IMA Journal of Numerical Analysis (2003) **23**, 269–299

A parallel method for time discretization of parabolic equations based on Laplace transformation and quadrature

DONGWOO SHEEN†
Department of Mathematics, Seoul National University, 151–747, Seoul, Korea

IAN H. SLOAN‡
School of Mathematics, The University of New South Wales, 2052, Sydney, Australia

AND

VIDAR THOMÉE§
Department of Mathematics, Chalmers University of Technology, S-412 96, Göteborg, Sweden

[Received on 8 May 2002]

We consider the discretization in time of an inhomogeneous parabolic equation in a Banach space setting, using a representation of the solution as an integral along a smooth curve in the complex left half-plane which, after transformation to a finite interval, is then evaluated to high accuracy by a quadrature rule. This reduces the problem to a finite set of elliptic equations with complex coefficients, which may be solved in parallel. The paper is a further development of earlier work by the authors, where we treated the homogeneous equation in a Hilbert space framework. Special attention is given here to the treatment of the forcing term. The method is combined with finite-element discretization in spatial variables.

Keywords: parabolic problems; parallel algorithm; Laplace transform; quadrature.

1. Introduction

In this paper, which is a further development of our earlier paper (Sheen *et al.*, 1999), we consider the approximate solution of parabolic problems of the form

$$u_t + Au = f(t), \quad \text{for } t > 0, \quad \text{with } u(0) = u_0, \tag{1.1}$$

where u_0 and $f(t)$ are given. Having in mind the case that A is a second-order elliptic differential operator with Dirichlet boundary conditions in a spatial domain Ω, we consider the problem in the framework of a Banach space $\mathbb{B}$. We assume that A is a closed operator in $\mathbb{B}$ such that $-A$ generates a bounded analytic semigroup $E(t) = \mathrm{e}^{-At}$. More precisely, we assume that the spectrum $\sigma(A)$ of A is contained in a sector of the right half-plane,

$$\sigma(A) \subset \Sigma_\delta := \{z \in \mathbb{C} : |\arg z| < \delta, z \neq 0\},$$

†Email: sheen@snu.ac.kr
‡Email: i.sloan@unsw.edu.au
§Email: thomee@math.chalmers.se

with $\delta \in (0, \pi/2)$, and that the resolvent $(z\,I + A)^{-1}$ of $-A$ satisfies

$$\|(z\,I + A)^{-1}\| \leqslant M(1 + |z|)^{-1}, \quad \text{for } z \in S = \Sigma_{\pi-\delta} \cup B, \tag{1.2}$$

with B a neighbourhood of the origin and with M independent of z. In Sheen *et al.* (1999) we restricted our attention to operators A which are symmetric and positive definite in a Hilbert space $\mathbb{H}$. In this case δ could be chosen as an arbitrary number in $(0, \pi/2)$, with $M = O(\delta^{-1})$, and this growth of M as $\delta \to 0$ then had to be taken into account in the analysis. Here, for simplicity, we shall consider δ and M fixed. For the elliptic differential operator case and $\mathbb{B} = C(\bar{\Omega})$, (1.2) was shown in Stewart (1974).

The first step in our approach is to represent the solution $u(t)$ as a contour integral of the form

$$u(t) = \frac{1}{2\pi \mathrm{i}} \int_\Gamma \mathrm{e}^{zt} w(z)\,\mathrm{d}z, \tag{1.3}$$

where, for $\mathrm{Re}\, z \geqslant \alpha$ with α sufficiently large, $w(z)$ is the Laplace transform of u,

$$w(z) := \widehat{u}(z) = \int_0^\infty \mathrm{e}^{-zt} u(t)\,\mathrm{d}t, \tag{1.4}$$

and where initially Γ is an appropriately chosen line Γ_0 in the complex plane parallel to the imaginary axis, with $\mathrm{Im}\, z$ increasing. In (1.3), u is then just the inverse Laplace transform of w. For our purposes, however, assuming that $w(z)$ may be continued analytically in an appropriate way, we shall want to take for Γ a deformed contour in the set S in (1.2), which behaves asymptotically as a pair of straight lines in the left half-plane, with $\mathrm{Re}\, z \to -\infty$ when $\mathrm{Im}\, z \to \pm\infty$, forcing the factor e^{zt} to decay exponentially towards both ends of the deformed contour. We permit Γ to cut the real axis either to the right of the origin or to the left, but sufficiently close to the origin. (In Sheen *et al.* (1999) we used $-z$ instead of z in (1.3) and (1.4), and Γ thus had to be transformed into the right half-plane rather than the left.) More precisely, we take

$$\Gamma = \{z : z = \varphi(y) + \mathrm{i}\sigma y,\ y \in \mathbb{R},\ y \text{ increasing}\} \subset S, \tag{1.5}$$

where σ is a positive parameter, and $\varphi : \mathbb{R} \to \mathbb{R}$ is a fixed smooth function satisfying

$$\varphi(y) \approx -|y| \quad \text{for large } |y|, \quad \text{and } \varphi(y) \leqslant \gamma - |y| \quad \text{for } y \in \mathbb{R}, \tag{1.6}$$

with $\gamma \in \mathbb{R}$. Thus the two asymptotes of Γ have slope $\pm\sigma$, and $\sigma \geqslant \tan\delta$, and from the point of view of enhancing the exponential decay in (1.3) σ should be taken as small as possible. In Sheen *et al.* (1999) we chose for φ (after the replacement of z by $-z$) the piecewise linear function $\varphi(y) = \gamma - |y|$, with γ the intersection of Γ with the real axis. At the crossing point on the real axis Γ thus had a corner. In this paper we assume instead that $\varphi \in C^\infty(\mathbb{R})$ (with bounded derivatives), because, as was suggested to us by M. Crouzeix, this has benefits over our treatment in Sheen *et al.* (1999) when we subsequently apply quadrature to the contour integral.

A choice of the function φ which we shall use in calculations is

$$\varphi(y) = \gamma - \sqrt{y^2 + \nu^2}, \quad \text{for } y \in \mathbb{R}, \tag{1.7}$$

for suitable parameters $\gamma \in \mathbb{R}$ and $\nu > 0$. The curve Γ is then the left-hand branch of a hyperbola, with asymptotes having slopes $\pm\sigma$, which crosses the real axis at $\gamma - \nu$.

Taking Laplace transforms in (1.1), we obtain the transformed equation

$$(z\,I + A)w(z) = u_0 + \widehat{f}(z), \tag{1.8}$$

and thus $w(z)$ may be represented as

$$w(z) = (z\,I + A)^{-1}(u_0 + \widehat{f}(z)), \quad \text{for } z \in \Gamma \subset S. \tag{1.9}$$

A minimum requirement on $f(t)$, which we shall subsequently weaken, is therefore that the Laplace transform $\widehat{f}(z)$ must have an analytic continuation, also denoted by $\widehat{f}(z)$, from Γ_0 to our deformed contour Γ, and that all singularities of $\widehat{f}(z)$ should lie to the left of Γ. The same property will then apply to $w(z)$ in (1.9).

With the deformed contour represented by (1.5), the integral (1.3) can be written as an infinite integral with respect to a real variable,

$$u(t) = \int_{-\infty}^{\infty} v(t, y)\,\mathrm{d}y, \text{ with } v(t, y) = \frac{1}{2\pi \mathrm{i}} \mathrm{e}^{z(y)t} w(z(y))z'(y), \;\; z(y) := \varphi(y) + \mathrm{i}\,\sigma y. \tag{1.10}$$

The integrand in this integral decays exponentially for large t, because of the assumed behaviour of $\mathrm{Re}\, z(y) = \varphi(y)$.

As in Sheen *et al.* (1999), the integral is now approximated by a suitable quadrature scheme, but this time one for the whole interval $(-\infty, \infty)$, instead of one for each of $(-\infty, 0]$ and $[0, \infty)$ separately. Our approximate solution will thus be of the form

$$U_N(t) = \sum_{j=-N+1}^{N-1} \omega_j v(t, y_j) = \sum_{j=-N+1}^{N-1} \widetilde{\omega}_j \mathrm{e}^{z_j t} w(z_j), \; z_j = z(y_j), \; \widetilde{\omega}_j = \frac{1}{2\pi \mathrm{i}} z'(y_j)\omega_j, \tag{1.11}$$

with the quadrature points $y_j \in \mathbb{R}$ and the nonnegative weights ω_j depending on the particular quadrature scheme. In Section 2 we shall consider in detail one particular scheme, obtained by mapping the infinite integral to $(-1, 1)$ and then applying the trapezoidal rule to the resulting finite integral. The mapping contains a time scaling parameter τ, and because of the exponential decay of the integrand in (1.10), this quadrature rule will have accuracy of order essentially t/τ, thus the order grows linearly with t for $t > \tau$. In the particular case of the homogeneous equation we may think of $U_N(t)$ as defined by a rational function $r(t; A)$ in A applied to u_0,

$$U_N(t) = r(t; A)u_0, \quad \text{where } r(t; a) = \sum_{j=-N+1}^{N-1} \widetilde{\omega}_j \mathrm{e}^{z_j t}(z_j + a)^{-1}.$$

The values of the Laplace transform $w(z)$ needed in (1.11) satisfy, from (1.8),

$$(z_j I + A)w(z_j) = g_j := u_0 + \widehat{f}(z_j), \quad j = -N + 1, \ldots, N - 1. \tag{1.12}$$

As already emphasized in Sheen *et al.* (1999), a central feature of our method is that the $2N-1$ values $w(z_j) \in \mathbb{B}$ in (1.11) can be computed in parallel, since (1.12) can be solved independently for each value of j.

