ are computed from the differential equation in terms of data as in (2.7).

3. Incomplete Iteration. In Section 2 we considered algorithms which required the exact solution of the backward difference equations at each time step. In this section, we consider the extension of such algorithms to the case where the backward difference equations are only 'approximately' solved. We shall limit our discussion here to approximation of smooth solutions of (2.12) and consider the case of nonsmooth initial data in the following section. Moreover, we shall only consider the case where the backward difference is applied to the equation which has already been discretized in space (2.13). The incomplete iterative technique has important computational advantages in applications where efficient preconditioners are available.

As already indicated, the incomplete iteration backward difference algorithm is defined by only approximately solving the time step equations. Again, we are to define a sequence of functions $\{U^i\} \subset S_h$. Given $U^{n-1}, \dots, U^{n-q}$, we use an iterative process to approximate the solution $\bar{U}^n$ of

$$(\alpha_0 + kA)\bar{U}^n = kf^n - \sum_{j=1}^{q} \alpha_j U^{n-j}. \tag{3.1}$$

We assume that the iterative process uses a starting guess $U^{n,0}$ (which we are to provide) and gives rise to a sequence of iterates $U^{n,m}$ converging to the solution $\bar{U}^n$ of (3.1) as m tends to infinity. The incomplete iteration algorithm is then defined by setting $U^n = U^{n,M(n)}$ for some integer $M(n)$ which may vary with n and is to be specified.

In addition to the number $M(n)$, we have to define $U^{n,0}$ in order to make our procedure precise. For the purpose of accuracy, we shall need $U^{n,0}$ to be a $q+1$'st order approximation to u^n. The $q+1$'st order extrapolation approximating u^n in terms of $u^{n-1},\dots,u^{n-q-1}$ is defined from

$$u^n = -\sum_{l=1}^{q+1} \binom{q+1}{l} (-1)^l u^{n-l} + O(k^{q+1}).$$

Hence we define

$$U^{n,0} = -\sum_{l=1}^{q+1} \binom{q+1}{l} (-1)^l U^{n-l}. \tag{3.2}$$

Since (3.2) may be used only for $n \geq q+1$, we define U^q equal to $\bar{U}^q$.

Remark 3.1. As will be demonstrated by the theory, it is possible to choose $M(n)$ *a priori* so that the incomplete iteration scheme is stable and convergent. However, this choice of $M(n)$ involves *a priori* constants which are not explicitly available in practice. Nevertheless, numerical examples given in Section 6 indicate that the threshold values of $M(n)$ necessary for stability are rather low. These results suggest that the incomplete iteration technique can be used to develop robust time stepping algorithms.

To study the stability and convergence properties of the above method, we must make some additional assumptions on the iterative procedure. We assume that there exist positive constants c_0 and $\kappa < 1$ such that

$$|||U^{n,j} - \bar{U}^n||| \leq c_0 \kappa^j |||U^{n,0} - \bar{U}^n||| \quad \text{for } j = 1,2,\dots, \tag{3.3}$$

where

$$|||v||| \equiv (\|v\|^2 + k(A_h v, v))^{1/2} = (\|v\|^2 + kA(v,v))^{1/2}.$$

Estimates of the form of (3.3) are rather typical in the theory of preconditioned iterative methods. Values of κ are generally related to the condition number of the preconditioned system. As an example, the case of preconditioned conjugate gradient iteration will be discussed in Section 6.

The next theorem gives an error estimate for the incomplete iteration backward difference method described above, applied to the semidiscrete equation (2.13).

THEOREM 3.1. *Let $q \leq 6$ and u be the solution of* (2.12). *Further, let $U^q = \bar{U}^q$ and U^n for $n > q$ be defined by incomplete iteration for the solution of* (3.1) *as described above. Assume that the starting values $\{U^j\}$ have been chosen so that*

$$|||U^j - R_h u^j||| \leq C(u)(h^r + k^q) \quad \textit{for } j = 0,\dots,q-1. \tag{3.4}$$

Then, provided u is sufficiently smooth, there exists a positive constant δ (independent of u) such that if $M(n)$ is large enough so that

$$\kappa^{M(n)} \leq \delta t_n^{1/2}, \tag{3.5}$$

then

$$\|U^n - u^n\| \leq C(u)(h^r + k^q) \quad \textit{for } n \geq q \textit{ and } t_n \leq T.$$

Proof. The result of the theorem for $n = q$ is contained in Theorem 2.2. We thus only consider $n > q$. Let ε be a small positive constant which will be defined later in the proof. By the triangle inequality and (3.3),

$$|||U^n - \bar{U}^n||| \le c_0 \kappa^{M(n)} (|||U^{n,0} - U^n||| + |||U^n - \bar{U}^n|||).$$

We choose $\delta = \delta(\varepsilon)$ small enough so that

$$\frac{c_0 \kappa^{M(n)}}{1 - c_0 \kappa^{M(n)}} \le \varepsilon t_n^{1/2}. \tag{3.6}$$

Then

$$|||U^n - \bar{U}^n||| \le \varepsilon t_n^{1/2} |||U^n - U^{n,0}|||. \tag{3.7}$$

But (3.2) can be rewritten as

$$U^n - U^{n,0} = k^{q+1} \bar{\partial}_t^{q+1} U^n$$

and hence

$$|||\omega^n||| \le \varepsilon t_n^{1/2} k^q |||\bar{\partial}_t^{q+1} U^n||| \quad \text{for } n \ge q+1, \tag{3.8}$$

where $\omega^n \equiv (U^n - \bar{U}^n)/k$.

We now proceed as in the proof of Theorem 2.2 and decompose the error as in (2.16), i.e., $U^n - u^n = \theta^n + \rho^n$. Once again, ρ^n is bounded by (2.17) and we are left to estimate θ^n. Note that the sequence $\tilde{\theta}^n = \bar{U}^n - R_h u^n$, $\tilde{\theta}^{n-j} = \theta^{n-j}$ for $j = 1, \dots, n-q$ satisfies (2.18). Hence

$$L_k \theta^n = \sigma^n + (\alpha_0 + kA_h)\omega^n, \tag{3.9}$$

and thus Lemma 1 (with $p = 0$) yields

$$\|\theta^n\|^2 \le Ck \sum_{j=q}^{n} \|\sigma^j + (\alpha_0 + kA_h)\omega^j\|_{-1,h}^2 + \sum_{j=0}^{q-1} \|\theta^j\|^2.$$

By (2.19), we clearly have

$$\|\sigma^j + (\alpha_0 + kA_h)\omega^j\|_{-1,h} \le C(u)(h^r + k^q) + C|||\omega^j|||. \tag{3.10}$$

Thus, (2.20) and the fact that $\omega^q = 0$ give

$$\|\theta^n\|^2 \le C(u)(h^r + k^q)^2 + Ck \sum_{j=q+1}^{n} |||\omega^j|||^2. \tag{3.11}$$

From (3.8) and the triangle inequality,

$$|||\omega^n||| \le \varepsilon t_n^{1/2} k^q (|||\bar{\partial}_t^{q+1} \theta^n||| + |||R_h \bar{\partial}_t^{q+1} u^n|||).$$

Now R_h is bounded in the $H^1(\Omega)$ norm and clearly, the norm $|||\cdot|||$ is bounded by the $H^1(\Omega)$ norm and hence

$$|||R_h \bar{\partial}_t^{q+1} u^n||| \le \|\bar{\partial}_t^{q+1} u^n\|_1.$$

But $\bar{\partial}_t^{q+1}$ annihilates polynomials up to degree q, and hence by Taylor's formula

$$\|\bar{\partial}_t^{q+1} u^n\|_1 \le C \left\| \left(\bar{\partial}_t^{q+1} \int_{t_{n-q-1}}^{t} (t-s)^q D^{q+1} u(s)\, ds \right)_{t=t_n} \right\|_1 \le C(u).$$

Thus

$$|||\omega^n||| \le C\varepsilon t_n^{1/2} \sum_{j=0}^{q} |||\bar{\partial}_t \theta^{n-j}||| + C(u)k^q \quad \text{for } n > q. \tag{3.12}$$

Hence

$$\|\theta^n\|^2 \le C(u)(h^r + k^q)^2 + C\varepsilon^2 k \sum_{j=1}^{n} t_j |||\bar{\partial}_t \theta^j|||^2. \tag{3.13}$$

We now need an estimate for the last term in (3.13). We introduce the following estimate for the original time stepping scheme in the Hilbert space framework. In this estimate, the norm $|||\cdot|||$ is defined in terms of $\mathscr{A}$. Moreover, we define the norms $|||\cdot|||_{*,s}$ by

$$|||v|||_{*,s} = \left\| (I + k\mathscr{A})^{-1/2} \mathscr{A}^{s/2} v \right\|.$$

When $s = 0$, we denote the above norm by $|||\cdot|||_*$.

