MATHEMATICS OF COMPUTATION
VOLUME 39, NUMBER 159
JULY 1982, PAGES 53–68

Time Discretization in the Backward Solution of Parabolic Equations. I*

By Lars Eldén

Abstract. The problem of solving a parabolic partial differential equation backwards in time by a method related to the Tikhonov-Phillips regularization method is considered. Time discretizations based on Padé approximations of the exponential function are studied, and a priori estimates of the step length are given, which guarantee an almost optimal error bound. The computational efficiency of different discretizations is discussed. Some numerical examples are given.

In Part II of this paper we study the backward beam method, and the same error estimates are obtained. A new scheme for time descretization based on Padé approximation is discussed.

1. Introduction. Consider the problem of solving a parabolic partial differential equation backwards in time. For convenience we write the equation in the following abstract form

$$\begin{cases} u_t = -Lu, & 0 \leqslant t \leqslant 1, \\ u(1) = w. \end{cases} \tag{1.1}$$

Here $w(x)$ is a given function in $L^2(\Omega)$, and Ω is a bounded domain in R^n with a smooth boundary $\partial\Omega$. L is the unbounded, nonnegative operator in $L^2(\Omega)$ corresponding to a selfadjoint, elliptic boundary value problem in Ω with zero Dirichlet data on $\partial\Omega$. The coefficients of L are assumed to be smooth and independent of time.

It is well known that (1.1) is ill-posed in the sense that the solution does not depend continuously on the data. One possible way to overcome this difficulty is to impose a bound on the solution at $t = 0$ and at the same time allow for some imprecision in the data. Thus we are led to the following constrained problem.

Find any solution of

$$\begin{cases} u_t = -Lu, & 0 < t \leqslant 1, \\ \|u(1) - w\| \leqslant \delta, \\ \|u(0)\| \leqslant M, \end{cases} \tag{1.2}$$

where the norm is the $L^2(\Omega)$-norm, and δ and M are given positive constants, $\delta \ll M$. Using logarithmic convexity [1], [11, p. 11], it is easy to show that any two solutions of (1.2), u_1 and u_2, satisfy

$$\|u_1(t) - u_2(t)\| \leqslant 2\delta^t M^{1-t}. \tag{1.3}$$

Thus for $0 < t \leqslant 1$ we have continuous dependence on the data.

Received April 30, 1981; revised October 8, 1981.
1980 *Mathematics Subject Classification.* Primary 65M30.
*This work was supported by the Swedish Natural Science Research Council.

0025-5718/81/0000-0441/$04.25

It is difficult to solve (1.2), one reason being that, in general, solutions are not unique. There are methods for approximating solutions of (1.2), which are optimal in the sense that Hölder type error estimates (1.3) can be obtained for them.

We consider two such methods: firstly a method related to the *regularization method* of Tikhonov and Phillips [14], [12], [5], and secondly the *backward beam method* of Buzbee, Carasso [2]. These methods are discussed in Parts I and II of this paper, respectively.

In the regularization method an approximate solution of (1.2) is given by

$$\begin{cases} v(t) = (\exp(-L) + \mu(t)I)^{-1}\exp(-Lt)w, \\ \mu(t) = (\delta/M)(1-t)/t. \end{cases} \tag{1.4}$$

Let u denote any solution of (1.2). Then, for $0 \leqslant t \leqslant 1$,

$$\|u(t) - v(t)\| \leqslant \delta^t M^{1-t}. \tag{1.5}$$

This result is due to Strakhov [13] (see also [9]). The proof is quite simple and we repeat it in Section 2.

We now raise the following question. Can we discretize (1.4) in such a way that for the discrete approximation v_a we get an error estimate of the type (1.5)

$$\|u(t) - v_a(t)\| \leqslant C\delta^t M^{1-t}, \tag{1.6}$$

for some constant C?

The answer to this question will have significance for the possibilities of solving numerically problems in two (or more) space dimensions, with nonrectangular geometry or nonconstant coefficients, since for such problems we must discretize in time and space.

In this paper we give a partial answer to the above question. We consider approximating the exponential function in (1.4) in a way which corresponds to a time discretization. In Section 3 we show that if $\exp(-\lambda)$ is approximated well enough for $0 \leqslant \lambda \leqslant \log(M/\delta)$, we can get error estimates of the form (1.6) with $C = 2$.

The results of Section 3 are used in Section 4 in connection with a class of approximations

$$e^{-\lambda} \approx \big(Q(\lambda/N)/P(\lambda/N)\big)^N, \tag{1.7}$$

where $Q(z)/P(z)$ is a Padé approximation of e^{-z}. Note that, e.g., the backward Euler and Crank-Nicolson approximations are members of this class. We derive explicit, a priori estimates of the largest step length in time $k = 1/N$, which ensures that (1.6) holds. It is shown that higher order approximations allow a much larger step length than, e.g., the low order backward Euler approximation.

In Section 5 we briefly discuss the efficiency of different Padé approximations. It is shown that the solution of the time-discrete problem can be obtained by solving a sequence of equations of the type

$$\big(\alpha_i L^2 + \beta_i L + \gamma_i I\big)v_i = w_i. \tag{1.8}$$

The number of equations (1.8) that have to be solved is taken as a measure of the efficiency of a Padé approximation.

It turns out that higher order Padé approximations are more efficient than low order approximations. This is also verified numerically.

In Part II of this paper we consider the backward beam method, and we obtain the same error estimates for the time-discrete versions of that method. Our scheme for time discretization is based on Padé approximations and it is conceptually different from that in [2].

Some numerical results for both methods are given in Part II.

The problem of solving a parabolic equation backward in time is also discussed in [8], where a completely different approach is made.

Unless otherwise stated the norm $\|\cdot\|$ is the $L^2(\Omega)$-norm. Throughout we shall write $\exp(-Lt)$ to denote an element of the strongly continuous semigroup generated by L (see, e.g., [6]). The semigroup is easily defined in terms of the spectral representation of L. Also from the spectral representation it is seen that for $\mu(t) > 0$ the operator $(\exp(-L) + \mu(t)I)$ has a bounded inverse, so that (1.4) is well defined.