The stability and convergence properties of our method are studied in Section 3. The basic error estimate shows essentially that, for any positive integer r,

$$\|U_N(t) - u(t)\| \leqslant \frac{C}{(N\tau)^r}\big(\|u_0\| + \sup_{z\in\Gamma}\max_{k\leqslant r}\|\widehat{f}^{(k)}(z)\|\big), \quad \text{for } t > r\tau,$$

with the constant C depending on various parameters. Thus, our time discretization method is of arbitrarily high-order, provided t is appropriately bounded below. We remark, as is also borne out by our numerical experiments, that high-order convergence is not possible uniformly down to $t = 0$. This results from the slow decay of the integrand in (1.10) when t is small and, in contrast to the case of time stepping methods, will not be helped by high regularity of u_0, such as $u_0 \in D(A^k)$ for some $k > 0$.

We emphasize again that in (1.12) $\widehat{f}(z_j)$ denotes the value at z_j of the Laplace transform obtained by analytic continuation from the one defined in (1.4), and that approximate values of $\widehat{f}(z_j)$ cannot in general be found by direct quadrature in (1.4). For the scalar case, examples of functions $f(t)$ for which $\widehat{f}(z)$ may be determined analytically on suitable contours Γ are those belonging to the family

$$\mathcal{F} := \left\{ f : \ f(t) = \sum_{j=1}^{J} P_j(t)\mathrm{e}^{-\lambda_j t}, \ \text{ where } \ \mathrm{Re}\lambda_j \geqslant 0, \ P_j(t) = \sum_{l=0}^{q_j} a_{jl}t^l, \ \ a_{jl} \in \mathbb{C} \right\}, \tag{1.13}$$

where $\mathbb{C}$ denotes the complex numbers.

In fact, for $f(t) = t^l\mathrm{e}^{-\lambda t}$, we have

$$\widehat{f}(z) = l!(\lambda + z)^{-l-1}, \quad \text{for } z \neq -\lambda.$$

Because the pole singularity of $\widehat{f}$ lies in the left half-plane, the contour Γ can be chosen so that the singularity is to the left of Γ. For our applications to partial differential equations, where it is necessary to allow f to depend also on a spatial variable $x \in \Omega \subset \mathbb{R}^d$, we shall make use of a generalization of $\mathcal{F}$,

$$\mathcal{F}(\Omega) := \{ f : \ f(x,\cdot) \in \mathcal{F} \text{ for each } x \in \Omega \ \text{ and } \ f(\cdot,t) \in \mathbb{B} \text{ for each } t \geqslant 0\}.$$

Thus in (1.13) we will now have $\lambda_j = \lambda_j(x)$ and $a_{jl} = a_{jl}(x)$.

In terms of the analytic semigroup $E(t)$ the solution of (1.1) may be written as

$$u(t) = E(t)u_0 + \int_0^t E(t-s)f(s)\,\mathrm{d}s, \quad \text{for } t \geqslant 0.$$

Since $\|E(t)\| \leqslant C_0$ for $t \geqslant 0$ and for some $C_0 \geqslant 1$, one has the stability property

$$\|u(t)\| \leqslant C_0\left(\|u_0\| + \int_0^t \|f(s)\|\,\mathrm{d}s\right). \tag{1.14}$$

In particular, to solve (1.1) approximately to a given tolerance, it suffices to solve (1.1) approximately for sufficiently close approximations of u_0 and f. In the case that the solution of a parabolic problem is needed only for a restricted time interval $[0, T]$, the stability result (1.14) allows us to replace the inhomogeneous term $f(t)$ by an approximate inhomogeneous term $F(t)$ on $[0, T]$, in the knowledge that the resulting error in the solution $u(t)$ is bounded by $C_0\|F - f\|_{L_1(0,T;\mathbb{B})}$. In this case the analyticity condition on $\widehat{f}$ is replaced by a corresponding condition on $\widehat{F}$.

In Section 4 we consider briefly a fully discrete approximation scheme for the heat equation in a suitable spatial domain based on piecewise linear finite-elements. We show error estimates in maximum-norm for the fully discrete scheme using a recent maximum-norm bound for the resolvent of the discrete Laplacian by Bakaev *et al.* (2001).

As indicated above, the possibility of approximation of f by a function F for which the $\widehat{F}(z_j)$ may be computed analytically and have acceptable analytic structure is an important consideration. If $f(x, t)$ is a product of the form $p(x)q(t)$, or a sum of such terms, then it suffices to consider just the scalar approximation of $q(t)$ by $F \in \mathcal{F}$, and in Section 5 we describe three different approaches to this problem; more work needs to be done in this regard to make our method a generally viable one. We emphasize, as a key aspect of our method, that it is not necessary that $\widehat{F}(z)$ be a good approximation to $\widehat{f}(z)$, or even that $\widehat{f}(z)$ exists on the contour Γ. It is enough that F be a good approximation to f in $L_1(0, T; \mathbb{B})$.

In Section 6 we give several examples of application to parabolic equations in 0, 1, and 2 space dimensions, combined with finite-element discretization in the spatial variable when $d = 1$ and 2. We use the various methods described in Section 5 to approximate the inhomogeneous term f; a general such method is to determine a separate approximation of $f(x, \cdot)$ by $F = F_x \in \mathcal{F}$ for each vertex x of the finite-element partition.

In practice, the elliptic problems in (1.12) themselves need to be solved approximately, for example by the finite-element method. Such fully discrete schemes then require the solution of large systems of linear equations with complex-valued matrices which are not Hermitian or positive definite. However, if A is a positive definite operator in a Hilbert space then the matrices are obtained from a positive definite matrix $\mathcal{A}$ by the addition of $z_j\mathcal{M}$ with $z_j \in \Gamma$, where $\mathcal{M}$ is the mass matrix (in the case of the finite-element method). In Section 7 we make some remarks about iterative solution of the finite-dimensional linear systems of equations associated with the method.

A representation of the form (1.3) with Γ a smooth path in the left complex half-plane, followed by quadrature based on the trapezoidal rule, was applied by Talbot (1979) for the numerical inversion of Laplace transforms of functions with suitably located singularities.

2. A quadrature scheme

In this section we develop a quadrature formula for an integral over the real axis $\mathbb{R} = (-\infty, \infty)$ with values in $\mathbb{B}$, by making a transformation to the finite interval $(-1, 1)$, and then applying the trapezoidal rule. Under appropriate conditions the resulting quadrature formula has a high order of accuracy. We shall then apply this formula to our representation (1.10) of the solution of the parabolic problem.

To define our quadrature formula we thus first make a change of variables $y = y(\eta)$, where $y(\eta)$ is a smooth increasing function mapping $(-1, 1)$ to $\mathbb{R}$, to obtain, for $v \in$

$C(\mathbb{R};\mathbb{B})$,

$$\mathcal{I}(v) := \int_{-\infty}^{\infty} v(y)\,\mathrm{d}y = \int_{-1}^{1} V(\eta)\,\mathrm{d}\eta, \quad \text{where } V(\eta) = v(y(\eta))y'(\eta). \tag{2.1}$$

Specifically, with τ a positive parameter, we choose $y(\eta)$ to be the odd function

$$y(\eta) = \tau^{-1}\chi(\eta), \quad \text{where } \chi(\eta) = \log((1+\eta)/(1-\eta)). \tag{2.2}$$

In our application τ will play the role of a threshold in t, in that our approximate solution will be accurate of order essentially t/τ for $t > \tau$. Applying the composite trapezoidal rule with spacing $1/N$ to the integral over $(-1, 1)$, and assuming $V(\pm 1) = 0$, we now define

$$Q_{N,\tau}(v) = \frac{1}{N}\sum_{j=-N+1}^{N-1} V(\eta_j) = \frac{1}{N\tau}\sum_{j=-N+1}^{N-1} \mu_j v(y_j),$$

where

$$\eta_j = j/N,\ \mu_j = \chi'(\eta_j) = 2/(1-\eta_j^2),\ y_j = y(\eta_j) = \tau^{-1}\chi(\eta_j). \tag{2.3}$$

We note that the quadrature points y_j are thus distributed over an interval with endpoints $\pm\tau^{-1}\log(2N-1)$. Provided $v(y)$ in (2.1) vanishes appropriately fast at infinity, this formula is then of an arbitrarily high order of accuracy, as expressed in the following lemma.