LEMMA 2. *Let $q \le 6$ and $p \ge 0$. Let $\{U^n\}$ be the solution of (2.2) and n be greater than or equal to q. Then*

$$k \sum_{j=q}^{n} t_j^p |||\bar{\partial}_t U^j|||^2 \le Ck \sum_{j=q}^{n} (t_j^p |||f^j|||_*^2 + |||f^j|||_{*,-p}^2) + C \sum_{j=0}^{q-1} (|||U^j|||_{*,-p+1}^2 + k^{p-1} |||U^j|||^2).$$

The proof of Lemma 2 will be given in Section 5. We complete the proof of the theorem assuming the lemma. Applying the lemma with $p = 1$ to θ^n satisfying (3.9) gives

$$k \sum_{j=q}^{n} t_j |||\bar{\partial}_t \theta^j|||^2 \le Ck \sum_{j=q}^{n} |||\sigma^j + (\alpha_0 + kA_h)\omega^j|||_*^2 + C \sum_{j=0}^{q-1} |||\theta^j|||^2.$$

By (2.19),

$$|||\sigma^j + (\alpha_0 + kA_h)\omega^j|||_* \le C(u)(h^r + k^q) + C|||\omega^j|||$$

and thus

$$k \sum_{j=q}^{n} t_j |||\bar{\partial}_t \theta^j|||^2 \le C(u)(h^r + k^q)^2 + Ck \sum_{j=q+1}^{n} |||\omega^j|||^2. \tag{3.14}$$

By (3.12),

$$\begin{aligned} k \sum_{j=q+1}^{n} |||\omega^j|||^2 &\le C\varepsilon^2 k \sum_{j=1}^{n} t_j |||\bar{\partial}_t \theta^j|||^2 + C(u)k^{2q} \\ &\le C\varepsilon^2 k \sum_{j=q}^{n} t_j |||\bar{\partial}_t \theta^j|||^2 + Ck^2 \sum_{j=1}^{q-1} |||\bar{\partial}_t \theta^j|||^2 + C(u)k^{2q}. \end{aligned} \tag{3.15}$$

By (3.4),

$$k|||\bar{\partial}_t \theta^j||| \le C(|||\theta^j||| + |||\theta^{j-1}|||) \le C(u)(h^r + k^q) \quad \text{for } j = 1, \dots, q-1. \tag{3.16}$$

Combining (3.14), (3.15) and (3.16) gives

$$k \sum_{j=q+1}^{n} |||\omega^j|||^2 \le C\varepsilon^2 k \sum_{j=q+1}^{n} |||\omega^j|||^2 + C(u)(h^r + k^q)^2.$$

Taking ε sufficiently small yields

$$k \sum_{j=q+1}^{n} |||\omega^j|||^2 \le C(u)(h^r + k^q)^2. \tag{3.17}$$

The theorem then follows combining (3.13), (3.14) and (3.17).

To satisfy (3.5), a larger number of preconditioned iterations must be taken in the earlier time steps. The next corollary (of the proof of Theorem 3.1) shows that it is possible to iterate with a fixed ($M(n)$ independent of n, k and h) number of preconditioned iterations if more accurate starting values are assumed.

COROLLARY 3.1. *Assume that the hypotheses of Theorem* 3.1 *hold with* (3.4) *replaced by*

$$|||U^j - R_h u^j||| \le C(u)k^{q+1/2} \quad \text{for } j = 0, \dots, q-1 \tag{3.18}$$

and (3.5) *replaced by* $M(n) \ge C$. *Then, provided that* u *is sufficiently smooth,*

$$\|U^n - u^n\| \le C(u)(h^r + k^q) \quad \text{for } n \ge q \text{ and } t_n \le T.$$

Proof. We follow the proof of Theorem 3.1, replacing $\varepsilon t_n^{1/2}$ by ε. Inequality (3.12) is replaced by

$$|||\omega^n||| \le C\varepsilon \sum_{j=0}^{q} |||\bar{\partial}_t \theta^{n-j}||| + C(u)k^q \quad \text{for } n > q.$$

Inequality (3.13) is replaced by

$$\|\theta^n\|^2 \le C(u)(h^r + k^q)^2 + C\varepsilon^2 k \sum_{j=1}^{n} |||\bar{\partial}_t \theta^j|||^2. \tag{3.19}$$

Applying Lemma 2 with $p = 0$ to θ^n gives

$$k \sum_{j=q}^{n} |||\bar{\partial}_t \theta^j|||^2 \le Ck \sum_{j=q}^{n} \|\sigma^j + (\alpha_0 + kA_h)\omega^j\|_{-1,h}^2 + Ck^{-1} \sum_{j=0}^{q-1} |||\theta^j|||^2.$$

For the second term above, we use the stronger assumption (3.18) and the arguments in the proof of Theorem 3.1 to derive (compare with (3.14))

$$k \sum_{j=q}^{n} |||\bar{\partial}_t \theta^j|||^2 \le C(u)(h^r + k^q)^2 + Ck \sum_{j=q+1}^{n} |||\omega^j|||^2.$$

The corollary then follows from the arguments after (3.14) and the above inequalities.

Remark 3.2. The condition (3.18) can be satisfied by choosing, for example,

$$U^j = R_h \sum_{l=0}^{q} \frac{(jk)^l}{l!} D_t^l u(0), \qquad j = 0, \dots, q-1,$$

i.e., by including one more term in the sum than in (2.22).

Remark 3.3. The condition $U^q = \bar{U}^q$ does not present additional difficulties in practice. We suggest using the preconditioned iterative scheme to solve for $\bar{U}^q$ up to computer roundoff accuracy. In general, this is of negligible computational cost compared to the work required for the remainder of the time stepping calculation.

4. Incomplete Iterations: The Case of Nonsmooth Initial Data. We consider incomplete iteration applied to the case of nonsmooth initial data in this section. The first theorem gives a result for the Hilbert space case, i.e., the case when there is no spatial discretization. The second theorem considers the fully discrete situation. All of our results apply to the homogeneous equation, i.e., $f = 0$.

For the Hilbert space case, it has been shown by spectral techniques, Le Roux [21], that under the appropriate assumptions about the choice of discrete initial data, the solution of (2.2) satisfies the error estimate

$$\|U^n - u^n\| \le Ck^q t_n^{-q} \|v\| \quad \text{for } n > 0. \tag{4.1}$$

Moreover, for the fully discrete approximation (i.e., the solution of (2.2) with $\mathscr{A} = A_h$), if $v_h = P_h v$ is used with the proper choice of the remaining starting values, then

$$\|U^n - u^n\| \le C(k^q t_n^{-q} + h^r t_n^{-1/2}) \|v\| \quad \text{for } n > 0. \tag{4.2}$$

In this section, we will generalize these results to the algorithms using incomplete iteration defined in Section 3.

For nonsmooth data estimates, we shall require some stronger hypotheses for the starting values. Specifically, we shall assume that

$$\|U^j - u^j\|_{-2q,h} + k^q |||U^j - u^j||| \le Ck^q \|v\| \quad \text{for } j = 0, \dots, q-1. \tag{4.3}$$

The development of starting values satisfying (4.3) will be discussed later.

We now state the theorem in the Hilbert space case.

THEOREM 4.1. *Let $q \le 6$ and u be the solution of* (2.1). *Assume that the starting values $\{U^i\}$, $i = 0, \dots q-1$, satisfy* (4.3), $U^q = \bar{U}^q$ *and* $U^{q+1} = \bar{U}^{q+1}$. *Let U^n for $n > q+1$ be the approximation generated using incomplete iteration as described in Section* 3. *There exists a positive constant δ such that, if $M(n)$ is chosen satisfying*

$$\kappa^{M(n)} \le \delta t_n^{q+1/2}, \tag{4.4}$$

then

$$\|U^n - u^n\| \le Ck^q t_n^{-q} \|v\| \quad \text{for } n > 0 \text{ and } t_n \le T. \tag{4.5}$$

The assumption (4.4) requires more iterations at earlier time steps than in the smooth data case (see Theorem 3.1 and Corollary 3.1).