I am indebted to Professor V. A. Morozov for making me aware of the paper by Strakhov [13].

2. Error Estimate for the Regularization Method (1.4). In this section we shall show that the estimate (1.5) holds for the regularization method (1.4). The proof is quite simple and we shall use the same technique in connection with discretizations of (1.4). We shall also show that the same error estimate is valid if we use (1.4) in a step-by-step manner.

Throughout we shall assume that δ and M have been chosen so that there exist solutions of (1.2).

THEOREM 2.1 (STRAKHOV [13], SEE ALSO [9]). *Let $u(t)$ denote an arbitrary solution of* (1.2), *and for $0 \leq t \leq 1$ let $v(t)$ be defined by* (1.4). *Then*

$$\|u(t) - v(t)\| \leq \delta^t M^{1-t}. \tag{2.1}$$

Proof. The assumption about the existence of solutions of (1.2) is equivalent to there being functions u_0 and Ψ such that (see, e.g., [4])

$$\|u_0\| \leq M, \quad \|\Psi\| \leq \delta, \quad w = \exp(-L)u_0 + \Psi. \tag{2.2}$$

Putting $u(t) = \exp(-Lt)u_0$, we get

$$\begin{aligned}
&\|u(t) - v(t)\| \\
&\quad = \|\exp(-Lt)u_0 - (\exp(-L) + \mu(t)I)^{-1}\exp(-Lt)(\exp(-L)u_0 + \Psi)\| \\
&\quad \leq \|\exp(-Lt) - (\exp(-L) + \mu(t)I)^{-1}\exp(-L(t+1))\|\,\|u_0\| \\
&\qquad + \|(\exp(-L) + \mu(t)I)^{-1}\exp(-Lt)\|\,\|\Psi\|,
\end{aligned}$$

where the operator norm is defined $\|A\| = \sup\{\|Au\|: \|u\| = 1\}$. We now use (2.2) and the fact that L is selfadjoint and nonnegative to get

$$\|u(t) - v(t)\| \leq \sup_{\lambda \geq 0} A(\lambda)M + \sup_{\lambda \geq 0} B(\lambda)\delta, \tag{2.3}$$

where

$$\begin{cases} A(\lambda) = \left| e^{-\lambda t} - \dfrac{e^{-\lambda(1+t)}}{e^{-\lambda} + \mu(t)} \right| = \mu(t)B(\lambda), \\ B(\lambda) = \dfrac{e^{-\lambda t}}{e^{-\lambda} + \mu(t)}. \end{cases}$$

Now for $0 \leq p, t \leq 1$ the inequality

$$p^t(1-p)^{1-t} \leq t^t(1-t)^{1-t} \tag{2.4}$$

is valid, and, putting $p = \exp(-\lambda)/(\exp(-\lambda) + \mu(t))$, we get

$$B(\lambda) = p^t(1-p)^{1-t}(\mu(t))^{t-1} \leq t^t(1-t)^{1-t}(\mu(t))^{t-1} = t(M/\delta)^{1-t},$$

(remember the definition (1.4) of $\mu(t)$). Thus we can estimate (2.3)

$$\begin{aligned}\|u(t) - v(t)\| &\leq \mu(t)t(M/\delta)^{1-t}M + t(M/\delta)^{1-t}\delta \\ &= (1-t)\delta^t M^{1-t} + t\delta^t M^{1-t} = \delta^t M^{1-t}. \quad \text{Q.E.D.}\end{aligned}$$

In view of the estimate (1.3) we cannot hope for a better error estimate than (2.1) for the regularization method (1.4), so that in this sense the method (1.4) is optimal.

The numerical solution of a *forward* parabolic problem is usually computed by a marching procedure, i.e., a procedure which is recursive in time. We next show that the method (1.4) for the backward problem can be generalized to a recursive formula in such a way that the procedure remains optimal in the above sense.

Make a (possibly nonuniform) partitioning of the interval [0, 1]

$$0 < t_1 < t_2 < \cdots < t_s < 1,$$

and define the recursion

$$\text{(2.5a)} \quad \begin{cases} v_s = v(t_s), \\ v_{i-1} = (\exp(-Lt_i) + \mu_i I)^{-1}\exp(-Lt_{i-1})v_i, \qquad i = s, s-1, \ldots, 2, \end{cases}$$

where $v(t_s)$ is given by (1.4) and

$$\text{(2.5b)} \quad \begin{cases} \mu_i = (\delta_i/M)(t_i - t_{i-1})/t_{i-1}, \\ \delta_i = \delta^{t_i} M^{1-t_i}. \end{cases}$$

COROLLARY 2.2. *Let $u(t)$ denote an arbitrary solution of* (1.2), *and let $(v_i)_{i=1}^s$ be defined by* (2.5). *Then*

$$\|u(t_i) - v_i\| \leq \delta^{t_i} M^{1-t_i}. \tag{2.6}$$

Proof. The result is obviously true for $i = s$. Then assume that it is true for $i = k$, and consider

$$u_t = -Lu \quad \text{for } 0 < t \leq t_k, \quad \|u(t_k) - v_k\| \leq \delta_k, \quad \|u(0)\| \leq M.$$

The recursion formula (2.5) is a straightforward generalization of (1.4) to the interval $[0, t_k]$, and, putting $\tau_k = t_{k-1}/t_k$, we then get

$$\|u(t_{k-1}) - v_{k-1}\| \leq \delta_k^{\tau_k} M^{1-\tau_k} = \delta^{t_{k-1}} M^{1-t_{k-1}} = \delta_{k-1}. \quad \text{Q.E.D.}$$

We conclude this section with some general remarks on the backwards problem (1.1) and on the formula (1.4).