LEMMA 2.1 Let $r \geqslant 1$ be given and assume that $v \in C^r(\mathbb{R};\mathbb{B})$ and that

$$\|v^{(j)}(y)\| = O(\mathrm{e}^{-r\tau|y|}), \quad \text{for } j \leqslant r \quad \text{as } |y| \to \infty, \tag{2.4}$$

and if $r = 1$ also that $\|v(y)\| = o(\mathrm{e}^{-\tau|y|})$, where τ is the parameter in (2.2). We then have the error estimate

$$\|Q_{N,\tau}(v) - \mathcal{I}(v)\| \leqslant C_r \frac{1}{N^r}\left(1 + \frac{1}{\tau^r}\right)\int_{-\infty}^{\infty} \mathrm{e}^{r\tau|y|}\sum_{j=0}^{r}\|v^{(j)}(y)\|\,\mathrm{d}y.$$

Proof. We shall use the following easy consequence of the Euler–Maclaurin summation formula (see Davis & Rabinowitz, 1975; Stoer, 1992): Assume that $V \in W_1^r((-1,1);\mathbb{B})$ and that

$$V(\pm 1) = V^{(2k-1)}(\pm 1) = 0 \quad \text{for } 2k-1 \leqslant r-2. \tag{2.5}$$

Then

$$\left\|\frac{1}{N}\sum_{j=-N+1}^{N-1} V(\eta_j) - \int_{-1}^{1} V(\eta)\,\mathrm{d}\eta\right\| \leqslant \frac{C_r}{N^r}\int_{-1}^{1}\|V^{(r)}(\eta)\|\,\mathrm{d}\eta. \tag{2.6}$$

In order to apply this to $V(\eta) = v(y(\eta))y'(\eta)$ we need to bound the derivatives entering in (2.6), and to verify that (2.5) is satisfied. One first sees that

$$V'(\eta) = v'(y(\eta))[y'(\eta)]^2 + v(y(\eta))y''(\eta),$$
$$V''(\eta) = v''(y(\eta))[y'(\eta)]^3 + 3v'(y(\eta))y'(\eta)y''(\eta) + v(y(\eta))y'''(\eta).$$

More generally, it follows easily by induction that

$$V^{(k)}(\eta) = \sum_{j=0}^{k} v^{(j)}(y(\eta))\, P_{k,j}(y'(\eta), y''(\eta), \dots, y^{(k+1)}(\eta)), \quad \text{for } k \geqslant 0, \tag{2.7}$$

where $P_{k,j}(y_1, \dots, y_{k+1})$ is a polynomial of degree $\leqslant k+1$ in $y_1, \dots, y_{k+1}$. More precisely, $P_{k,j}(y_1, \dots, y_{k+1})$ is a linear combination of monomials $\prod_{m=1}^{k+1} y_m^{j_m}$ with $\sum_{m=1}^{k+1} m j_m = k+1$. Since by (2.2)

$$y^{(m)}(\eta) = \tau^{-1}(m-1)!\Big((-1)^{m-1}(1+\eta)^{-m} + (1-\eta)^{-m}\Big) \quad \text{for } m \geqslant 1,$$

and since $(1 \mp \eta)^{-1} = (\mathrm{e}^{\pm\tau y} + 1)/2 \leqslant \mathrm{e}^{\tau|y|}$, we find that $|y^{(m)}| \leqslant C_m \tau^{-1} \mathrm{e}^{m\tau|y|}$ and hence

$$\left| P_{k,j}(y', y'', \dots, y^{(k+1)}) \right| \leqslant C_k(\tau^{-1} + \tau^{-k-1})\mathrm{e}^{(k+1)\tau|y|}. \tag{2.8}$$

Since $\mathrm{d}\eta/\mathrm{d}y = \tau(1-\eta^2)/2 \leqslant 2\tau \mathrm{e}^{-\tau|y|}$, we obtain

$$\begin{aligned}\int_{-1}^{1} \|V^{(r)}(\eta)\|\, \mathrm{d}\eta &= \int_{-\infty}^{\infty} \|V^{(r)}(\eta(y))\| \frac{\mathrm{d}\eta}{\mathrm{d}y}(y)\, \mathrm{d}y \\ &\leqslant C_r\left(1 + \frac{1}{\tau^r}\right) \int_{-\infty}^{\infty} \mathrm{e}^{r\tau|y|} \sum_{j=0}^{r} \|v^{(j)}(y)\|\, \mathrm{d}y.\end{aligned}$$

Finally, by (2.7), (2.8), and (2.4) we have

$$\|V^{(k)}(\eta)\| \leqslant C_{r,\tau} \sum_{j=0}^{k} \|v^{(j)}(y(\eta))\|\, \mathrm{e}^{(k+1)\tau|y(\eta)|},$$

giving $V^{(k)}(\eta) \to 0$ as $\eta \to \pm 1$ for $0 \leqslant k \leqslant r-2$, as well as $V(\eta) \to 0$ as $\eta \to \pm 1$ if $r = 1$, and hence showing (2.5). This completes the proof. □

We are now in a position to apply our quadrature scheme to our representation of the solution of (1.1). In our approach the quadrature scheme approximates the exact solution $u(t) = \mathcal{I}(v(t, \cdot))$, where $v(t, y)$ is given in terms of w by (1.10). Explicitly, the approximation to $u(t)$ is given by

$$U_{N,\tau}(t) = Q_{N,\tau}(v(t, \cdot)) = \frac{1}{N\tau} \sum_{j=-N+1}^{N-1} \widetilde{\mu}_j \mathrm{e}^{z_j t} w(z_j), \quad \widetilde{\mu}_j = \frac{1}{2\pi \mathrm{i}} z'(y_j)\mu_j, \tag{2.9}$$

where $w(z)$ is defined by (1.9) and $z_j = z(y_j) = \varphi(y_j) + \mathrm{i}\sigma y_j$, and where $\tau > 0$ is the parameter in the transformation (2.2).

We remark that in the important special case in which $w(\bar{z}) = \overline{w(z)}$ (which is the case, for instance, for the heat equation, or if $\mathbb{B}$ is a Hilbert space and A positive definite), and if we choose the function φ to be even, then we have $\widetilde{\mu}_{-j} = \overline{\widetilde{\mu}_j}$ and $z_{-j} = \bar{z}_j$, allowing (2.9) to be rewritten as

$$U_{N,\tau}(t) = 2\,\mathrm{Re}\left(\frac{1}{N\tau}\sum_{j=0}^{N-1}{}' \widetilde{\mu}_j \mathrm{e}^{z_j t} w(z_j)\right),$$

where the prime indicates that the term with $j = 0$ is to be halved. The practical effect is that the number of elliptic problems of the form (1.12) to be solved is approximately halved.

3. Stability and error estimates

In this section we analyse the stability and error of the approximation of the solution of (1.1) defined in (2.9). We begin with a simple stability estimate.

THEOREM 3.1 Let $U_{N,\tau}(t)$ be the approximate solution of (1.1) defined by (2.9). Assume that $\widehat{f}(z)$ is analytic in z to the right of the contour Γ and continuous onto $\Gamma\cdot$ Then

$$\|U_{N,\tau}(t)\| \leqslant C\mathrm{e}^{\gamma t}\left(1+\frac{\sigma}{N\tau}+\log_+\frac{\sigma}{t}\right)\left(\|u_0\|+\sup_{z\in\Gamma}\|\widehat{f}(z)\|\right), \quad \text{for } t>0.$$

Proof. By (2.9) and (1.6) and since $|\widetilde{\mu}_j| \leqslant (2\pi)^{-1}(\kappa_1+\sigma)\mu_j$ where $\kappa_1 = \|\varphi'\|_{L_\infty(R)}$, we find at once

$$\begin{aligned}\|U_{N,\tau}(t)\| &\leqslant (N\tau)^{-1}\sum_{j=-N+1}^{N-1}|\widetilde{\mu}_j|\mathrm{e}^{\varphi(y_j)t}\|w(z_j)\| \\ &\leqslant (2\pi)^{-1}(\kappa_1+\sigma)(N\tau)^{-1}\mathrm{e}^{\gamma t}\sum_{j=-N+1}^{N-1}\mu_j\,\mathrm{e}^{-|y_j|t}\|w(z_j)\|.\end{aligned}$$