Proof. By (4.1) and (4.3), there is nothing to prove for $n \le q+1$. We set $e^n = u^n - U^n$ and note that the sequence $\tilde{e}^n = u^n - \bar{U}^n$, $\tilde{e}^{n-j} = u^{n-j} - U^{n-j}$ for $j = 1, \dots q$ satisfies (2.8). Hence

$$L_k e^n = \tau^n + (\alpha_0 + k\mathscr{A})\omega^n = \tau^n + \tilde{\omega}^n = \varphi^n.$$

Applying Lemma 1 with $p = 2q$ and (4.3), we obtain

$$t_n^{2q} \|e^n\|^2 \le Ck \sum_{j=q}^{n} (t_j^{2q} \|\varphi^j\|_{-1}^2 + \|\varphi^j\|_{-2q-1}^2) + Ck^{2q} \|v\|^2.$$

Now for $t_n \le T$,

$$k\sum_{j=q}^{n}(t_j^{2q}\left\|\tilde{\omega}^j\right\|_{-1}^2+\left\|\tilde{\omega}^j\right\|_{-2q-1}^2)\le Ck\sum_{j=q}^{n}\left\|\tilde{\omega}^j\right\|_{-1}^2$$

and

$$\begin{aligned}\left\|\tilde{\omega}^j\right\|_{-1}=\left\|(\alpha_0+k\mathscr{A})\omega^j\right\|_{-1}&=\left\|\mathscr{A}^{-1/2}(\alpha_0+k\mathscr{A})\omega^j\right\|\\&\le C\left\|(I+k\mathscr{A})^{1/2}\omega^j\right\|=C|||\omega^j|||.\end{aligned}\tag{4.6}$$

Thus

$$t_n^{2q}\left\|e^n\right\|^2\le Ck\sum_{j=q}^{n}(t_j^{2q}\left\|\tau^j\right\|_{-1}^2+\left\|\tau^j\right\|_{-2q-1}^2)+Ck\sum_{j=q+2}^{n}|||\omega^j|||+Ck^{2q}\left\|v\right\|^2.\tag{4.7}$$

We next show that

$$k\sum_{j=q}^{n}(t_j^{2q}\left\|\tau^j\right\|_{-1}^2+\left\|\tau^j\right\|_{-2q-1}^2)\le Ck^{2q}\left\|v\right\|^2.\tag{4.8}$$

Let $s\ge -2q-1$. Using (2.11), we clearly have

$$\left\|\tau^j\right\|_s^2\le Ck^{2q-1}\int_{t_{j-q}}^{t_j}\left\|D^{q+1}u(y)\right\|_s^2\,dy,$$

from which it follows that

$$kt_j^{2q+1+s}\left\|\tau^j\right\|_s^2\le Ck^{2q}\int_{t_{j-q}}^{t_j}y^{2q+1+s}\left\|D^{q+1}u(y)\right\|_s^2\,dy\tag{4.9}$$

holds for $j>q$ when $s>-2q-1$ and for $j\ge q$ when $s=-2q-1$.

Let $\{\varphi_j\}_1^\infty$ and $\{\lambda_j\}_1^\infty$ be respectively the eigenfunctions and eigenvalues of the operator $\mathscr{A}$. Then, using the eigenfunction expansion of the solution u, we get

$$\begin{aligned}&\int_0^\infty y^{2q+1+s}\left\|D^{q+1}u(y)\right\|_s^2\,dy\\&\le\int_0^\infty y^{2q+1+s}\sum_{l=1}^{\infty}\lambda_l^{2q+2+s}e^{-2\lambda_l y}(v,\varphi_l)^2\,dy\\&\le C\sum_{l=1}^{\infty}(v,\varphi_l)^2=C\left\|v\right\|^2.\end{aligned}\tag{4.10}$$

Combining (4.9) and (4.10) shows that

$$k\sum_{j=j_0}^{n}t_j^{2q+1+s}\left\|\tau^j\right\|_s^2\le Ck^{2q}\left\|v\right\|^2,\tag{4.11}$$

where $j_0=q+1$ when $s>-2q-1$ and $j_0=q$ when $s=-2q-1$. Thus, to complete the proof of (4.8), it suffices to bound the $j=q$ term in the sum.

We write τ^q as in (2.11) with ς given by

$$\varsigma(t)=\int_0^t u_t(s)\,ds.$$

Then

$$kt_q^{2q}\,\|\tau^q\|_{-1}^2 \le Ck^{2q}\left(\int_0^{t_q}\|u_t(y)\|_{-1}^2\,dy + k\,\|u_t(t_q)\|_{-1}^2\right) \le Ck^{2q}\,\|v\|^2, \tag{4.12}$$

where the second inequality follows easily from techniques used in deriving (4.10). This completes the proof of (4.8).

Combining (4.8) with (4.7) gives

$$t_n^{2q}\,\|e^n\|^2 \le Ck^{2q}\,\|v\|^2 + Ck\sum_{j=q+2}^{n}|||\omega^j|||^2. \tag{4.13}$$

Let ε be a positive constant which is to be specified later. Then, by the argument preceding (3.8), there exists a positive δ such that (4.4) implies

$$\begin{aligned}|||\omega^j||| &\le \varepsilon t_j^{q+1/2}k^q|||\bar\partial_t^{q+1}U^j||| \le \varepsilon t_j^{q+1/2}\{k^q|||\bar\partial_t^{q+1}u^j||| + k^q|||\bar\partial_t^{q+1}e^j|||\}\\ &\le \varepsilon t_j^{q+1/2}\left\{k^q|||\bar\partial_t^{q+1}u^j||| + C\sum_{l=0}^{q}|||\bar\partial_t e^{j-l}|||\right\} \quad \text{for } j > q.\end{aligned}$$

Thus

$$k\sum_{j=q+2}^{n}|||\omega^j|||^2 \le Ck^{2q+1}\sum_{j=q+2}^{n}t_j^{2q+1}|||\bar\partial_t^{q+1}u^j|||^2 + C\varepsilon^2 k\sum_{j=2}^{n}t_j^{2q+1}|||\bar\partial_t e^j|||^2.$$

Using the fact that $\bar\partial_t^{q+1}$ annihilates polynomials of degree up to q, Taylor's formula gives

$$\begin{aligned}t_j^{2q+1}\left\|\bar\partial_t^{q+1}u^j\right\|^2 &\le Ct_j^{2q+1}\left\|\left(\bar\partial_t^{q+1}\int_{t_{j-q-1}}^{t}(t-s)^qD^{q+1}u(s)\,ds\right)_{t=t_j}\right\|^2\\ &\le Ck^{-1}\int_{t_{j-q-1}}^{t_j}s^{2q+1}\left\|D^{q+1}u(s)\right\|^2\,ds.\end{aligned}$$

Hence, using (4.10),

$$k\sum_{j=q+2}^{n}t_j^{2q+1}\left\|\bar\partial_t^{q+1}u^j\right\|^2 \le C\,\|v\|^2.$$

A similar argument, using one less term in the Taylor series gives

$$k^3\sum_{j=q+2}^{n}t_j^{2q+1}\left\|\bar\partial_t^{q+1}u^j\right\|_2^2 \le C\,\|v\|^2.$$

Thus, by interpolation,

$$k\sum_{j=q+2}^{n}t_j^{2q+1}|||\bar\partial_t^{q+1}u^j|||^2 \le C\,\|v\|^2.$$

Consequently,

$$k\sum_{j=q+2}^{n}|||\omega^j|||^2 \le Ck^{2q}\,\|v\|^2 + C\varepsilon^2 k\sum_{j=2}^{n}t_j^{2q+1}|||\bar\partial_t e^j|||^2. \tag{4.14}$$

To estimate the last term of (4.14), we apply Lemma 2 to $\{e^n\}$ and derive

$$k\sum_{j=q}^{n} t_j^{2q+1}|||\bar{\partial}_t e^j|||^2 \le Ck\sum_{j=q}^{n}(t_j^{2q+1}|||\tau^j|||_*^2 + |||\tau^j|||_{*,-2q-1}^2) + Ck\sum_{j=q+2}^{n}|||\omega^j|||^2 + C\sum_{j=0}^{q-1}(|||e^j|||_{*,-2q}^2 + k^{2q}|||e^j|||^2), \tag{4.15}$$

where we used (4.6) to estimate the terms involving $\tilde{\omega}^j$. Applying (4.11) gives

$$k\sum_{j=q+1}^{n} t_j^{2q+1}|||\tau^j|||_*^2 \le k\sum_{j=q+1}^{n} t_j^{2q+1}\left\|\tau^j\right\|^2 \le Ck^{2q}\|v\|^2.$$

As in (4.12),

$$kt_q^{2q+1}|||\tau^q|||_*^2 \le Ck^{2q+1}\left(\int_0^{t_q}|||u_t(y)|||_*^2\,dy + k|||u_t(t_q)|||_*^2\right) \le Ck^{2q}\left(\int_0^{t_q}\|u_t(y)\|_{-1}^2\,dy + k\|u_t(t_q)\|_{-1}^2\right) \le Ck^{2q}\|v\|^2.$$

Clearly, by (4.8),

$$k\sum_{j=q}^{n}|||\tau^j|||_{*,-2q-1}^2 \le k\sum_{j=q}^{n}\left\|\tau^j\right\|_{-2q-1}^2 \le Ck^{2q}\|v\|^2.$$

Combining the above estimates with (4.3) gives

$$k\sum_{j=2}^{n} t_j^{2q+1}|||\bar{\partial}_t e^j|||^2 \le k\sum_{j=2}^{q-1} t_j^{2q+1}|||\bar{\partial}_t e^j|||^2 + Ck^{2q}\|v\|^2 + Ck\sum_{j=q+2}^{n}|||\omega^j|||^2 \le Ck^{2q}\|v\|^2 + Ck\sum_{j=q+2}^{n}|||\omega^j|||^2.$$

Together with (4.14) this shows

$$k\sum_{j=q+2}^{n}|||\omega^j|||^2 \le Ck^{2q}\|v\|^2 + C\varepsilon^2 k\sum_{j=q+2}^{n}|||\omega^j|||^2,$$

and hence, if ε is chosen small enough,

$$k\sum_{j=q+2}^{n}|||\omega^j|||^2 \le Ck^{2q}\|v\|^2.$$

Hence, (4.13) yields

$$t_n^{2q}\|e^n\|^2 \le Ck^{2q}\|v\|^2,$$

which completes the proof of the theorem.