Considering (1.1) as an ordinary differential equation on a Hilbert space, we can write the solution formally as

$$u(t) = \exp(L(1-t))w. \tag{2.7}$$

If we try to solve (1.1) numerically simply by applying a standard marching procedure for parabolic equations (backwards in time), then effectively we are trying to approximate (2.7). This will of course give a meaningless result, since the large

eigenvalues of L (the eigenvalues of L tend to plus infinity) will cause perturbations of the data to blow up catastrophically.

However, if we assume that the main part of the information about the solution of (1.1) is connected with the small eigenvalues of L, then we can solve our problem by approximating the exponential function in (2.7) by a function $g(\lambda, t)$ such that $g(\lambda, t) \approx \exp(\lambda(1 - t))$ for small λ, and $g(\lambda, t)$ is bounded for large λ.

In the regularization method (1.4) we have

$$g(\lambda, t) = (\exp(-\lambda) + \mu(t))^{-1}\exp(-\lambda t). \tag{2.8}$$

In Figure 2.1 we have plotted this function for a few values of t and $M/\delta = 10^6$. Note that for fixed t, $g(\lambda, t)$ has its maximum equal to $t \cdot (M/\delta)^{1-t}$ for $\lambda = \log(M/\delta)$. This observation will be of significance later when we approximate the exponential function in (2.8).

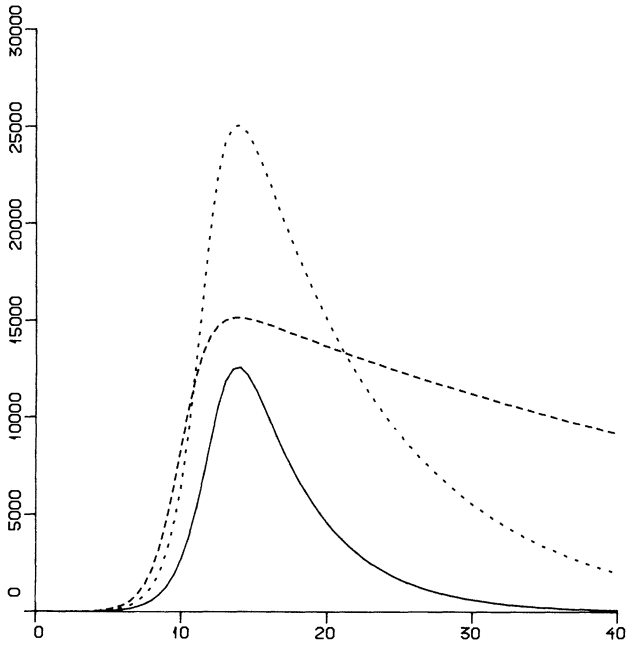

FIGURE 2.1

The function $g(\lambda, t)$ *defined by* (2.8) *plotted for* $t = 0.2$ (*solid line*), $t = 0.1$ (*dotted*), *and* $t = 0.02$ (*dashed*). $M/\delta = 10^6$.

3. Preliminary Error Estimates for the Approximate Method. In this section we derive error estimates for

$$v_a(t) = (f(L) + \mu(t)I)^{-1}(f(L))^t w, \tag{3.1}$$

which is (1.4) with the exponential function replaced by an approximation f, such that $f(\lambda) \approx e^{-\lambda}$.

In the next section $f(\lambda)$ will depend on N, where $k = 1/N$ is a step length parameter, but here this dependence is suppressed. There we shall be dealing

explicitly with the class of approximations defined by (1.7), but in this section it will be sufficient to distinguish between two subclasses characterized by the following inequalities

$$\text{(3.2)} \qquad \text{(i)} \qquad e^{-\lambda} \leqslant f(\lambda) \leqslant 1, \qquad \lambda \geqslant 0,$$

$$\text{(3.3a)} \qquad \text{(ii)} \qquad \begin{cases} 0 < f(\lambda) \leqslant e^{-\lambda}, & 0 \leqslant \lambda \leqslant \log(M/\delta), \\ 0 \leqslant f(\lambda) \leqslant 1, & \lambda \geqslant \log(M/\delta). \end{cases}$$

(3.3b)

First we give an error estimate for approximations satisfying (3.2).

THEOREM 3.1. *Let $u(t)$ denote an arbitrary solution of* (1.2), *let $v_a(t)$ be defined by* (3.1), *and assume that f satisfies* (3.2). *If*

$$\text{(3.4)} \qquad \lambda + \log f(\lambda) \leqslant (\delta/M)\frac{1}{t}e^{\lambda t} \quad \textit{for } 0 \leqslant \lambda \leqslant \log(M/\delta),$$

then

$$\text{(3.5)} \qquad \|u(t) - v_a(t)\| \leqslant (t + \max(1, 2(1-t)))\delta^t M^{1-t}.$$

Remark. The error estimate depends on how well $e^{-\lambda}$ is approximated by $f(\lambda)$. Theorem 3.1 shows that $f(\lambda)$ need only be a good approximation for $0 \leqslant \lambda \leqslant \log(M/\delta)$. This is not surprising in view of the remarks made at the end of Section 2. Note that the assumption (3.2) implies that $0 \leqslant \lambda + \log f(\lambda)$. If $f(\lambda)$ is a good approximation of $e^{-\lambda}$, then $\lambda + \log f(\lambda)$ is small.