Using (1.9) and (1.2) we have

$$\|w(z_j)\| \leqslant M(1+\sigma|y_j|)^{-1}(\|u_0\|+\max_l\|\widehat{f}(z_l)\|),$$

and it thus remains to bound

$$(N\tau)^{-1}\sum_{j=-N+1}^{N-1}\mu_j\mathrm{e}^{-|y_j|t}(1+\sigma|y_j|)^{-1}.$$

The term with $j = 0$ is $2(N\tau)^{-1}$, and for $j > 0$ we find, since $\mathrm{e}^{-yt}(1+\sigma y)^{-1}$ is decreasing in y for $y \geqslant 0$ and since $\eta_j = j/N$ and $\mathrm{d}\eta/\mathrm{d}y = (1-\eta^2)\tau/2$ (see (2.2) and (2.3)),

$$\begin{aligned}\mathrm{e}^{-y_jt}(1+\sigma y_j)^{-1} &\leqslant N\int_{\eta_{j-1}}^{\eta_j}\mathrm{e}^{-y(\eta)t}(1+\sigma y(\eta))^{-1}\mathrm{d}\eta \\ &= \tfrac{1}{2}N\tau\int_{y_{j-1}}^{y_j}(1-\eta(y)^2)\mathrm{e}^{-yt}(1+\sigma y)^{-1}\mathrm{d}y.\end{aligned}$$

Hence we obtain, by using $\mu_j = 2/(1-\eta_j^2)$ and the easily shown fact that $1-\eta^2 \leqslant 2(1-\eta_j^2)$ for $\eta_{j-1} \leqslant \eta \leqslant \eta_j$ and $1 \leqslant j \leqslant N-1$,

$$\begin{aligned}(N\tau)^{-1}\sum_{j=1}^{N-1}\mu_j\mathrm{e}^{-y_jt}(1+\sigma y_j)^{-1} &\leqslant 2\int_0^\infty \mathrm{e}^{-yt}(1+\sigma y)^{-1}\,\mathrm{d}y \\ &\leqslant C\sigma^{-1}(1+\log_+(\sigma/t)). \qquad (3.1)\end{aligned}$$

The argument for $j < 0$ is analogous. On recalling that $\sigma \geqslant \tan\delta$, the proof is thus complete. □

THEOREM 3.2 Let $u(t)$ be the solution of (1.1) and let $U_{N,\tau}(t)$ be its approximation defined by (2.9). Assume that $\widehat{f}(z)$ is analytic to the right of the contour Γ and continuous onto Γ, with $\widehat{f}^{(j)}(z)$ bounded on Γ for $j \leqslant r$ and r an integer $\geqslant 1$. Then, for $t > r\tau$,

$$\|U_{N,\tau}(t) - u(t)\| \leqslant \frac{C_{r,\sigma}}{N^r}\left(1 + t^r + \frac{1}{\tau^r}\right)\mathrm{e}^{\gamma t}\left(1 + \log_+ \frac{1}{t - r\tau}\right) \times\big(\|u_0\| + \max_{k\leqslant r}\sup_{z\in\Gamma}\|\widehat{f}^{(k)}(z)\|\big).$$

Proof. Recall from (1.10) and (2.9) that

$$U_{N,\tau}(t) - u(t) = Q_{N,\tau}(v(t,\cdot)) - \mathcal{I}(v(t,\cdot)), \quad \text{with } v(t,y) = (2\pi\mathrm{i})^{-1}\mathrm{e}^{z(y)t}z'(y)w(z(y)).$$

To apply Lemma 2.1, we need to bound the derivatives of the function $w(z)$ on Γ. Noting that $(\mathrm{d}/\mathrm{d}z)^j(z\,I + A)^{-1} = (-1)^j\,j!(z\,I + A)^{-j-1}$, we find by using (1.2) and the Leibniz rule applied to (1.9) that

$$\|w^{(j)}(z)\| \leqslant C_j(1 + |z|)^{-1}\big(\|u_0\| + \max_{k\leqslant j}\|\widehat{f}^{(k)}(z)\|\big), \quad \text{for } z \in \Gamma,$$

and hence from (1.10),

$$\|v^{(j)}(t,y)\| \leqslant C_r(1+\sigma)(1 + (1+\sigma)^r t^r)\mathrm{e}^{t\varphi(y)}(1 + \sigma|y|)^{-1} \times \big(\|u_0\| + \max_{k\leqslant r}\|\widehat{f}^{(k)}(z(y))\|\big), \quad \text{for } j \leqslant r,\ y \in \mathbb{R},$$

with C_r depending on $\|\varphi^{(j)}\|_{L_\infty(R)}$ for $j \leqslant r$. By (1.6) the assumptions of Lemma 2.1 are thus satisfied if $t > r\tau$, and the lemma gives

$$\|U_{N,\tau}(t) - u(t)\| \leqslant C_r(1+\sigma)N^{-r}(1 + \tau^{-r})(1 + (1+\sigma)^r t^r)\,\mathrm{e}^{\gamma t} \times \int_{-\infty}^{\infty}\mathrm{e}^{-|y|(t-r\tau)}(1+\sigma|y|)^{-1}\mathrm{d}y\,\big(\|u_0\| + \sup_{z\in\Gamma}\max_{k\leqslant r}\|\widehat{f}^{(k)}(z)\|\big).$$

Bounding the integral as in (3.1) shows the theorem. □

It is natural to think of τ as a threshold, in the sense of being the smallest value of the time t for which the method can be expected to yield first-order convergence in N^{-1}. More precisely, the theorem says that for any $r \geqslant 1$, and $t > r\tau$ there exists an error bound of order $O(N^{-r})$, provided $\widehat{f}$ is appropriately regular. Note that the parameter r does not appear in the method itself, only in the theorem: the larger we can choose r, the faster is the convergence.

As noted earlier, in most applications it will be necessary to replace the inhomogeneous term f by an approximation $F\cdot$ In this case the error can be further decomposed as

$$U^F_{N,\tau}(t) - u(t) = (U^F_{N,\tau}(t) - u^F(t)) + ((u^F(t) - u(t)),$$

where the label F is used to denote quantities with f replaced by $F\cdot$ Application of Theorem 3.2 to the first term and the stability estimate (1.14) to the second at once yield the following theorem.

THEOREM 3.3 Under the assumptions of Theorem 3.2, with f replaced by F, we have

$$\begin{aligned}\|U^F_{N,\tau}(t) - u(t)\| \leqslant & \frac{C_{r,\sigma}}{N^r}\left(1 + t^r + \frac{1}{\tau^r}\right)\mathrm{e}^{\gamma t}\left(1 + \log_+ \frac{1}{t - r\tau}\right) \\ & \times \left(\|u_0\| + \max_{k\leqslant r}\sup_{z\in\Gamma}\|\widehat{F}^{(k)}(z)\|\right) \\ & + C_0\int_0^t \|F(s) - f(s)\|\,\mathrm{d}s, \quad \text{for } t > r\tau.\end{aligned} \tag{3.2}$$

We recall again that when an error bound is needed on $[0, T]$ only, then it suffices that $C_0\|F - f\|_{L_1(0,T;\mathbb{B})}$ is small. We remark that a best approximation of f is not needed: all we need is an approximation F to f of the desired form such that $C_0\|F - f\|_{L_1(0,T;\mathbb{B})} < \epsilon/2$, say, where ϵ is the tolerance in our desired approximate solution. Note that since f is a known function, if we ignore C_0 then this condition is easy to check before we go to the next step in the procedure, which is to solve numerically the approximate parabolic problem, with F replacing f, to the remaining tolerance $\epsilon/2$.

4. Application to finite-element method for the heat equation

In this section we apply our above results to the discretization in both space and time of an initial-boundary value problem for the heat equation,

$$u_t - \Delta u = f(t) \quad \text{in } \Omega, \quad \text{with } u(\cdot, t) = 0 \quad \text{on } \partial\Omega \quad \text{for } t > 0, \quad u(\cdot, 0) = u_0 \quad \text{in } \Omega, \tag{4.1}$$

where Ω is a bounded convex domain with smooth boundary $\partial\Omega$ and Δ is the Laplacian. For simplicity we consider just the case of a two-dimensional domain Ω, but there is no difficulty in principle in extending it to any dimension $d \geqslant 1$. In Sheen *et al.* (1999) we analysed fully discrete schemes within the framework of the Hilbert space $L_2(\Omega)$. Since our present work is developed in a Banach space context, we now consider (4.1) in the Banach space $C(\bar{\Omega})$, and illustrate our theory by deriving error estimates in the maximum-norm $\|v\| = \sup_{x\in\Omega}|v(x)|$. This, however, requires more sophisticated results for the discretization of the associated elliptic problem and the spatial discretization of the parabolic problem.