Remark 4.1. The arguments up to (4.13) provide a proof of (4.1) in the case in which (3.1) is solved exactly, i.e., $\omega^j = 0$.

We shall briefly indicate by an example how initial data can be constructed to satisfy (4.3). Take a rational function $r(\lambda)$ satisfying

$$r(\lambda) = e^{-\lambda} + O(\lambda^q) \quad \text{as } \lambda \to 0 \tag{4.16}$$

and

$$|r(\lambda)| < 1 \quad \text{for } \lambda > 0,\ r(\infty) = 0. \tag{4.17}$$

Set

$$U^j = r(k\mathscr{A})^j v, \qquad j = 0, \dots, q-1.$$

Then, by spectral representation we have

$$\begin{aligned} \left\| U^j - u^j \right\|_{-2q} &= \| \mathscr{A}^{-q} (r(k\mathscr{A})^j - \exp(-jk\mathscr{A}))v \| \\ &\le k^q \sup_{\lambda > 0} |\lambda^{-q}(r(\lambda)^j - e^{-j\lambda})| \, \|v\| \le Ck^q \|v\| \end{aligned}$$

and

$$\begin{aligned} |||U^j - u^j||| &= \|(I + k\mathscr{A})^{1/2}(r(k\mathscr{A})^j - \exp(-jk\mathscr{A}))v\| \\ &\le \sup_{\lambda > 0} |(1+\lambda)^{1/2}(r(\lambda)^j - e^{-j\lambda})| \, \|v\| \le C \|v\|, \end{aligned}$$

from which (4.3) follows.

Note that the above choice of $U^1, \dots, U^{q-1}$ corresponds to applying the single step operator corresponding to $r(k\mathscr{A})$ for the first $q-1$ steps. The order of accuracy defined by (4.16) is only $q-1$, which suffices since it is only used a fixed number of times (independent of k). For instance, if $q = 2$, then U^1 may be computed by the first-order backward Euler method ($r(\lambda) = 1/(1+\lambda)$). More generally, we can choose $r(\lambda)$ to be the subdiagonal Padé approximation of the appropriate order to $e^{-\lambda}$.

We end this section by applying our above nonsmooth data error estimate to the solution of a parabolic equation which has already been discretized with respect to the space variables (defined by (2.13)). If $v_h = P_h v$, then the solution of (2.13) satisfies (cf. [11])

$$\|u_h(t) - u(t)\| \le Ch^r t^{-r/2} \|v\| \quad \text{for } t \ge 0. \tag{4.18}$$

We can now give the theorem for the fully discrete time stepping scheme.

THEOREM 4.2. *Let $q \le 6$. Consider the incomplete iteration scheme described in Section 3 applied to (2.13) with $f = 0$ and initial data v only in $L^2(\Omega)$. Let u_h solve (2.13) with $v_h = P_h v$ and assume that the starting procedure is such that*

$$\left\| U^j - u_h(jk) \right\|_{-2q,h} + k^q |||U^j - u_h(jk)||| \le Ck^q \|v\| \quad \text{for } j = 0, \dots, q-1. \tag{4.19}$$

Let $U^q = \bar{U}^q$ and $U^{q+1} = \bar{U}^{q+1}$. There exists a positive constant δ such that, if $M(n)$ is chosen satisfying

$$\kappa^{M(n)} \le \delta t_n^{q+1/2},$$

then

$$\|U^n - u(t_n)\| \le C(h^r t_n^{-r/2} + k^q t_n^{-q}) \|v\| \quad \text{for } n \ge q \text{ and } t_n \le T.$$

Proof. Using the triangle inequality this follows at once by Theorem 4.1 applied to the equation (2.13), together with the estimate (4.18).

Initial values satisfying (4.19) may now be chosen in the form

$$U^j = r(kA_h)^j P_h v \quad \text{for } j = 0, \dots, q-1, \tag{4.20}$$

with $r(\lambda)$ satisfying (4.16) and (4.17). Clearly, the argument following (4.17) implies (4.19).

In order to compute the values of $U^j, j = 0, \ldots, q+2$, one must solve algebraic systems of the form

$$(\gamma + \beta A_h)U^j = \text{data} \tag{4.21}$$

with appropriate data. Clearly, these equations may be solved iteratively. Moreover, the cost of iteratively solving a fixed number of problems of the form (4.21) is small compared to the computational effort required for the remainder of the time stepping scheme.

5. Proofs of Lemmas 1 and 2. This section gives the proofs of Lemmas 1 and 2. By eigenfunction expansions of U^n and f^n, we find that it suffices to show that the lemmas hold in the case where H is the set of real numbers. Then the operator $\mathscr{A}$ corresponds to multiplication by a scalar $\tilde{\lambda}$ and the solution U^n satisfies the recurence

$$(\alpha_0 + k\tilde{\lambda})U^n + \alpha_1 U^{n-1} + \cdots + \alpha_q U^{n-q} = \bar{f}^n \quad \text{for } n \geq q, \tag{5.1}$$

where $\bar{f}^n = kf^n$. The solution of (5.1) can be written

$$U^n = (\alpha_0 + k\tilde{\lambda})^{-1} \left(\sum_{j=0}^{n-q} \beta_j \bar{f}^{n-j} - \sum_{s=0}^{q-1} \left(\sum_{j=q-s}^{q} \beta_{n-s-j}\alpha_j \right) U^s \right), \tag{5.2}$$

where $\beta_j = 0$ for $j < 0$, $\beta_0 = 1$ and $\beta_j = \beta_j(k\tilde{\lambda})$ for $j > 0$ is defined recursively by

$$(\alpha_0 + k\tilde{\lambda})\beta_j + \alpha_1\beta_{j-1} + \cdots + \alpha_q\beta_{j-q} = 0. \tag{5.3}$$

The following estimates for β_j will be useful in the proof of the Lemmas 1 and 2.

LEMMA 5.1. *Let $q \leq 6$. There are positive constants c, C and λ_0 such that*

$$|\beta_j(\lambda)| \leq \begin{cases} Ce^{-cj\lambda} & \text{for } 0 < \lambda \leq \lambda_0, \\ Ce^{-cj} & \text{for } \lambda \geq \lambda_0. \end{cases}$$

Lemma 5.1 was proved in [14], [21]. We include a proof for completeness and since similar arguments will be used later in this section.

Proof. Consider the polynomial

$$P(\tau, \lambda) = \tau^q + (\alpha_1\tau_{q-1} + \ldots + \alpha_q)/(\alpha_0 + \lambda). \tag{5.4}$$

The solution of (5.3) can be written

$$\beta_j(\lambda) = \frac{1}{2\pi i} \int_\Gamma \frac{\tau^{j+q-1}}{P(\tau, \lambda)} d\tau, \tag{5.5}$$

where Γ is a closed path in the complex plane which winds once around each root of $P(\cdot, \lambda)$. Indeed, the sequence $\beta_j(\lambda)$ given by (5.5) clearly satisfies (5.3) for $j > 0$. Moreover, a straightforward application of Rouché's Theorem implies that the above expression exhibits the correct initial values.

Let $\tau_i(\lambda)$ denote the ith root of $P(\tau, \lambda) = 0$. It is known that $P(\tau) = P(\tau, 0)$ has a simple zero at $\tau = 1$ and that the remaining zeros are in the interior of the unit disk. Further, for any $\lambda > 0$, all roots of $P(\cdot, \lambda)$ are in the interior of the unit disk and tend to zero as λ tends to infinity. We order these roots so that $\tau_l(\lambda)$ is a

continuous function of λ for each l, and set $\tau_1(0) = 1$. Elementary manipulations give

$$\tau_1(\lambda) = 1 - \lambda/\tilde{P}'(1) + O(\lambda^2),$$

where

$$\tilde{P}(\tau) = \alpha_0 P(\tau, 0) = \sum_{j=1}^{q} \frac{(\tau - 1)^j}{j} \tau^{q-j}.$$

Clearly,

$$\tau_1(\lambda) = 1 - \lambda + O(\lambda^2) \leq 1 - \lambda/2 \tag{5.6}$$

for λ in some neighborhood of the origin. Hence, there exists a positive constant λ_0 such that (5.6) holds and $\tau_1(\lambda)$ is a simple root of $P(\tau, \lambda) = 0$ for $0 \leq \lambda \leq \lambda_0$. The remaining roots are bounded in absolute value by $1 - \delta$ for some positive constant δ independent of $\lambda \geq 0$, and hence we can assume that they are a bounded distance away from $\tau_1(\lambda)$ for $0 \leq \lambda \leq \lambda_0$.