Proof. Using the arguments of the proof of Theorem 2.1, we see that

$$\text{(3.6)} \qquad \|u(t) - v_a(t)\| \leqslant \sup_{\lambda \geqslant 0} A(\lambda) M + \sup_{\lambda \geqslant 0} B(\lambda)\delta,$$

where now

$$\text{(3.7)} \qquad \begin{cases} A(\lambda) = \left| e^{-\lambda t} - \dfrac{(f(\lambda))^t}{f(\lambda) + \mu(t)} e^{-\lambda} \right|, \\ B(\lambda) = \dfrac{(f(\lambda))^t}{f(\lambda) + \mu(t)}. \end{cases}$$

Using the inequality (2.4), we immediately get

$$\text{(3.8)} \qquad B(\lambda) \leqslant t(M/\delta)^{1-t}.$$

To estimate $A(\lambda)$ we first show that under the assumption (3.2)

$$A(\lambda) = e^{-\lambda t} - \frac{(f(\lambda))^t}{f(\lambda) + \mu(t)} e^{-\lambda}, \qquad \lambda \geqslant 0.$$

Since $\mu(t) \geqslant 0$ for $0 < t \leqslant 1$, we have

$$\frac{(f(\lambda))^t}{f(\lambda) + \mu(t)} \leqslant (f(\lambda))^{t-1} \leqslant e^{\lambda(1-t)},$$

which gives

$$e^{-\lambda t} - \frac{(f(\lambda))^t}{f(\lambda) + \mu(t)} e^{-\lambda} \geqslant e^{-\lambda t} - e^{\lambda(1-t)} e^{-\lambda} = 0.$$

Therefore, for large λ we can estimate

$$A(\lambda) \leqslant e^{-\lambda t} \leqslant (\delta/M)^t, \quad \text{for } \lambda \geqslant \log(M/\delta). \tag{3.9}$$

To estimate $A(\lambda)$ for $0 \leqslant \lambda \leqslant \log(M/\delta)$, we write

$$\begin{aligned} A(\lambda)(f(\lambda) + \mu(t)) &= \mu(t)e^{-\lambda t} + f(\lambda)e^{-\lambda t} - (f(\lambda))^t e^{-\lambda} \\ &= \mu(t)e^{-\lambda t} + (f(\lambda))^t e^{-\lambda t}\big((f(\lambda))^{1-t} - e^{-\lambda(1-t)}\big). \end{aligned}$$

Here both terms are positive. Using the mean value theorem and the assumption (3.2) it is easy to show that

$$(f(\lambda))^{1-t} - e^{-\lambda(1-t)} \leqslant (\log f(\lambda) + \lambda)(1-t),$$

and then, using (3.4), we can estimate

$$\begin{aligned} A(\lambda)(f(\lambda) + \mu(t)) &\leqslant \mu(t)e^{-\lambda t} + (f(\lambda))^t e^{-\lambda t}(1-t)/t(\delta/M)e^{\lambda t} \\ &= \mu(t)\big(e^{-\lambda t} + (f(\lambda))^t\big) \leqslant 2\mu(t)(f(\lambda))^t, \end{aligned}$$

where the last inequality follows from (3.2). Now by (2.4)

$$A(\lambda) \leqslant 2(\mu(t))^t t^t (1-t)^{1-t} = 2(1-t)(\delta/M)^t, \tag{3.10}$$

for $0 \leqslant \lambda \leqslant \log(M/\delta)$. Therefore, combining (3.9), (3.10) and (3.8), we get the desired result. Q.E.D.

We next give the corresponding theorem for approximations satisfying (3.3).

THEOREM 3.2. *Let $u(t)$ denote an arbitrary solution of* (1.2), *let $v_a(t)$ be defined by* (3.1), *and assume that f satisfies* (3.3). *If*

$$-\lambda - \log f(\lambda) \leqslant (\delta/M)\frac{\log 2e^{\lambda}}{t} \quad \text{for } 0 \leqslant \lambda \leqslant \log(M/\delta), \tag{3.11}$$

then

$$\|u(t) - v_a(t)\| \leqslant \big(t + \max(1, 2(1-t))\big)\delta^t M^{1-t}. \tag{3.12}$$

Remark. Note that the assumption (3.3) implies that $-\lambda - \log f(\lambda)$ is nonnegative.

Proof. As in the proof of Theorem 3.1 we get

$$\|u(t) - v_a(t)\| \leqslant \sup_{\lambda \geqslant 0} A(\lambda) \cdot M + \sup_{\lambda \geqslant 0} B(\lambda) \cdot \delta,$$

where $A(\lambda)$ and $B(\lambda)$ are given by (3.7), and $B(\lambda)$ can be estimated by (3.8).

We now show that under the assumptions (3.3) and (3.11) $a(\lambda)$ defined by

$$a(\lambda) = e^{-\lambda t} - \frac{(f(\lambda))^t}{f(\lambda) + \mu(t)}e^{-\lambda}, \qquad \lambda \geqslant 0,$$

is nonnegative. First we consider $\lambda \geqslant \log(M/\delta)$. The inequality $a(\lambda) \geqslant 0$ is equivalent to

$$\frac{(f(\lambda))^t}{f(\lambda) + \mu(t)} \leqslant e^{\lambda(1-t)}. \tag{3.13}$$

By (2.4) we have

$$\frac{(f(\lambda))^t}{f(\lambda) + \mu(t)} \leqslant t(M/\delta)^{1-t}$$

and for $\lambda \geq \log(M/\delta)$

$$e^{\lambda(1-t)} \geq (M/\delta)^{1-t},$$

so that (3.13) is satisfied. This means that for large λ we can estimate

$$A(\lambda) \leq e^{-\lambda t} \leq (\delta/M)^t \quad \text{for } \lambda \geq \log(M/\delta).$$

Then we consider $0 \leq \lambda \leq \log(M/\delta)$. $a(\lambda) \geq 0$ is equivalent to

$$(f(\lambda))^t e^{-\lambda t}\left(e^{-\lambda(1-t)} - (f(\lambda))^{1-t}\right) \leq \mu(t) e^{-\lambda t},$$

and, using the mean value theorem and (3.11), the left-hand side can be estimated

$$\begin{aligned}(f(\lambda))^t e^{-\lambda t}\left(e^{-\lambda(1-t)} - (f(\lambda))^{1-t}\right) &\leq (f(\lambda))^t e^{-\lambda}(1-t)(-\lambda - \log(f(\lambda))) \\ &\leq (f(\lambda))^t e^{-\lambda}(1-t)\frac{\log 2}{t} e^{\lambda}\delta/M = \log 2\mu(t)(f(\lambda))^t \\ &\leq \mu(t) e^{-\lambda t},\end{aligned}$$

where the last inequality follows by the assumption (3.3). Thus $a(\lambda) \geq 0$.