Let V_h denote standard piecewise linear finite-element spaces defined on a family of quasi-uniform triangulations of Ω and vanishing on $\partial\Omega$. A spatially semidiscrete problem corresponding to (4.1) is to find $u_h(t) \in V_h$ such that, with $(v, w) = \int_\Omega v(x)\overline{w(x)}\,\mathrm{d}x$,

$$(u_{h,t}, \chi) + (\nabla u_h, \nabla\chi) = (f, \chi), \quad \forall\chi \in V_h,\ t > 0, \quad \text{with } u_h(0) = P_h u_0 \in V_h, \tag{4.2}$$

where $P_h : L_2(\Omega) \to V_h$ is the orthogonal projection with respect to $(\cdot, \cdot)$. With $\Delta_h : V_h \to V_h$ being the discrete Laplacian defined by

$$(\Delta_h\psi, \chi) = -(\nabla\psi, \nabla\chi), \quad \forall\psi,\ \chi \in V_h,$$

this problem may also be written as

$$u_{h,t} - \Delta_h u_h = P_h f, \quad \text{for } t > 0, \quad \text{with } u_h(0) = P_h u_0. \tag{4.3}$$

This problem is of the form (1.1) when V_h, equipped with the maximum-norm, is considered as a Banach space.

It was shown in Thomée & Wahlbin (2000) that the semigroup $E_h(t) = \mathrm{e}^{\Delta_h t}$ on V_h generated by Δ_h is analytic, so that

$$\|E_h(t)v_h\| + t\|E_h'(t)v_h\| \leqslant C\|v_h\|, \quad \text{for } t > 0,$$

uniformly in h, and therefore that a maximum-norm resolvent estimate for Δ_h of the form (1.2) holds: i.e. for some $M \geqslant 1$ and some $\delta \in (0, \pi/2)$ independent of h,

$$\|(z\,I - \Delta_h)^{-1}\| \leqslant M(1 + |z|)^{-1}, \quad \text{for } z \in S = \Sigma_{\pi-\delta} \cup B.$$

It was subsequently shown in Bakaev *et al.* (2001) that δ may be chosen arbitrarily small (at the expense of a correspondingly large constant M), which implies that the resolvent estimate does not place any restriction on the slope parameter σ required for Γ.

The fully discrete solution obtained by application of our method to (4.3), with the forcing term f approximated by F, is thus defined from (2.9) by

$$U^F_{N,h,\tau}(t) = \frac{1}{N\tau} \sum_{j=-N+1}^{N-1} \widetilde{\mu}_j \mathrm{e}^{z_j t} w_h^F(z_j), \quad w_h^F(z) = (z\,I - \Delta_h)^{-1} P_h(u_0 + \widehat{F}(z)). \tag{4.4}$$

To find $U^F_{N,h,\tau}(t)$ for a range of values of t it is thus required to solve the $2N-1$ discrete elliptic problems

$$(\nabla w_h^F(z_j), \nabla\chi) + z_j(w_h^F(z_j), \chi) = (u_0 + \widehat{F}(z_j), \chi), \quad \forall \chi \in V_h, \quad |j| \leqslant N-1. \tag{4.5}$$

To estimate the error in $U^F_{N,h,\tau}$ we write, with u_h^F and u^F the solutions of (4.2) and (1.1) with f replaced by F,

$$U^F_{N,h,\tau}(t) - u(t) = (U^F_{N,h,\tau}(t) - u_h^F(t)) + (u_h^F(t) - u^F(t)) + (u^F(t) - u(t)).$$

For the first term we obtain by Theorem 3.2, uniformly in h, with C depending on r, τ, t, and σ as in the error bound in Theorem 3.2, under the appropriate assumptions on $\widehat{F}(z)$,

$$\|U^F_{N,h,\tau}(t) - u_h^F(t)\| \leqslant CN^{-r}\big(\|u_0\| + \max_{k\leqslant r} \sup_{z\in\Gamma} \|\widehat{F}^{(k)}(z)\|\big), \quad \text{for } t > r\tau, \tag{4.6}$$

where we have used the fact (see, for example, Thomée, 1997) that P_h is bounded in the maximum-norm. For the third term we have by the stability estimate (1.14) with $C_0 = 1$ (by the maximum principle),

$$\|u^F(t) - u(t)\| \leqslant \int_0^t \|F(s) - f(s)\|\,\mathrm{d}s. \tag{4.7}$$

Finally, in order to bound the second term we show the following result for the semidiscrete problem.

LEMMA 4.1 We have for the solutions of (4.1) and (4.2), for small h, with $C > 0$ depending on Γ,

$$\|u_h(t) - u(t)\| \leqslant Ch^2 \log^2(1/h)\, t^{-1}\mathrm{e}^{\gamma t}(\|u_0\| + \sup_{z\in\Gamma} \|\widehat{f}(z)\|), \quad \text{for } t > 0.$$

Proof. We have the representation

$$u_h(t) - u(t) = \frac{1}{2\pi \mathrm{i}} \int_\Gamma \mathrm{e}^{tz} \widehat{G}_h(z)(u_0 + \widehat{f}(z))\,\mathrm{d}z,$$

where, with $A = -\Delta$ and $A_h = -\Delta_h$,

$$\widehat{G}_h(z) = (zI + A_h)^{-1} P_h - (zI + A)^{-1}.$$

We shall show below that, for h small, (see also Lubich *et al.*, 1996)

$$\|\widehat{G}_h(z)\| \leqslant Ch^2 \log^2(1/h), \quad \text{for } z \in \Gamma. \tag{4.8}$$

Since $\mathrm{Re}\, z = \varphi(y) \leqslant \gamma - |y|$ and $|\mathrm{d}z/\mathrm{d}y| \leqslant C$ on Γ, this implies

$$\begin{aligned}\|u_h(t) - u(t)\| &\leqslant C\,\mathrm{e}^{\gamma t} h^2 \log^2(1/h) \int_0^\infty \mathrm{e}^{-ty}\mathrm{d}y\, (\|u_0\| + \sup_{z\in\Gamma} \|\widehat{f}(z)\|) \\ &\leqslant Ch^2 \log^2(1/h)\, t^{-1}\mathrm{e}^{\gamma t}(\|u_0\| + \sup_{z\in\Gamma} \|\widehat{f}(z)\|),\end{aligned}$$

which completes the proof.

To prove (4.8) we write

$$\widehat{G}_h(z) = \big((zI + A_h)^{-1} P_h - P_h(zI + A)^{-1}\big) + (P_h - I)(zI + A)^{-1} = G_h'(z) + G_h''(z).$$

Here

$$\begin{aligned}G_h'(z) &= (zI + A_h)^{-1} P_h (zI + A)(zI + A)^{-1} - (zI + A_h)^{-1}(zI + A_h) P_h (zI + A)^{-1} \\ &= (zI + A_h)^{-1}(P_h A - A_h P_h)(zI + A)^{-1} = (zI + A_h)^{-1} A_h (R_h - P_h)(zI + A)^{-1},\end{aligned}$$

where $R_h : H_0^1(\Omega) \to V_h$ is the elliptic projection defined by

$$(\nabla R_h v, \nabla \chi) = (\nabla v, \nabla \chi), \quad \forall \chi \in V_h,$$

and where we have used $P_h A = A_h R_h$. We shall need the maximum-norm error estimate by Schatz & Wahlbin (1982) for the elliptic projection R_h, with $L_p = L_p(\Omega)$ and $W_p^2 = W_p^2(\Omega)$,

$$\|(R_h - I)v\|_{L_p} \leqslant Ch^2 \log(1/h) \|v\|_{W_p^2} \quad \text{if } v = 0 \quad \text{on } \partial\Omega,$$

and also the Agmon *et al.* (1950) regularity estimate

$$\|v\|_{W_p^2} \leqslant Cp \|Av\|_{L_p} \quad \text{if } v = 0 \quad \text{on } \partial\Omega. \tag{4.9}$$

In Agmon *et al.* (1950) this is stated without the explicit dependence on p, which may be traced through the proof to the Calderòn–Zygmund lemma. Using also the inverse estimate $\|\chi\| \leqslant Ch^{-2/p}\|\chi\|_{L_p}$ on V_h and the maximum-norm boundedness of $A(zI + A)^{-1} =$

$I - z(zI + A)^{-1}$ for $z \in S$ (note that (1.2) now holds by Stewart, 1974), and similarly for $(zI + A_h)^{-1}A_h = A_h(zI + A_h)^{-1}$, we find that for $v \in C(\bar{\Omega})$

$$\begin{aligned}\|G_h'(z)\, v\| &\leqslant C\|(R_h - P_h)(zI + A)^{-1}v\| \\ &\leqslant Ch^{-2/p}\|P_h(R_h - I)(zI + A)^{-1}v\|_{L_p} \leqslant Ch^{2-2/p}\log(1/h)\|(zI + A)^{-1}v\|_{W_p^2} \\ &\leqslant Ch^{2-2/p}\log(1/h)\, p\, \|A(zI + A)^{-1}v\|_{L_p} \leqslant Ch^{2-2/p}\log(1/h)\, p\, \|v\|\cdot\end{aligned}$$