Let $P(\tau, \lambda) = (\tau - \tau_1(\lambda))Q(\tau, \lambda)$; then

$$\beta_j(\lambda) = \frac{[\tau_1(\lambda)]^{j+q-1}}{Q(\tau_1(\lambda), \lambda)} + \frac{1}{2\pi i} \int_\Gamma \frac{R(\tau, \lambda)\tau^{j+q-1}}{Q(\tau, \lambda)}\, d\tau, \tag{5.7}$$

where

$$R(\tau, \lambda) = \frac{Q(\tau_1(\lambda), \lambda) - Q(\tau, \lambda)}{(\tau - \tau_1(\lambda))Q(\tau_1(\lambda), \lambda)}.$$

In view of the above discussion, it is easily seen that the first term in (5.7) is bounded by $Ce^{-cj\lambda}$ for λ in $(0, \lambda_0]$.

For the second term of (5.7), we note that for each λ, $R(\tau, \lambda)$ is a polynomial in τ whose roots depend continuously on λ. Consequently, the roots of $R(\cdot, \lambda)$ can be bounded independent of λ in the interval $[0, \lambda_0]$, and hence

$$|R(\tau, \lambda)| \leq C \quad \text{for } 0 \leq \lambda \leq \lambda_0 \text{ and } |\tau| \leq 1.$$

Taking Γ in (5.7) to be the circle centered at the origin of radius $1 - \delta/2$ implies that the second term in (5.7) can be bounded by $Ce^{-\delta j/2}$. This verifies the lemma for $0 < \lambda \leq \lambda_0$.

For $\lambda \geq \lambda_0$, we note that all roots of $P(\cdot, \lambda)$ are bounded in absolute value by $1 - \delta$ for some positive constant δ independent of λ. The lemma in this case easily follows, taking Γ in (5.5) to be the circle centered at the origin of radius $1 - \delta/2$.

LEMMA 5.2. *Let $q \leq 6$ and $\tilde{\beta}_j \equiv \tilde{\beta}_j(\lambda) = (\alpha_0 + \lambda)^{-1}\beta_j(\lambda)$. We then have*

$$j^p|\tilde{\beta}_j(\lambda)| \leq C(1 + \lambda^{-p}) \quad \textit{for } j \geq 0$$

and

$$\lambda \sum_{j=0}^{\infty} j^p|\tilde{\beta}_j(\lambda)| \leq C(1 + \lambda^{-p}).$$

Proof. For $\lambda \leq \lambda_0$,

$$j^p|\tilde{\beta}_j| \leq Cj^p|\beta_j| \leq Cj^p e^{-c\lambda j} \leq C\lambda^{-p}$$

and

$$\lambda \sum_{j=0}^{\infty} j^p|\tilde{\beta}_j| \leq C\lambda \sum_{j=0}^{\infty} j^p e^{-c\lambda j} \leq C\lambda \int_0^{\infty} t^p e^{-c\lambda t} dt \leq C\lambda^{-p}.$$

For $\lambda \geq \lambda_0$,

$$j^p|\tilde{\beta}_j| \leq Cj^p|\beta_j| \leq Cj^p e^{-cj} \leq C$$

and

$$\lambda \sum_{j=0}^{\infty} j^p|\tilde{\beta}_j| \leq C(1+\lambda)^{-1}\lambda \sum_{j=0}^{\infty} j^p e^{-cj} \leq C,$$

which proves the lemma.

Proof of Lemma 1. We are to show that $\{U^n\}$ given by (5.2) satisfies

$$t_n^p(U^n)^2 + k\sum_{j=q}^{n} t_j^p\tilde{\lambda}(U^j)^2 \leq Ck\sum_{j=q}^{n}(t_j^p\tilde{\lambda}^{-1} + \tilde{\lambda}^{-p-1})(f^j)^2 + C\sum_{j=0}^{q-1}(\tilde{\lambda}^{-p} + k^p)(U^j)^2. \tag{5.8}$$

Putting $\lambda = k\tilde{\lambda}$, (5.8) becomes

$$n^p(U^n)^2 + \lambda\sum_{j=q}^{n} j^p(U^j)^2 \leq C\sum_{j=q}^{n}(j^p\lambda^{-1} + \lambda^{-p-1})(\bar{f}^j)^2 + C\sum_{j=0}^{q-1}(1+\lambda^{-p})(U^j)^2. \tag{5.9}$$

We shall prove the lemma by considering two cases. The first case is when $U^0 = \cdots = U^{q-1} = 0$, and the second is when $f^j = 0$ for $j = q,\ q+1, \ldots$. Clearly, the proof of the lemma will be complete when we show that (5.9) holds in each of these cases.

For the first case, we have

$$U^n = \sum_{j=0}^{n-q} \tilde{\beta}_j \bar{f}^{n-j} \quad \text{for } n \geq q.$$

Using the Schwarz inequality and Lemma 5.2 with $p = 0$, we obtain

$$(U^n)^2 \leq \left(\sum_{j=0}^{n-q}|\tilde{\beta}_j|\right)\left(\sum_{j=0}^{n-q}|\tilde{\beta}_j|(\bar{f}^{n-j})^2\right) \leq C\lambda^{-1}\sum_{j=0}^{n-q}|\tilde{\beta}_j|(\bar{f}^{n-j})^2.$$

Now

$$n^p \leq C(j^p + (n-j)^p), \tag{5.10}$$

and hence

$$n^p(U^n)^2 \leq C\lambda^{-1}\sum_{j=0}^{n-q}\{j^p|\tilde{\beta}_j|(\bar{f}^{n-j})^2 + |\tilde{\beta}_j|(n-j)^p(\bar{f}^{n-j})^2\}. \tag{5.11}$$

Applying Lemma 5.2 gives

$$n^p(U^n)^2 \leq C\lambda^{-1}\sum_{j=0}^{n-q}(\lambda^{-p} + (n-j)^p)(\bar{f}^{n-j})^2,$$

which is the desired estimate for the first term in (5.9) in this case.

We now turn to the second term of (5.9). By summation of (5.11), we obtain

$$\lambda \sum_{n=q}^{N} n^p (U^n)^2 \le C \sum_{n=q}^{N} \sum_{j=0}^{n-q} \{ j^p |\tilde{\beta}_j| (\bar{f}^{n-j})^2 + |\tilde{\beta}_j| (n-j)^p (\bar{f}^{n-j})^2 \}$$
$$\le C \sum_{n=q}^{N} (\bar{f}^n)^2 \sum_{j=0}^{N-q} j^p |\tilde{\beta}_j| + C \sum_{n=q}^{N} n^p (\bar{f}^n)^2 \sum_{j=0}^{N-q} |\tilde{\beta}_j|.$$

Applying Lemma 5.2 gives

$$\lambda \sum_{n=q}^{N} n^p (U^n)^2 \le C\lambda^{-1} \sum_{n=q}^{N} (\lambda^{-p} + n^p)(\bar{f}^n)^2$$

which verifies (5.9) for the first case.

We now consider the second case. In addition, assume that $U^1 = \cdots = U^{q-1} = 0$. Then

$$U^n = -\tilde{\beta}_{n-q} \alpha_q U^0,$$

and hence by (5.10) and Lemma 5.2,

$$n^p (U^n)^2 \le C n^p \tilde{\beta}_{n-q}^2 (U^0)^2 \le C(1 + (n-q)^p) |\tilde{\beta}_{n-q}| (U^0)^2 \tag{5.12}$$
$$\le C(1 + \lambda^{-p})(U^0)^2.$$

By summation of (5.12) we obtain

$$\lambda \sum_{n=q}^{N} n^p (U^n)^2 \le C\lambda \sum_{n=q}^{N} (1 + (n-q)^p) |\tilde{\beta}_{n-q}| (U^0)^2$$
$$\le C(1 + \lambda^{-p})(U^0)^2.$$

This proves (5.9) when $U^1 = \cdots = U^{q-1} = 0$. The arguments verifying (5.9) when $U^i \neq 0$, $i = 1, \ldots, q-1$, are similar and will not be given. This completes the proof of Lemma 1.