We now have

$$A(\lambda)(f(\lambda) + \mu(t)) = \mu(t) e^{-\lambda t} + (f(\lambda))^t e^{-\lambda t}\left((f(\lambda))^{1-t} - e^{-\lambda(1-t)}\right).$$

Here the second term on the right-hand side is nonpositive, and we can estimate

$$A(\lambda)(f(\lambda) + \mu(t)) \leq \mu(t) e^{-\lambda t}. \tag{3.14}$$

Now, since for $0 \leq \lambda \leq \log(M/\delta)$

$$e^{\lambda}\delta/M \leq 1,$$

the assumption (3.11) gives

$$-\lambda - \log(f(\lambda)) \leq \frac{1}{t}\log 2$$

or, equivalently,

$$e^{-\lambda t} \leq 2(f(\lambda))^t.$$

If we use this in (3.14), we get

$$A(\lambda) \leq 2\mu(t)\frac{(f(\lambda))^t}{f(\lambda) + \mu(t)} \leq 2(1-t)(\delta/M)^t,$$

where the last inequality follows by (2.4). Combining the estimates of $A(\lambda)$ and $B(\lambda)$, we get (3.12). Q.E.D.

Note that it is possible to get the error estimate (3.12) under somewhat less restrictive assumptions than (3.11) (essentially it is not necessary to have $a(\lambda) \geq 0$ for $0 \leq \lambda \leq \log(M/\delta)$). However, we have chosen this form in order to get a more straightforward analysis in the next section.

In the next section we shall see that certain approximations to the exponential functions do not satisfy (3.3b) but rather

$$|f(\lambda)| \leq 1 \quad \text{for } \lambda \geq \log(M/\delta).$$

This is the case if we have

$$f(\lambda) = (h(\lambda))^N, \qquad N \text{ even}, \tag{3.15a}$$

where

$$0 < h(\lambda) \leq e^{-\lambda/N} \quad \text{for } 0 \leq \lambda \leq \log(M/\delta), \tag{3.15b}$$

$$|h(\lambda)| \leq 1 \quad \text{for } \lambda \geq \log(M/\delta). \tag{3.15c}$$

If $t = n/N$, where n is odd, then $(h(\lambda))^n$ can be negative for large λ.

In this case the error estimate is not quite as good as (3.12).

COROLLARY 3.3. *Assume that N is even and $t = n/N$, where n is an integer. Let $h(\lambda)$ satisfy* (3.15), *and assume that $f(\lambda)$ in* (3.15a) *satisfies* (3.11). *Define*

$$v_a(t) = \left((h(L))^N + \mu(t)I\right)^{-1}(h(L))^n w.$$

Then

$$\|u(t) - v_a(t)\| \leq \left(t + \max(1 + t, 2(1 - t))\right)\delta^t M^{1-t}.$$

Proof. The assumption N even is necessary for $v_a(t)$ to be well defined.

We only need to show that for $\lambda \geq \log(M/\delta)$ we have

$$0 \leq e^{-\lambda t} - \frac{(h(\lambda))^n}{(h(\lambda))^N + \mu(t)} e^{-\lambda} \leq (1 + t)e^{-\lambda t}.$$

But this is satisfied if

$$\frac{|h(\lambda)|^n}{(h(\lambda))^N + \mu(t)} \leq te^{\lambda(1-t)}, \tag{3.16}$$

and from the proof of Theorem 3.2 (3.16) can be seen to hold. Q.E.D.

4. Padé Approximations and Discretization in Time. The two theorems of Section 3 show that when the exponential function in (1.4) is replaced by an approximation it is necessary to approximate $e^{-\lambda}$ well only for small λ. This leads us naturally to considering Padé approximations, which by definition are best approximations in the neighborhood of the origin.

Assume that the interval $[0, 1]$ has been divided into N equal subintervals, put $k = 1/N$, and assume that $t = nk$ for some integer n. Put

$$f_{pq}^N(\lambda) = \left(F_{pq}(k\lambda)\right)^N, \tag{4.1a}$$

where

$$F_{pq}(\lambda) = Q_{pq}(\lambda)/P_{pq}(\lambda), \tag{4.1b}$$

and $Q_{pq}(\lambda)/P_{pq}(\lambda)$ is the Padé approximation to $e^{-\lambda}$ defined [15], [10] by

$$Q_{pq}(\lambda) = \sum_{\nu=0}^{q} \frac{(p + q - \nu)!q!}{(p + q)!\nu!(q - \nu)!}(-\lambda)^\nu, \tag{4.1c}$$

$$P_{pq}(\lambda) = \sum_{\nu=0}^{p} \frac{(p + q - \nu)!p!}{(p + q)!\nu!(p - \nu)!}\lambda^\nu. \tag{4.1d}$$

Two simple approximations of this type are

$$Q_{10}(\lambda)/P_{10}(\lambda) = 1/(1 + \lambda),$$
$$Q_{11}(\lambda)/P_{11}(\lambda) = (1 - \lambda/2)/(1 + \lambda/2),$$

which in connection with ordinary differential equations correspond to the backward Euler and trapezoidal (Crank-Nicolson) methods, respectively.

Then we approximate (1.4)

$$v^N(t) = \big((F(kL))^N + \mu(t)I\big)^{-1}(F(kL))^n w. \tag{4.2}$$

For convenience we omit the indices pq.

In this section we shall examine the inequalities (3.4) and (3.11) with $f(\lambda) = f^N(\lambda)$ and use error estimates for Padé approximations to derive lower bounds on N such that (3.4) and (3.11) are satisfied.