Thus, with $p = \log(1/h)$,

$$\|G_h'(z)\, v\| \leqslant Ch^2\log^2(1/h)\|v\|.$$

To bound $G_h''(z)v$ we introduce the piecewise linear interpolant $I_h : C(\bar{\Omega}) \to V_h$ and note that, for any triangle K of the triangulation, the Bramble–Hilbert lemma implies

$$\|I_h v - v\|_{L_\infty(K)} \leqslant Ch^{2-2/p}\|v\|_{W_p^2(K)},$$

from which the corresponding estimate follows with K replaced by Ω. Hence, since P_h is bounded in maximum-norm, and using (4.9),

$$\|(P_h - I)v\| = \|(P_h - I)(I_h - I)v\| \leqslant Ch^{2-2/p}\|v\|_{W_p^2} \leqslant Ch^{2-2/p}\, p\, \|Av\|_{L_p}.$$

Using $\|A(zI + A)^{-1}\| \leqslant C$, it follows that

$$\|G_h''(z)v\| \leqslant Ch^{2-2/p}\, p\, \|A(zI + A)^{-1}v\|_{L_p} \leqslant Ch^{2-2/p}\, p\, \|v\|,$$

and, again with $p = \log(1/h)$,

$$\|G_h''(z)v\| \leqslant Ch^2\log(1/h)\|v\|,$$

which completes the proof of (4.8), and thus the lemma. □

This result has the character of a nonsmooth data error estimate in that it requires no regularity of data with respect to the spatial variable x, at the expense of the factor t^{-1} on the right. For solutions which are smoother in x, this factor and one of the factors $\log(1/h)$ can be removed. When the Banach space is chosen as the Hilbert space $L_2(\Omega)$ the factors $\log(1/h)$ are superfluous. We refer to Thomée (1997) for details.

Together (4.6), (4.7), and Lemma 4.1 applied with f replaced by F, show the following error bound for the fully discrete solution.

THEOREM 4.2 Let $u(t) \in C(\overline{\Omega})$ be the solution of (4.1), and let $U_{N,h,\tau}^F(t)$ be the approximation defined by (4.4). Assume that $\widehat{F}(z)$ is analytic to the right of the contour Γ and continuous on Γ, and that $\widehat{F}^{(j)}(z)$ is bounded on Γ in the maximum-norm for $j \leqslant r$, with r an integer $\geqslant 1\cdot$ Then for $t > \tau r$, for small h, with C independent of N and h, but depending on t, τ and Γ,

$$\begin{aligned}\|U_{N,h,\tau}^F(t) - u(t)\| &\leqslant C\, N^{-r}\Big(\|u_0\| + \max_{k\leqslant r}\sup_{z\in\Gamma}\|\widehat{F}^{(k)}(z)\|\Big) \\ &\quad + C\, h^2\log^2\frac{1}{h}\Big(\|u_0\| + \sup_{z\in\Gamma}\|\widehat{F}(z)\|\Big) + \int_0^t \|F(s) - f(s)\|\, \mathrm{d}s\cdot\end{aligned} \tag{4.10}$$

In the implementation of our method we need to be able to calculate the inner products $(\widehat{F}(z_j), \chi)$ occurring in (4.5), where F is the approximation of f, and $\chi \in V_h$. In principle, this requires us to determine first the approximation $F(x, t)$ and then its Laplace transform $\widehat{F}(z_j)$ at each $x \in \Omega$, or at least at a sufficiently large number of quadrature points for the inner products to be well approximated. One way to reduce this work is to replace f by its interpolant $I_h f \in V_h$ in which case we need only approximate $f_\xi(t) = f(\xi, t)$ at the vertices $\{\xi\}$ of the triangulation by suitable scalar functions $F_\xi(t)$. We then define $F(x, t)$ as the piecewise linear interpolant $F(x, t) = \sum_\xi F_\xi(t)\Phi_\xi(x) \in V_h$, where the Φ_ξ are the standard basis functions for V_h. Now $F(x, t) \in \mathcal{F}(\Omega)$ and $\widehat{F}(x, z) = \sum_\xi \widehat{F}_\xi(z)\Phi_\xi(x)$, so that only the Laplace transforms of the scalar functions $F_\xi(t)$ are needed. We note that

$$(\widehat{F}(z_j), \Phi_\xi) = \sum_{\xi'} \widehat{F}_{\xi'}(z_j)(\Phi_{\xi'}, \Phi_\xi),$$

where the $(\Phi_{\xi'}, \Phi_\xi)$ are the elements of the mass matrix which is already needed in the left-hand side of (4.5). Writing $F - f = (F - I_h f) + (I_h f - f)$ we may bound the last term in (4.10) by

$$\int_0^t \|F(s) - f(s)\| \, ds \leqslant \int_0^t \max_\xi |F_\xi(s) - f_\xi(s)| \, ds + \int_0^t \|I_h f(s) - f(s)\| \, ds. \quad (4.11)$$

The first term contains the approximation error at the vertices and the last term is clearly of $O(h^2)$ if f is sufficiently smooth.

5. Approximation of a scalar function f

In this section we give examples of methods for finding approximations F of a scalar function f on a finite interval $[0, T]$, in such a way that F has an analytically known Laplace transform with simple analytic structure. In practice we take $F \in \mathcal{F}$ (see Section 1). If all of the exponents λ_j in (1.13) are real and positive for the function F then the singularities of $\widehat{F}$ are poles on the negative real axis, and therefore impose little constraint on the choice of the deformed contour Γ. In theory such an approximation is *always* possible on $[0, T]$, even with the q_j in (1.13) equal to zero, if f is continuous there, because the transformation $s = e^{-t}$ converts the approximation problem $f(t) \approx \sum_{j=1}^J a_j e^{-jt}$ to one of polynomial approximation on the finite interval $[e^{-T}, 1]$. Thus the existence of a uniform approximation of any desired accuracy follows from the Weierstrass theorem. However, it could happen that the number of terms and the coefficients in this approximation are prohibitively large. Explicit Laplace transforms are known in terms of special functions for our examples, but approximation of f in $\mathcal{F}$ still provides convenient tests of the various approaches.

5.1 *Least-squares approximation*

One way to find an approximation $F \in \mathcal{F}$ of a given function f on $[0, T]$ is to prescribe both a finite set $\{\lambda_j\}_{j=1}^J$ with $\operatorname{Re}\lambda_j \geqslant 0$ and the degrees q_j of the corresponding polynomials $P_j(t)$ in (1.13) and then, with $\bar{\mathcal{F}}$ denoting the class $\mathcal{F}$ thus restricted, to

determine these polynomials so that F is the best approximation of f in an appropriate weighted L_2-norm, i.e. so that F minimizes

$$\|F - f\|_{L_{2,\omega}(0,T)} = \left(\int_0^T \omega(t)(F(t) - f(t))^2 \mathrm{d}t\right)^{1/2},$$

where ω is a weight function with $\omega^{-1} \in L_1(0, T)$. This minimization problem is a standard linear least-squares problem which has a unique solution $F \in \bar{\mathcal{F}}$. Moreover, we have the L_1 bound

$$\|F - f\|_{L_1(0,T)} \leqslant \|\omega^{-1}\|^{1/2}_{L_1(0,T)} \|F - f\|_{L_{2,\omega}(0,T)}.$$

The weight function may be chosen as $\omega(t) \equiv 1$ on $[0, T]$, or we may choose it small near $t = 0$, such as $\omega(t) = t^{1/2}$, to allow $F(t) - f(t)$ to be larger there.

We illustrate this approach for the function

$$f_0(t) = \frac{(1-t)^2}{(1+t^2)^2}, \quad \text{for } t \in [0, 4],$$

using the weight function $\omega(t) \equiv 1$ on $[0, T]$.