The remainder of this section is devoted to the proof of Lemma 2. We note that Eq. (2.2) may be written

$$\sum_{j=0}^{q-1} \gamma_j \bar{\partial}_t U^{n-j} + \mathscr{A} U^n = f^n \quad \text{for } n \ge q.$$

Clearly,

$$\alpha_0 P(x, 0) = (x-1) \sum_{j=0}^{q-1} \gamma_j x^{q-j-1} \equiv (x-1) Q(x).$$

Hence, the roots of $Q(x)$ are in the interior of the unit disk. For the proof of Lemma 2, we shall use the following lemma which gives estimates for solutions to the difference equation with characteristic polynomial Q.

LEMMA 3. *Let $q \le 6$, $p \ge 0$, and $\{\gamma_j\}$ be as above. Let $\{W^n\}$ be the solution of the difference equation*

$$\sum_{j=0}^{q-1} \gamma_j W^{n-j} = F^n \quad \text{for } n \ge q.$$

Then

$$\sum_{j=q}^{n} j^p (W^j)^2 \le C \sum_{j=q}^{n} j^p (F^j)^2 + C \sum_{j=1}^{q-1} (W^j)^2.$$

Proof. As in (5.2),

$$W^n = \gamma_0^{-1} \left(\sum_{j=0}^{n-q} \mu_j F^{n-j} - \sum_{s=1}^{q-1} \left(\sum_{j=q-s}^{q-1} \mu_{n-s-j} \gamma_j \right) W^s \right) \quad \text{for } n \ge q.$$

Here the $\mu_j = 0$ for $j < 0$, $\mu_0 = 1$ and

$$\gamma_0 \mu_j + \gamma_1 \mu_{j-1} + \cdots + \gamma_{q-1} \mu_{j-q+1} = 0 \quad \text{for } j \ge 1.$$

Clearly, $|\mu_j| \le Ce^{-cj}$ (see the proof of Lemma 5.1), from which it obviously follows that $j^p |\mu_j| \le C$ and

$$\sum_{j=0}^{\infty} j^p |\mu_j| \le C.$$

The lemma now follows from the arguments given in the proof of Lemma 1.

Proof of Lemma 2. We are to show

$$(1 + k\tilde{\lambda}) k \sum_{j=q}^{n} t_j^p (\bar{\partial}_t U^j)^2 \le Ck(1 + k\tilde{\lambda})^{-1} \sum_{j=q}^{n} (t_j^p + \tilde{\lambda}^{-p})(f^j)^2 + C \sum_{j=0}^{q-1} ((1 + k\tilde{\lambda})^{-1} \tilde{\lambda}^{-p+1} + k^{p-1}(1 + k\tilde{\lambda}))(U^j)^2,$$

which can be rewritten ($\lambda = k\tilde{\lambda}$)

$$\sum_{j=q}^{n} j^p (U^j - U^{j-1})^2 \le C(1 + \lambda)^{-2} \sum_{j=q}^{n} (j^p + \lambda^{-p})(\bar{f}^j)^2 + C \sum_{j=0}^{q-1} (1 + (1 + \lambda)^{-2} \lambda^{-p+1})(U^j)^2. \tag{5.13}$$

Again, the proof of this lemma is reduced to verifying (5.13) for the two cases in the proof of Lemma 1.

We consider the first case, i.e., $U^0 = \cdots = U^{q-1} = 0$. From the definition of Q, we have

$$(1 + \lambda) \sum_{j=0}^{q-1} \gamma_j (U^{n-j} - U^{n-j-1}) = \bar{f}^n - \lambda U^n + \lambda \sum_{j=0}^{q-1} \gamma_j (U^{n-j} - U^{n-j-1}) = \bar{f}^n + \lambda \sum_{j=0}^{q} \tilde{\gamma}_j U^{n-j}. \tag{5.14}$$

By Lemma 3,

$$\sum_{j=q}^{n} j^p (U^j - U^{j-1})^2 \le C(1 + \lambda)^{-2} \sum_{j=q}^{n} j^p (\bar{f}^j)^2 + C\lambda^2 (1 + \lambda)^{-2} \sum_{j=q}^{n} j^p (U^j)^2.$$

By (5.9), we have

$$\lambda^2 \sum_{j=q}^{n} j^p (U^j)^2 \leq C \sum_{j=q}^{n} (j^p + \lambda^{-p})(\bar{f}^j)^2.$$

Combining the above two inequalities verifies (5.13) for the first case.

For the second case, once again applying Lemma 3 to (5.14) gives

$$\begin{aligned}\sum_{j=q}^{n} j^p (U^j - U^{j-1})^2 &\leq C\lambda^2 (1+\lambda)^{-2} \left((U^0)^2 + \sum_{j=1}^{n} j^p (U^j)^2 \right) \\ &\quad + C \sum_{j=1}^{q-1} (U^j - U^{j-1})^2 \\ &\leq C\lambda^2 (1+\lambda)^{-2} \sum_{j=q}^{n} j^p (U^j)^2 + C \sum_{j=0}^{q-1} (U^j)^2.\end{aligned}$$

By (5.9),

$$\lambda^2 \sum_{j=q}^{n} j^p (U^j)^2 \leq C\lambda \sum_{j=0}^{q-1} (1 + \lambda^{-p})(U^j)^2.$$

Combining the above two inequalities implies (5.13) for the second case. This completes the proof of Lemma 2.

6. Preconditioning and Numerical Experiments. In this section, we shall describe the results of computational experiments illustrating the theory presented earlier. To more fully describe the algorithms employed, we first discuss the preconditioning techniques used to define the iterative process (3.1). We shall also demonstrate how these techniques can be used in the computation of the starting values defined by (4.20). We then give numerical results for the algorithms applied to the smooth as well as nonsmooth initial value problems.

The iterative approximation $U^{n,m}$ for the solution $\bar{U}^n$ of (3.1) will be defined by preconditioned conjugate gradients [13]. The preconditioner B_{kh} is a symmetric positive definite linear operator defined on S_h which, to be computationally effective, should satisfy

(1) The action of B_{kh} on arbitrary functions in S_h should be computationally less expensive than that of $(\alpha_0 + kA_h)^{-1}$.

(2) The operator B_{kh} should approximately invert $(\alpha_0 + kA_h)$ in the sense that there are positive constants κ_0, κ_1 satisfying

$$\kappa_0 (B_{kh}^{-1} v, v) \leq ((\alpha_0 + kA_h)v, v) \leq \kappa_1 (B_{kh}^{-1} v, v) \quad \text{for all } v \in S_h \tag{6.1}$$

with κ_1/κ_0 close to one.

In our computational examples, we shall use preconditioners which are based on multigrid iteration [4] and lead to constants κ_0 and κ_1 satisfying (6.1) with $\kappa_1/\kappa_0 \leq C$. For additional techniques for the construction of preconditioners see [3]–[9], [13], [17], [25]. Note that we require a family of preconditioners since the operators $\{B_{kh}\}$ are indexed by k and h.

Remark 6.1. We note that the inequalities (6.1) are equivalent to the inequalities

$$\kappa_0((\alpha_0 + kA_h)^{-1}v, v) \leq (B_{kh}v, v) \leq \kappa_1((\alpha_0 + kA_h)^{-1}v, v) \quad \text{for all } v \in S_h.$$

It is well known (cf. [23]) that the sequence of iterates $\{U^{n,m}\}$ defined by the preconditioned conjugate gradient method with preconditioner B_{kh} as above satisfies (3.3) with

$$\kappa = \frac{\sqrt{\kappa_1/\kappa_0} - 1}{\sqrt{\kappa_1/\kappa_0} + 1}$$

and $c_0 = 2$.

We next consider the problem of computing the starting values U^j given by (4.20). In the case of $q = 2$, the starting value U^1 is determined by the backward Euler ($q = 1$) method. We can obviously use the preconditioner B_{kh} and solve for U^1 to computer round off. This only involves a number of iterations proportional to the number of significant digits on the given machine. Similar techniques are used to compute U^q and U^{q+1} (for general q) when they are defined to be the exact solutions of (3.1).