We first identify some properties of Padé approximations. It is easily seen that if we require that $f^N(\lambda)$ satisfies either of (3.2) and (3.3), we must restrict ourselves to the case $q \leqslant p$ [15, Lemma 2]. Further [15], [7]

$$e^{-z} = Q(z)/P(z) + (-1)^{q+1}R(z), \tag{4.3a}$$

where for some θ

$$\begin{cases} R(z) = \sigma z^{p+q+1}e^{-\theta}/P(z), & z \geqslant \theta \geqslant 0, \qquad \text{(4.3b)} \\ \sigma = \sigma_{pq} = \dfrac{p!q!}{(p+q)!\,(p+q+1)!}. & \text{(4.3c)} \end{cases}$$

Since for $z \geqslant 0$ $P(z) \geqslant 1$, $R(z)$ can be estimated

$$0 \leqslant R(z) \leqslant \sigma z^{p+q+1}e^{-\theta} \leqslant \sigma z^{p+q+1}. \tag{4.3d}$$

From (4.3) we see that for $z \geqslant 0$ and

$$q \text{ even}, \quad Q(z)/P(z) \geqslant e^{-z}, \tag{4.4a}$$

$$q \text{ odd}, \quad Q(z)/P(z) \leqslant e^{-z}. \tag{4.4b}$$

It is obvious that all the approximations (4.1) with q even, $q \leqslant p$, constitute the subclass characterized by (3.2). Similarly, the approximations with q odd are (under certain conditions) the subclass characterized by (3.3).

We now give the two theorems which correspond to Theorems 3.1 and 3.2.

THEOREM 4.1. *Let $f^N(\lambda)$ be defined by* (4.1), *and let q be even, $q \leqslant p$.*

(a) *For all $\lambda \geqslant 0$,*

$$e^{-\lambda} \leqslant f^N(\lambda) \leqslant 1.$$

(b) *Let $u(t)$ denote any solution of* (1.2), *and let $v^N(t)$ be defined by* (4.2). *If*

$$N \geqslant \max(1/t, N_1), \tag{4.5a}$$

where

$$N_1 = \log(M/\delta)\big[t\sigma \log(M/\delta)M/\delta\big]^{1/(p+q)}, \tag{4.5b}$$

then

$$\|u(t) - v^N(t)\| \leqslant \big(t + \max(1, 2(1-t))\big)\delta^t M^{1-t}. \tag{4.6}$$

Proof. (a) follows immediately from (4.4a) and the assumption $q \leqslant p$. For (b) we use (4.3) and rewrite

$$\begin{aligned} \lambda + \log f^N(\lambda) &= \lambda + N\log(Q(\lambda/N)/P(\lambda/N)) \\ &= \lambda + N\log\big(e^{-\lambda/N} + R(\lambda/N)\big) = N\log\big(1 + e^{\lambda/N}R(\lambda/N)\big). \end{aligned}$$

The inequality $\log(1+x) \leq x$ and (4.3d) now give

$$\lambda + \log f^N(\lambda) \leq Ne^{\lambda/N}R(\lambda/N) \leq Ne^{\lambda/N}\sigma(\lambda/N)^{p+q+1},$$

and further, using (4.5),

$$\begin{aligned}\lambda + \log f^N(\lambda) &\leq e^{\lambda/N}\sigma\lambda^{p+q+1}\left[(\log(M/\delta))^{p+q+1}t\sigma M/\delta\right]^{-1} \\ &= \left(\frac{\lambda}{\log(M/\delta)}\right)^{p+q+1}\frac{1}{t}\delta/Me^{\lambda/N} \leq \frac{1}{t}\delta/Me^{\lambda t},\end{aligned}$$

where the last inequality is a consequence of the assumption $N \geq 1/t$ and the fact that we consider only the interval $0 \leq \lambda \leq \log(M/\delta)$. (b) now follows from Theorem 3.1. Q.E.D.

THEOREM 4.2. *Let $f^N(\lambda)$, $u(t)$, and $v^N(t)$ be defined as in Theorem* 4.1. *Assume that q is odd, $q \leq p$, N is even and $t = n/N$, where n is an even integer.*

If

$$N \geq \max(1/t, N_2), \tag{4.7a}$$

where

$$N_2 = \log(M/\delta)\left[\frac{2}{\log 2}t\sigma\log(M/\delta)M/\delta\right]^{1/(p+q)}, \tag{4.7b}$$

then

$$\text{(a)} \quad \begin{cases} 0 < f^N(\lambda) \leq e^{-\lambda} & \text{for } 0 \leq \lambda \leq \log(M/\delta), \quad (4.8a) \\ 0 \leq f^N(\lambda) \leq 1 & \text{for } \lambda \geq \log(M/\delta), \quad (4.8b) \end{cases}$$

and

$$\text{(b)} \quad \|u(t) - v^N(t)\| \leq (t + \max(1, 2(1-t)))\delta^t M^{1-t}. \tag{4.9}$$

Proof. From (4.4b) we see that $Q(z)$ has a zero for some positive z. In order that $f^N(\lambda) \geq 0$ we must then have N even (note that v^N is not well defined if q and N are odd), and to be able to use Theorem 3.2 we must also have n even. Now if we can show that $e^{\lambda/N}R(\lambda/N) < 1$ for $0 \leq \lambda \leq \log(M/\delta)$, we see that $f^N(\lambda) > 0$ for $0 \leq \lambda \leq \log(M/\delta)$, since by (4.3) we then have

$$(f^N(\lambda)) = (Q(\lambda/N)/P(\lambda/N))^N = e^{-\lambda}(1 - e^{\lambda/N}R(\lambda/N))^N > 0.$$

We now show that (4.7) implies that for $0 \leq \lambda \leq \log(M/\delta)$

$$e^{\lambda/N}R(\lambda/N) \leq \frac{\log 2}{2} < 1/2. \tag{4.10}$$

Using (4.3d) and (4.7), we get

$$\begin{aligned}e^{\lambda/N}R(\lambda/N) &\leq e^{\lambda/N}\sigma(\lambda/N)^{p+q+1} \\ &\leq e^{\lambda/N}\sigma\lambda^{p+q+1}\frac{1}{N}\left[\frac{2}{\log 2}t\sigma(\log(M/\delta))^{p+q+1}M/\delta\right]^{-1} \\ &= e^{\lambda/N}\frac{\log 2}{2}\frac{1}{Nt}\left(\frac{\lambda}{\log(M/\delta)}\right)^{p+q+1}\delta/M.\end{aligned}$$

Since we consider $0 \leq \lambda \leq \log(M/\delta)$ and since $N \geq 1/t$, (4.10) follows immediately, and we have proved (4.8).