Since we do not wish to address here the nonlinear problem of finding good choices for the exponents λ_j in (1.13), we have restricted our attention to the special case in which they are integers; methods for finding both exponents and coefficients in special cases will be considered in Sections 5.2 and 5.3. Numerical exploration showed that a least-squares approximation of the form

$$F(t) = \sum_{j=1}^{5} \sum_{l=0}^{3} a_{jl} t^l \mathrm{e}^{-jt} \tag{5.1}$$

yielded the L_1-error shown in the first column of Table 1. It should be said, however, that the resulting linear system is strongly ill-conditioned, and thus the resulting numbers depend sensitively on precise details of the computation. The inner products of the basis functions over $[0, 4]$ were calculated exactly by using closed formulae, while those involving the function f were calculated by the composite Simpson rule over 1000 uniform subintervals. (Since Simpson's rule is fourth-order accurate, the resulting error is of order $h^4 = (4 \times 10^{-3})^4 = 2{\cdot}56 \times 10^{-10}$.) The computations in Table 1 were obtained on a 64-bit machine by using a singular value decomposition, with singular values less than 10^{-15} times the largest singular value ignored, and correspond to the approximation

$$\begin{aligned} F(t) = F_0(t) = {} & (-31{\cdot}8 + 94{\cdot}9t - 25{\cdot}1t^2 + 1{\cdot}9t^3)\mathrm{e}^{-t} \\ & + (-25{\cdot}8 - 41{\cdot}6t + 77{\cdot}1t^2 - 68{\cdot}8t^3)\mathrm{e}^{-2t} \\ & + (24{\cdot}1 - 49{\cdot}2t + 53{\cdot}6t^2 - 188{\cdot}4t^3)\mathrm{e}^{-3t} \\ & + (68{\cdot}8 + 29{\cdot}8t - 290{\cdot}2t^2 + 104{\cdot}0t^3)\mathrm{e}^{-4t} \\ & + (-34{\cdot}3 + 56{\cdot}8t + 288{\cdot}1t^2 + 121{\cdot}1t^3)\mathrm{e}^{-5t}. \end{aligned}$$

TABLE 1 $L_1(0, 4)$ *approximation errors and the quantities* $F_{\Gamma,k}$ *defined in* (5·1)

	$L_1(0,4)$-error	$F_{\Gamma,0}$	$F_{\Gamma,1}$	$F_{\Gamma,2}$	$F_{\Gamma,3}$	$F_{\Gamma,4}$
$f_0(t)$	0·767E−04	0·61	0·84	4·16	26·17	177·08
$f_1(t)$	0·102E−04	0·94	1·10	3·08	18·68	141·80
$f_2(t)$	0·114E−05	0·56	0·40	0·73	2·44	13·85
$f_3(t)$	0·105E−05	0·39	0·19	0·23	0·53	2·06
$f(t)$	0·104E−04	0·93	0·73	1·22	3·62	17·38

(For brevity, coefficients are here rounded to one decimal place, whereas to produce the results in Table 1 accuracy to about eight decimal places is needed.) The ill-conditioning is seen in spectacular fashion if the smallest singular values are not ignored: the L_1 error in Table 1 is then improved by a factor approaching 2, but the coefficients in F_0 are larger by an order of magnitude or more. It should be emphasized, however, that either version of the approximation can be used if the L_1 error is small enough and if enough significant figures are carried through to the calculation of the Laplace transform

$$\widehat{F}(z) = \sum_{j=1}^{5} \sum_{l=0}^{3} a_{jl}\, l!(j+z)^{-l-1} \quad \text{for } z \neq -j,\ j = 1, \ldots, 5. \tag{5.2}$$

In our error estimates we also need bounds for the quantities in (3.2), i.e. in this case

$$F_{\Gamma,k} = \|\widehat{F}^{(k)}\|_{L_\infty(\Gamma)}.$$

These are also entered in Table 1. The derivatives were computed directly from (5.2).

Table 1 also shows the analogous results for the application of this approach to the functions

$$f_l(t) = (1+t)^{-l}, \quad l = 1, 2, 3, \quad \text{and } f(t) = (1+t^2)^{-1}. \tag{5.3}$$

5.2 *Completely monotonic functions and Gauss–Laguerre approximation*

One class of functions f for which approximation in $\mathcal{F}$ is very natural is that of 'completely monotonic functions' in t. Considered as functions in t, these are C^∞ functions on $[0, \infty)$ with the property that $(-1)^k f^{(k)}(t) > 0$ for $t \geqslant 0$ for all $k \geqslant 0$. It is known (Bernstein's theorem, see Widder, 1941) that the completely monotonic functions are exactly those that can be represented in the form of a Stieltjes integral

$$f(t) = \int_0^\infty \mathrm{e}^{-st} \mathrm{d}\omega(s), \tag{5.4}$$

with $\mathrm{d}\omega$ a positive measure. The prototypical example is $f(t) = \mathrm{e}^{-\lambda t}$, in which case ω is a point mass at $s = \lambda$; this function is already in $\mathcal{F}$. For a function representable in the form (5.4) the Laplace transform is

$$\widehat{f}(z) = \int_0^\infty \frac{1}{z+s}\, \mathrm{d}\omega(s),$$

which is analytic in the complex plane cut along the negative real axis, with bounded derivatives as $|z| \to \infty$, provided $\int_0^\infty d\omega(s) < \infty$. A function of the form (5.4) may be approximated arbitrarily closely on a finite interval $[0, T]$ by a finite exponential sum $F(t) = \sum_{j=1}^J a_j e^{-\lambda_j t}$ with a_j positive and $\lambda_j \geqslant 0$, thus, in particular belonging to $\mathcal{F}$. This will be illustrated by an example presently. In this case the Laplace transform becomes $\widehat{F}(z) = \sum_{j=1}^J a_j (z + \lambda_j)^{-1}$.

More generally, we can allow functions f of the form $f_1 - f_2$, where f_1 and f_2 are completely monotonic. This corresponds to the representation (5.4) with the positive measure replaced by a bounded real measure $d\omega$, which is the difference between the positive measures corresponding to f_1 and f_2.

As examples of completely monotonic functions which are not themselves in $\mathcal{F}$ we consider the functions f_l in (5.3), for $l = 1, 2, 3$. We shall approximate these functions by finite exponential sums on $[0, T]$ in two different ways, one in this section and the other in Section 5.3. For our first approach, we note that

$$f_l(t) = \frac{1}{(l-1)!} \int_0^\infty s^{l-1} e^{-(1+t)s}\, ds, \tag{5.5}$$

which is of the form (5.4) with $d\omega(s) = s^{l-1} e^{-s} ds/(l-1)!$. The result of applying a quadrature formula to the latter integral will produce a function in $\mathcal{F}$. We shall make use of the Gauss–Laguerre quadrature formula

$$\int_0^\infty g(s) e^{-s}\, ds \approx \sum_{j=1}^J w_j g(s_j), \tag{5.6}$$

where s_j, $j = 1, \ldots, J$, are the zeros of the Laguerre polynomial of degree J, and w_j, $j = 1, \ldots, J$, the corresponding quadrature weights. This formula is exact for g a polynomial of degree $2J - 1$. Before applying (5.6), we choose to make a linear transformation of the integrand in the integral (5.5). Specifically, we choose a number $\xi \in [0, T]$ and write

$$\begin{aligned} f_l(t) &= \frac{1}{(l-1)!} \int_0^\infty s^{l-1}\, e^{-(t-\xi)s} e^{-(1+\xi)s}\, ds \\ &= (1+\xi)^{-l} \frac{1}{(l-1)!} \int_0^\infty s^{l-1}\, e^{-(t-\xi)s/(1+\xi)} e^{-s}\, ds, \end{aligned}$$

where in the last step we replaced $(1 + \xi)s$ by a new integration variable. Applying (5.6) with $g(s) = (1+\xi)^{-l} s^{l-1}\, e^{-(t-\xi)s/(1+\xi)}/(l-1)!$, we now obtain the explicit formula

$$f_l(t) \approx \sum_{j=1}^J \mu_j e^{-\lambda_j t}, \text{ where } \lambda_j = s_j (1+\xi)^{-1}, \ \mu_j = w_j s_j^{l-1} (1+\xi)^{-l} e^{\xi \lambda_j}/(l-1)!.$$

The motivation for using the Gauss–Laguerre quadrature formula in this way is that the approximation is now exact for $t = \xi$ provided $J \geqslant l/2$, since $g(s)$ is then a polynomial of degree $l - 1$; the error grows as t moves away from ξ. If the approximation is required in $[0, T]$ it would seem undesirable to choose ξ as one of the endpoints of this interval, since

TABLE 2 *$L_1(0,4)$ approximation errors for $f_l(t), l = 1,2,3$, on $[0,4]$ by 12-point Gauss–Laguerre approximation using (5·7) with $\xi = 2$, and maximum absolute values $F_{\Gamma,k}$ of $\widehat{F}^{(k)}(z), k = 1,\ldots,4$, over the contour Γ*

	$L_1(0,4)$-error	$F_{\Gamma,0}$	$F_{\Gamma,1}$	$F_{\Gamma,2}$	$F_{\Gamma,3}$	$F_{\Gamma,4}$
$f_1(t)$	0·134E−07	0·93	1·08	2·93	13·09	82·96
$f_2(t)$	0·246E−06	0·56	0·40	0·70	2·25	10·89
$f_3(t)$	0·223E−05	0·39	0·19	0·22	0·50	1·78

the error then would be relatively large at the other endpoint. Although it might not be the optimal choice, it appears natural and simple to take $\xi = T/2$.