We now describe the computation of the starting values for $q = 3$ and $q = 4$. The starting values for higher q can be developed in a similar manner. As discussed earlier, it suffices to use the subdiagonal Padé approximation of order three, i.e.,

$$r(\tau) = \frac{1 - \tau/3}{1 + 2\tau/3 + \tau^2/6}.$$

As observed in [1], r can be written

$$r(\tau) = 1 - \mathrm{Re}\left(\frac{\gamma\tau}{1 + \beta\tau}\right),$$

where $\beta = (1 + i\sqrt{2}/2)/3$ and $\gamma = 1 - i\sqrt{2}/2$. Consequently, $U^j = r(kA_h)U^{j-1}$ can be written $U^j = U^{j-1} - \mathrm{Re}(W)$, where W is the solution to

$$(W, \chi) + k\beta A(W, \chi) = k\gamma A(U^{j-1}, \chi) \quad \text{for all } \chi \in S_h. \tag{6.2}$$

We next show how the preconditioner B_{kh} can be used to efficiently solve (6.2). We first set up a simple iteration for the solution of (6.2) which involves the inversion of the operator $A_k^\delta = I + k\delta A_h$ and subsequently show that A_k^δ can be replaced by a preconditioner without significant loss of efficiency. Here, δ is a positive number which we are to provide. Starting from an initial guess W^0 (e.g., $W^0 = 0$) for the solution W, we define a sequence of iterates $\{W^l\}$ for $l > 0$ by

$$W^l = W^{l-1} + (A_k^\delta)^{-1}R^{l-1}, \tag{6.3}$$

where R^{l-1} is the residual defined by

$$R^{l-1} = kA_h(\gamma U^{j-1} - \beta W^{l-1}) - W^{l-1}.$$

The iterative scheme (6.3) can be analyzed by estimating the components of the error in terms of the eigenvectors and eigenvalues of A_h. In fact,

$$\left\|(A_k^\delta)^{1/2}(I - (A_k^\delta)^{-1}(I + k\beta A_h))V\right\|^2 \leq C_\delta \left\|(A_k^\delta)^{1/2}V\right\|^2 \quad \text{for all } V \in S_h \tag{6.4}$$

holds for

$$C_\delta = \sup_{\lambda \geq 0}\left|1 - \frac{1 + k\beta\lambda}{1 + k\delta\lambda}\right|^2.$$

Inequality (6.4) implies that iteration (6.3) converges in the norm $(A_k^\delta \cdot, \cdot)^{1/2}$ at a rate bounded by $C_\delta^{1/2}$. Clearly, $C_\delta \le |1 - \beta/\delta|^2$ so, for example, $C_{1/2} \le 1/3$.

We now describe a similar iteration replacing A_k^δ by a preconditioner B_{kh}. We assume that (6.1) holds with A_k^δ replacing $(\alpha_0 + kA_h)$ and that δ is chosen so that $C_\delta < 1$. We replace (6.3) with

$$W^l = W^{l-1} + \mu B_{kh} R^{l-1}, \tag{6.5}$$

where μ is a positive iteration parameter. Let $E^l = W - W^l$; then

$$\begin{aligned}(B_{kh}^{-1}E^l, E^l) = (B_{kh}^{-1}E^{l-1}, E^{l-1}) - 2\mu\mathrm{Re}((I + k\beta A_h)E^{l-1}, E^{l-1}) \\ + \mu^2(B_{kh}(I + k\beta A_h)E^{l-1}, (I + k\beta A_h)E^{l-1}).\end{aligned}$$

By (6.4),

$$\begin{aligned}(A_k^\delta E^{l-1}, E^{l-1}) + ((A_k^\delta)^{-1}(I + k\beta A_h)E^{l-1}, (I + k\beta A_h)E^{l-1}) \\ \le C\mathrm{Re}((I + k\beta A_h)E^{l-1}, E^{l-1})\end{aligned}$$

holds for $C = 2/(1 - C_\delta)$. Hence (6.1) implies

$$\begin{aligned}(B_{kh}^{-1}E^l, E^l) &\le (B_{kh}^{-1}E^{l-1}, E^{l-1}) - 2\mu\mathrm{Re}((I + k\beta A_h)E^{l-1}, E^{l-1}) \\ &\quad + C\mu^2((A_k^\delta)^{-1}(I + k\beta A_h)E^{l-1}, (I + k\beta A_h)E^{l-1}) \\ &\le (B_{kh}^{-1}E^{l-1}, E^{l-1}) - 2\mu(1 - C\mu)\mathrm{Re}((I + k\beta A_h)E^{l-1}, E^{l-1}) \\ &\le [1 - c\mu(1 - C\mu)](B_{kh}^{-1}E^{l-1}, E^{l-1}).\end{aligned}$$

By choosing μ small enough (independently of k and h), we can make $(1 - c\mu(1 - C\mu))$ less than one. Hence, each iteration of (6.3) will reduce the error by a fixed factor independent of k and h. Thus, the computation of W only requires a number of iterations proportional to the number of significant digits on the given computer.

Example 1. The first example which we shall consider is the one-dimensional problem

$$\begin{aligned}u_t - u_{xx} &= f \quad \text{for } (x,t) \in (0,1) \times (0,1], \\ u(0,t) = u(1,t) &= 0 \quad \text{for } t \in (0,1], \\ u(x,0) &= 0 \quad \text{for } x \in (0,1).\end{aligned} \tag{6.6}$$

The function f is defined so that (6.6) has the solution

$$u(x,t) = \sin(40t)e^{-1/x^2 - 1/(x-1)^2}.$$

Accordingly, the smooth data results of this paper apply.

For this example, we define S_h to be the set of continuous piecewise linear functions on a uniform mesh of size h (which vanish at $x = 0$ and $x = 1$). For these subspaces, (3.1) can be solved trivially, and hence the results presented for this example will only illustrate the theory of Section 2. Results for incomplete iteration will be given in later examples. Starting values U^i, $i = 1, \ldots, q-1$, were generated by either the backward Euler method or the (2,1) Padé approximation discussed earlier, with appropriate modification to take into account the forcing function f.

Table 6.1 gives the normalized discrete L^2 error $E(q,k)$ as a function of the order of time step approximation q and the time step size k. This error is defined by

$$E(q,k) = \left(\frac{\sum_i (U^j(x_i) - u(x_i,1))^2}{\sum_i u(x_i,1)^2}\right)^{1/2}. \tag{6.7}$$

The sums in (6.7) are over the nodes x_i of the subspace, and the time level j corresponds to $t_j = 1$. For the results reported in Table 6.1, the mesh parameter h was chosen sufficiently small so that the time step error dominated the computation. The results illustrate the higher-order convergence suggested by Theorem 2.2. Note that for higher-accuracy approximation, the higher-order schemes are always more efficient.

TABLE 6.1
Discrete L^2 error for Example 1.

k	$q = 2$	$q = 3$	$q = 4$
1/10	1.5	1.4	1.5
1/20	.59	.48	.41
1/40	.21	.068	.12
1/80	.055	.013	.011
1/160	.013	.0025	.0006

Example 2. For our second example, we consider approximating the solution of the problem

$$\begin{aligned} u_t - \Delta u &= f \quad \text{in } \Omega \times (0,1], \\ u(x,t) &= 0 \quad \text{on } \partial\Omega \times (0,1], \\ u(x,0) &= 0 \quad \text{for } x \in \Omega, \end{aligned} \tag{6.8}$$

where Ω is the unit square in R^2. The function f is defined so that (6.8) has the solution

$$u(x,y,t) = \sin(40t)e^{-1/x^2-1/(x-1)^2-1/y^2-1/(y-1)^2}.$$

Even though $\partial\Omega$ is not smooth, in this case, it is possible to prove results similar to those given earlier.

To define the approximation subspaces, we first break Ω into $n \times n$ square subregions and partion each subregion into two triangles by the diagonal connecting the bottom left corner with the top right. We define S_h to be the set of continuous piecewise linear functions (which vanish on $\partial\Omega$) on this mesh and set $h = 1/n$. We shall report results for this scheme which use incomplete iteration approximating the solution of (3.1) at each time step. Even though there are 'fast' direct methods for the solution of the corresponding system (3.1), we feel that the incomplete iteration results presented in this example are important since they are representative of the type of results expected in more general applications. As in the previous example, starting values $U^i, i = 1, \ldots, q-1$, were generated by the backward Euler method or the (2,1) Padé approximation discussed earlier, with appropriate modification to take into account the forcing function f.

We use preconditioned conjugate gradient to define the iterative approximation $U^{n,m}$ for the solution of (3.1). The preconditioner B_{kh} is defined in terms of a multigrid iteration which we will not describe here (see for example, [4]). Note, however, that Table 6.2 gives the condition number K for the preconditioned system (defined to be the smallest ratio $K = \kappa_1/\kappa_0$ satisfying (6.1)). The reported condition numbers were for the case $q = 2$; the condition numbers for $q = 3$ and

TABLE 6.2
Discrete L^2 error for Example 2.