In the same way as in the proof of Theorem 4.1, we get

$$-\lambda - \log f^N(\lambda) = N \log\bigl(1 - e^{\lambda/N} R(\lambda/N)\bigr)^{-1},$$

(we only consider this for $0 \leq \lambda \leq \log(M/\delta)$, where $f^N(\lambda) > 0$). Using (4.10) and the inequality

$$\frac{1}{1-x} = 1 + \frac{x}{1-x} \leq e^{x/(1-x)},$$

we estimate

$$-\lambda - \log f^N(\lambda) \leq N \frac{e^{\lambda/N} R(\lambda/N)}{1 - e^{\lambda/N} R(\lambda/N)} < 2N e^{\lambda/N} R(\lambda/N).$$

In the same way as above, we now get

$$-\lambda - \log f^N(\lambda) \leq \frac{1}{t} \log 2 e^{\lambda/N} \delta/M \leq \frac{1}{t} \log 2 e^{\lambda} \delta/M,$$

and (b) now follows from Theorem 3.2. Q.E.D.

COROLLARY 4.3. *If in Theorem* 4.2 *the restriction n even is dropped, the error estimate becomes*

$$\|u(t) - v^N(t)\| \leq \bigl(t + \max(1 + t, 2(1 - t))\bigr) \delta^t M^{1-t}.$$

In Tables 4.1–4.4 we give the values of N_1 and N_2 for a few values of M/δ and t. Even though from (4.5b) and (4.7b) it appears that approximations with q even are better than those with q odd (the constant $2/\log 2$ is missing in (4.5b)), the tables show that this is of almost no practical significance. For both classes of approximations increasing p and q leads to a drastically lower value of N, and soon the requirement $N \geq 1/t$ becomes the most restrictive. Note that for q odd, N is also restricted by the requirement that N must be even; e.g., for $t = 0.2$ we cannot have N smaller than 10.

TABLE 4.1

N_1 and N_2 defined by (4.5b) *and* (4.7b) *are given for $M/\delta = 10^6$, $t = 0.2$*

q	0	1	2	3	4	5	6	7	8
p									
1	$\approx 19 \cdot 10^6$	11262							
2	9376	664	109						
3	673	157	46	29					
4	171	65	26	18	11				
5	72	35	17	13	8	7			
6	40	23	12	10	7	6	5		
7	26	16	10	8	6	5	4	4	
8	18	12	8	6	5	4	4	3	3

TABLE 4.2

N_1 and N_2 given for $M/\delta = 10^6$, $t = 0.5$

q / p	0	1	2	3	4	5	6	7	8	9	10
1	$\approx 47 \cdot 10^6$	17806									
2	14824	901	137								
3	913	198	55	34							
4	214	78	30	21	12						
5	87	41	19	14	9	8					
6	47	26	14	11	7	6	5				
7	29	18	10	8	6	5	4	4			
8	20	14	8	7	5	5	4	4	3		
9	15	11	7	6	5	4	3	3	3	3	
10	12	9	6	5	4	4	3	3	3	2	2
11	10	7	5	5	4	3	3	3	2	2	2
12	8	6	5	4	3	3	3	3	2	2	2

TABLE 4.3

N_1 and N_2 given for $M/\delta = 10^4$, $t = 0.2$

q / p	0	1	2	3	4	5	6
1	84831	613					
2	511	84	21				
3	85	30	12	9			
4	33	16	8	6	4		
5	18	11	6	5	3	3	
6	12	8	5	4	3	3	2

TABLE 4.4

N_1 and N_2 given for $M/\delta = 10^4$, $t = 0.5$

q / p	0	1	2	3	4	5	6	7	8
1	212076	970							
2	807	113	27						
3	115	38	14	10					
4	41	19	9	7	5				
5	22	12	7	5	4	3			
6	14	9	5	4	3	3	2		
7	10	7	4	4	3	3	2	2	
8	8	5	4	3	3	2	2	2	2

The price that must be paid for a smaller value of N is, of course, a more complicated approximation, and it is seen that the work for computing v^N depends also on p and q. This question is discussed in the next section.

5. Efficiency Considerations. In this section we briefly discuss how to compute $v^N(t)$ from (4.2). We still assume that the problem has not been discretized in space.

Using (4.1) and putting $k = 1/N$, $t = nk$, we can rewrite,

$$(5.1) \qquad \big(Q^N(kL) + \mu(t)P^N(kL)\big)v^N(nk) = Q^n(kL)P^{N-n}(kL)w.$$

On the left-hand side we have a polynomial in L of degree Np (remember $q \leq p$) and this can be factorized in quadratic factors (if N and p are odd there will be one linear factor; for simplicity we assume N even in the sequel). If we factorize also the right-hand side of (5.1), we can solve (5.1) by a recursion

$$\text{(5.2)} \qquad \begin{cases} z_0 = w, \\ (\alpha_i L^2 + \beta_i L + \gamma_i I) z_i = S_i(L) z_{i-1}, & i = 1, 2, \ldots, Np/2, \\ v^N(nk) = z_{Np/2}, \end{cases}$$

where the S_i are quadratic or linear polynomials.

Thus, if L is a second order elliptic operator, $v^N(nk)$ can be computed essentially by solving a sequence of $Np/2$ fourth order elliptic equations.

To compare the different time discretizations considered in Section 4, we take Np as a measure of their efficiency, where, for given values of p, q, t and M/δ, N is the smallest integer which satisfies the conditions of Theorems 4.1 or 4.2. Here it is important to note that the conditions of these theorems are only *sufficient* conditions. Therefore the approximation with the smallest value of Np need not be the optimal approximation. This will be verified numerically.

In Tables 5.1–5.4 we give the values of Np for different Padé approximations and for the same values of M/δ and t as in Tables 4.1–4.4.

TABLE 5.1

Np given for $M/\delta = 10^6$, $t = 0.2$

q	0	1	2	3	4	5	6	7	8
p									
1	$19 \cdot 10^6$	11270							
2	18760	1340	220						
3	2025	480	150	90					
4	700	280	120	80	60				
5	375	200	100	100	50	50			
6	240	180	90	60	60	60	30		
7	210	140	70	70	70	70	35	70	
8	160	160	80	80	40	80	40	80	40

TABLE 5.2

Np given for $M/\delta = 10^6$, $t = 0.5$

q	0	1	2	3	4	5	6	7	8	9	10
p											
1	$47 \cdot 10^6$	17806									
2	29648	1804	276								
3	2742	594	168	102							
4	856	312	120	88	48						
5	440	210	100	70	50	40					
6	288	156	84	72	48	36	36				
7	210	126	70	56	42	42	28	28			
8	160	112	64	64	48	48	32	32	32		
9	144	108	72	54	54	36	36	36	36	36	
10	120	100	60	60	40	40	40	40	40	20	20
11	110	88	66	66	44	44	44	44	22	22	22
12	96	72	72	48	48	48	48	48	24	24	24

TABLE 5.3

Np given for $M/\delta = 10^4$, $t = 0.2$

q / p	0	1	2	3	4	5	6
1	84835	620					
2	1030	180	50				
3	255	90	45	30			
4	140	80	40	40	20		
5	100	100	50	50	25	50	
6	90	60	30	60	30	60	30

TABLE 5.4

Np given for $M/\delta = 10^4$, $t = 0.5$

q / p	0	1	2	3	4	5	6	7
1	212076	970						
2	1616	228	56					
3	348	114	42	30				
4	168	80	40	32	24			
5	110	60	40	30	20	20		
6	84	60	36	24	24	24	12	
7	70	56	28	28	28	28	14	14

In Tables 5.1 and 5.3 some adjacent values of Np differ very much. E.g., in Table 5.1 the $(6,6)$ and $(7,7)$ approximations have $Np = 30$ and $Np = 70$, respectively. This is because of the requirement that N must be even when q is odd. Such differences do not occur when $1/t$ is an even integer.

From Tables 5.1–5.4 we see that with Np as a measure of efficiency the most efficient approximations are those with $p = q$. We also see that using a higher order approximation may reduce the work substantially. This is verified numerically in Part II of this paper. However, the value of Np cannot be reduced under a certain level because of the restriction $t = n/N$.

Space discretization and the efficient solution of linear algebraic systems corresponding to (5.2) are treated in [3] for the special case when the geometry is rectangular in two dimensions and the coefficients of L are nonconstant but allow separation of variables.

Department of Mathematics
University of Linköping
S-581 83 Linköping, Sweden

1. S. AGMON & L. NIRENBERG, "Properties of solutions of ordinary differential equations in Banach space," *Comm. Pure Appl. Math.*, v. 16, 1963, pp. 121–239.

2. B. L. BUZBEE & A. CARASSO, "On the numerical computation of parabolic problems for preceding times," *Math. Comp.*, v. 27, 1973, pp. 237–266.

3. L. ELDÉN, "Regularization of the backward solution of parabolic problems," *Inverse and Improperly Posed Problems in Differential Equations* (G. Anger, ed.), Akademie-Verlag, Berlin, 1979.

4. R. E. EWING, "The approximation of certain parabolic equations backward in time by Sobolev equations," *SIAM J. Math. Anal.*, v. 6, 1975, pp. 283–294.

5. J. N. FRANKLIN, "Minimum principles for ill-posed problems," *SIAM J. Math. Anal.*, v. 9, 1978, pp. 638–650.

6. A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

7. P. M. HUMMEL & C. L. SEEBECK, JR., "A generalization of Taylor's expansion," *Amer. Math. Monthly*, v. 56, 1949, pp. 243–247.

8. P. MANSELLI & K. MILLER, "Dimensionality reduction methods for efficient numerical solution, backward in time, of parabolic equations with variable coefficients," *SIAM J. Math. Anal.*, v. 11, 1980, pp. 147–159.

9. V. A. MOROZOV, "On the restoration of functions with guaranteed accuracy," *Numerical Analysis in Fortran*, Moscow Univ. Press, Moscow, 1979, pp. 46–65. (Russian)

10. H. PADÉ, "Sur la représentation approchée d'une fonctions par des fractions rationelles," Thesis, *Ann. École Norm.* (3), v. 9, 1892.

11. L. E. PAYNE, *Improperly Posed Problems in Partial Differential Equations*, SIAM, Philadelphia, Pa., 1975.

12. D. L. PHILLIPS, "A technique for the numerical solution of certain integral equations of the first kind," *J. Assoc. Comput. Mach.*, v. 9, 1962, pp. 84–97.

13. V. N. STRAKHOV, "Solution of incorrectly-posed problems in Hilbert space," *Differential Equations*, v. 6, 1970, pp. 1136–1140. (Russian)

14. A. N. TIKHONOV, "Solution of incorrectly formulated problems and the regularization method," *Dokl. Akad. Nauk SSSR*, v. 151, 1963, pp. 501–504; English transl. in *Soviet Math. Dokl.*, v. 4, 1963, pp. 1035–1038.

15. R. S. VARGA, "On higher order stable implicit methods for solving parabolic partial differential equations," *J. Math. and Phys.*, v. 40, 1961, pp. 220–231.