$M(n) = 3$					
k	h	$q = 2$	$q = 3$	$q = 4$	K
1/10	1/16	.89	.93	.99	2.2
1/20	1/32	.38	.30	.25	2.3
1/40	1/64	.12	.046	.11	2.4
1/80	1/64	.028	.033	.032	2.4
1/160	1/64	.024	.033	.032	2.4
$M(n) = 5$					
k	h	$q = 2$	$q = 3$	$q = 4$	K
1/40	1/64	.122	.048	.097	2.4
1/80	1/64	.029	.015	.0086	2.4
1/160	1/64	.0073	.0053	.0042	2.4
$M(n) = 1$					
k	h	$q = 2$	$q = 3$	$q = 4$	K
1/40	1/64	.17	.22	.30	2.4
1/80	1/64	.28	.34	.38	2.4
1/160	1/64	.99	1.2	1.3	2.4

$q = 4$ were almost identical. We also give results as a function of $M(n)$, the number of preconditioned conjugate gradient steps in the incomplete iteration for the solution of (3.1). For $n \leq q + 1$, sufficiently many steps were taken to essentially solve the problem.

Good convergence results were obtained for $M(n) = 3$ and $M(n) = 5$. However, the results obtained for $M(n) = 1$ seem to suggest instability. This is in agreement with the theory, which requires enough iterations to at least beat a threshold error. Three iterations were sufficient in this example.

Example 3. Our last example will illustrate model computations on a problem with nonsmooth initial data. We consider the problem

$$\begin{aligned} u_t - \frac{1}{12}\Delta u &= 0 && \text{in } \Omega \times (0,1], \\ u(x,y,t) &= 0 && \text{on } \partial\Omega \times (0,1], \\ u(x,y,0) &= v(x,y) && \text{for } x \in \Omega, \end{aligned} \tag{6.9}$$

where Ω is the unit square in R^2 and

$$v(x,y) = \begin{cases} 1 & \text{if } 1/4 \leq x,\ y \leq 3/4, \\ 0 & \text{otherwise.} \end{cases}$$

The solution of (6.9) is given by

$$u(x,y,t) = \frac{8}{\pi^2} \sum_{i,j=0}^{\infty} c_i c_j \sin(\pi x(2i+1)) \sin(\pi y(2j+1)) \times \exp\left(-\pi^2 t \frac{(2i+1)^2 + (2j+1)^2}{12}\right),$$

where

$$c_i = \begin{cases} (-1)^{(i/2)}(2i+1)^{-1} & \text{if } i \text{ is even,} \\ (-1)^{(i+1)/2}(2i+1)^{-1} & \text{otherwise.} \end{cases}$$

It is easy to see that $v \in H^\alpha$ only for $\alpha < 1/2$.

For this example, we use the same approximation subspaces as those used in Example 2 and again, we use a multigrid preconditioner. For the computations given in Table 6.3, we use $M(n) = 3$ for $n > q+1$ and sufficiently many iterations to guarantee convergence for lower values of n. The method worked reasonably well for large k and h but failed to show improvement for smaller values. This is probably because (4.4) was not enforced.

Table 6.3 gives the discrete L^2 error as a function of the time step size k, the spatial mesh size h, and the order q. We also include the condition number $K = \kappa_1/\kappa_0$ of the preconditioned system.

TABLE 6.3
Discrete L^2 error for Example 3 with $M(n) = 3$.

k	h	$q=2$	$q=3$	$q=4$	K
1/10	1/16	.0036	.0079	.010	2.2
1/20	1/32	.0011	.0024	.0025	2.3
1/40	1/64	.0010	.0011	.0011	2.4
1/80	1/64	.00094	.00092	.00090	2.4

The final table gives convergence results when the number of iterations for the solution of (3.1) was given by

$$M(n) = 3 + 10\log_2(t_n^{-1}) \quad \text{for } n > q+1. \tag{6.10}$$

This choice of $M(n)$ satisfies (4.4) for some δ. Note that we get improved convergence results as long as we decrease h. Moreover, for this example, we see no improvement with larger q or smaller k with h fixed. This seems to suggest that for these runs, the error due to spatial discretization is the dominant term.

TABLE 6.4
Discrete L^2 error for Example 3 with $M(n)$ given by (6.10).

k	h	$q=2$	$q=3$	$q=4$	K
1/10	1/16	.0039	.0080	.010	2.2
1/20	1/32	.0010	.0021	.0024	2.3
1/40	1/64	.00034	.00054	.00056	2.4
1/80	1/64	.00048	.00053	.00054	2.4

Cornell University
Ithaca, New York 14853
E-mail: bramble@mathvax.msi.cornell.edu

Brookhaven National Laboratory
Upton, New York 11973
E-mail: pasciak@bnl.gov

BP Canada
Oil and Gas Division
Calgary, Alberta Canada

Chalmers University of Technology
S-41296 Göteborg, Sweden
E-mail: cmsvt@seguc21.bitnet

1. G. A. BAKER, J. H. BRAMBLE & V. THOMÉE, "Single step Galerkin approximations for parabolic problems," *Math. Comp.*, v. 31, 1977, pp. 818–847.

2. J. H. BRAMBLE, "Discrete methods for parabolic equations with time-dependent coefficients," in *Numerical Methods for PDE's*, Academic Press, New York, 1979, pp. 41–52.

3. J. H. BRAMBLE, R. E. EWING, J. E. PASCIAK & A. H. SCHATZ, "A preconditioning technique for the efficient solution of problems with local grid refinement," *Comput. Methods Appl. Mech. Engrg.*, v. 67, 1988, pp. 149–159.

4. J. H. BRAMBLE & J. E. PASCIAK, "New convergence estimates for multigrid algorithms," *Math. Comp.*, v. 49, 1987, pp. 311–329.

5. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "An iterative method for elliptic problems on regions partitioned into substructures," *Math. Comp.*, v. 46, 1986, pp. 361–369.

6. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "The construction of preconditioners for elliptic problems by substructuring, I," *Math. Comp.*, v. 47, 1986, pp. 103–134.

7. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "The construction of preconditioners for elliptic problems by substructuring, II," *Math. Comp.*, v. 49, 1987, pp. 1–16.

8. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "The construction of preconditioners for elliptic problems by substructuring, III," *Math. Comp.*, v. 51, 1988, pp. 415–430.

9. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "The construction of preconditioners for elliptic problems by substructuring, IV," *Math. Comp.* (To appear.)

10. J. H. BRAMBLE & P. H. SAMMON, "Efficient higher order single step methods for parabolic problems: Part I," *Math. Comp.*, v. 35, 1980, pp. 655–677.

11. J. H. BRAMBLE, A. H. SCHATZ, V. THOMÉE & L. B. WAHLBIN, "Some convergence estimates for Galerkin type approximation for parabolic equations," *SIAM J. Numer. Anal.*, v. 14, 1977, pp. 218–241.

12. J. H. BRAMBLE & V. THOMÉE, "Discrete time Galerkin methods for a parabolic boundary value problem," *Ann. Mat. Pura Appl.*, v. 101, 1974, pp. 115–152.

13. R. CHANDRA, *Conjugate Gradient Methods for Partial Differential Equations*, Yale University, Dept. of Comp. Sci. Rep. No. 129, 1978.

14. M. CROUZEIX & P. A. RAVIART, "Approximation d'équations d'évolution linéaires par des méthodes multi-pas," in *Étude Numérique des Grands Systèmes*, Proc. Sympos. Novosibirsk, Dunod, Paris, 1978, pp. 133–150.

15. C. W. CRYER, "On the instability of high order backward-difference multistep methods," *BIT*, v. 12, 1972, pp. 17–25.

16. J. DOUGLAS, JR., T. DUPONT & R. EWING, "Incomplete iterations for time-stepping a Galerkin method for a quasilinear parabolic problem," *SIAM J. Numer. Anal.*, v. 16, 1979, pp. 503–522.

17. T. DUPONT, R. P. KENDALL & H. H. RACHFORD, "An approximate factorization procedure for solving self-adjoint elliptic difference equations," *SIAM J. Numer. Anal.*, v. 5, 1968, pp. 559–573.

18. C. W. GEAR, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, N. J., 1971.

19. F. B. HILDEBRAND, *Introduction to Numerical Analysis*, McGraw-Hill, New York, 1956.

20. S. L. KEELING, "Galerkin/Runge-Kutta discretizations for parabolic equations with time dependent coefficients," Preprint, 1987.

21. M.-N. LE ROUX, "Semidiscretization in time for parabolic problems," *Math. Comp.*, v. 33, 1979, pp. 919–931.

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

23. W. M. PATTERSON, 3RD, *Iterative Methods for the Solution of a Linear Operator Equation in Hilbert Space—A Survey*, Lecture Notes in Math., vol. 394, Springer-Verlag, New York, 1974.

24. V. THOMÉE, *Galerkin Finite Element Methods for Parabolic Problems*, Lecture Notes in Math., vol. 1054, Springer-Verlag, New York, 1984.

25. H. YSERENTANT, "On the multi-level splitting of finite element spaces," *Numer. Math.*, v. 49, 1986, pp. 379–412.

26. M. ZLÁMAL, "Finite element multistep discretizations of parabolic boundary value problems," *Math. Comp.*, v. 29, 1975, pp. 350–359.