For example, for $T = 4$ we choose $\xi = 2$. For $J = 12$ we then find

$$\begin{aligned} f_2(t) \approx\ & 0{\cdot}004\,\mathrm{e}^{-0{\cdot}039t} + 0{\cdot}039\,\mathrm{e}^{-0{\cdot}204t} + 0{\cdot}112\,\mathrm{e}^{-0{\cdot}504t} + 0{\cdot}188\,\mathrm{e}^{-0{\cdot}945t} \\ & + 0{\cdot}220\,\mathrm{e}^{-1{\cdot}533t} + 0{\cdot}194\,\mathrm{e}^{-2{\cdot}282t} + 0{\cdot}133\,\mathrm{e}^{-3{\cdot}207t} + 0{\cdot}070\,\mathrm{e}^{-4{\cdot}335t} \\ & + 0{\cdot}029\,\mathrm{e}^{-5{\cdot}706t} + 0{\cdot}009\,\mathrm{e}^{-7{\cdot}348t} + 0{\cdot}002\,\mathrm{e}^{-9{\cdot}496t} + 0{\cdot}0002\,\mathrm{e}^{-12{\cdot}366t}, \end{aligned} \tag{5.7}$$

where neither the exponents nor the coefficients are given to sufficient accuracy.

The quality of the approximation (5.7) and its analogues for $l = 1$ and 3 is explored in Table 2 with $T = 4$, $\xi = 2$, and $J = 12$.

5.3 *Prony's method*

We now describe an approach which is based on Prony's method for exponential fitting (see de Prony, 1795; Kammler, 1977). We want to determine real or complex numbers a_j, λ_j such that

$$f(t) \approx \sum_{j=1}^{J} a_j \mathrm{e}^{-\lambda_j t} =: F_J(t) \quad \text{on } [0,T]. \tag{5.8}$$

Setting $t_l = lT/m,\ l = 0,\ldots,m$, we try to accomplish this approximation by exact or approximate interpolation at $t = t_l,\ l = 0,\ldots,m$, i.e.

$$f_l := f(t_l) \approx \sum_{j=1}^{J} a_j \mu_j^l, \quad l = 0,\ldots,m, \quad \text{where } \mu_j = \mathrm{e}^{-\lambda_j T/m}.$$

We discuss the choice of J and m below.

Assume that $m \geqslant 2J-1$ and that the μ_j are the J roots of an algebraic equation

$$\mu^J - \beta_1 \mu^{J-1} - \beta_2 \mu^{J-2} - \cdots - \beta_J = 0, \tag{5.9}$$

where $\beta_1,\ldots,\beta_J$ are as yet to be chosen. If the approximation is an exact interpolant, i.e. if

$$\sum_{j=1}^{J} a_j \mu_j^l = f_l,\ \ l = 0,\ldots,m, \tag{5.10}$$

then on multiplying $J+1$ consecutive equations (with l running from $k-J$ to k for $k \geqslant J$) by $-\beta_J, \ldots, -\beta_1, 1$ and summing, we find by (5.9)

$$f_{k-1}\beta_1 + \cdots + f_{k-J}\beta_J = f_k, \quad k = J, \ldots, m, \tag{5.11}$$

or, in matrix form,

$$\mathcal{M}\beta = g, \tag{5.12}$$

where $\mathcal{M}$ is an $(m - J + 1) \times J$ matrix.

Exact interpolation corresponds to $m = 2J-1$, in which case (5.12) gives J equations to solve for the J unknowns $\beta_1, \ldots, \beta_J$. These can then be inserted in the algebraic equation (5.9), which we may now solve for the μ_j, and then finally (5.10) may be used to determine the $a_1, \ldots, a_J$.

We may also want to use $m > 2J - 1$. In this case we cannot expect to have equality in (5.10) and hence also not in (5.11). In this case we may instead determine β in (5.11) or (5.12) such that $|\mathcal{M}\beta - g|$ is as small as possible in the l_2 sense. This least-squares problem may also be written as

$$\mathcal{M}^*\mathcal{M}\beta = \mathcal{M}^* g. \tag{5.13}$$

Note that $\mathcal{M}^*\mathcal{M}$ is a $J \times J$ matrix. One could think that $m > 2J - 1$ is the preferred option, because it uses more information about the function $f(t)$. On the other hand, if (as is usual) $\mathcal{M}$ is badly conditioned, then the use of the normal equation (5.13) makes the problem of ill-conditioning even worse. In the calculations below we used $m = 2J - 1$, and solved the $J \times J$ linear system by using again a singular value decomposition, with small singular values treated as in Section 5.1.

Note that if f is itself an exponential sum with J terms and $m > 2J - 1$, then the equations in the system (5.11) are linearly dependent.

It may be proved (see Kammler, 1977) that if f is completely monotonic, and not itself an exponential sum with less than J terms, then there is a unique function F_J of the form (5.8) with all λ_j real and non-negative that interpolates f at the points t_l, $l = 0, \ldots, 2J - 1$. It is also shown in Kammler (1977) that in this situation F_J converges to f as $J \to \infty$, and that $|F_J - f| \leqslant |f'(0)|T/(2J-1) = O(J^{-1})$.

We now illustrate Prony's method by applying it to the completely monotonic functions $f_l(t) = 1/(1+t)^l$, $l = 1, 2, 3$, with $J = 8$. For instance, for $l = 2$ our calculations give

$$\begin{aligned} f_2(t) \approx\ & 0{\cdot}049\,\mathrm{e}^{-0{\cdot}185t} + 0{\cdot}203\,\mathrm{e}^{-0{\cdot}639t} + 0{\cdot}330\,\mathrm{e}^{-1{\cdot}410t} + 0{\cdot}282\,\mathrm{e}^{-2{\cdot}602t} \\ & + 0{\cdot}120\,\mathrm{e}^{-4{\cdot}439t} + 0{\cdot}017\,\mathrm{e}^{-7{\cdot}527t} \\ & - 0{\cdot}187\,10^{-9}\,\mathrm{e}^{-0{\cdot}641t}\cos(2{\cdot}155t) - 0{\cdot}288\,10^{-8}\,\mathrm{e}^{-0{\cdot}641t}\sin(2{\cdot}155t). \end{aligned}$$

The errors in the approximations (if the coefficients and exponents are retained to full accuracy rather than in short form as here) are shown in Table 3. We note that the last two terms correspond to a complex λ_j in (5.8) and so by the theory of Kammler (1977) should not appear. Since the coefficients in this case are small, they can be omitted without changing the quality of the approximate solution shown in the second line of the table.

TABLE 3 *$L_1(0,4)$ approximation errors for f_l, $l = 1, 2, 3$, by Prony's method with $J = 8$ terms, together with the $F_{\Gamma,k}$, $k = 0, \ldots, 4$*

	$L_1(0,4)$-error	$F_{\Gamma,0}$	$F_{\Gamma,1}$	$F_{\Gamma,2}$	$F_{\Gamma,3}$	$F_{\Gamma,4}$
$f_1(t)$	0·266E−06	0·93	1·08	2·93	13·05	82·18
$f_2(t)$	0·125E−05	0·56	0·40	0·71	2·25	10·83
$f_3(t)$	0·370E−05	0·39	0·19	0·22	0·50	1·79

TABLE 4 *$L_1(0,4)$ approximation errors for $f(t) = 1/(1 + t^2)$ by the J-term Prony method, together with the $F_{\Gamma,k}$, $k = 0, \ldots, 4$*

J	$L_1(0,4)$-error	$F_{\Gamma,0}$	$F_{\Gamma,1}$	$F_{\Gamma,2}$	$F_{\Gamma,3}$	$F_{\Gamma,4}$
8	0·159E−03	0·93	0·71	1·18	3·37	14·73
12	0·408E−03	0·96	0·72	1·18	3·37	14·75
16	0·871E−04	0·97	0·72	1·18	3·37	14·57

We next apply Prony's method to a case where $f(t)$ is not completely monotonic. We consider the function $f(t) = 1/(1 + t^2)$. For this function the approximations with 8, 12 and 16 exponential terms were computed; the approximation obtained for $J = 12$ is

$$\begin{aligned}\frac{1}{1+t^2} \approx\ & 0{\cdot}560\,10^{-8}\,\mathrm{e}^{-0{\cdot}783t}\cos(3{\cdot}826t) - 0{\cdot}459\,10^{-7}\,\mathrm{e}^{-0{\cdot}783t}\sin(3{\cdot}826t)\\ & -0{\cdot}485\,10^{-7}\,\mathrm{e}^{-0{\cdot}767t}\cos(7{\cdot}963t) + 0{\cdot}329\,10^{-7}\,\mathrm{e}^{-0{\cdot}767t}\sin(7{\cdot}963t)\\ & -0{\cdot}776\,\mathrm{e}^{-4{\cdot}848t}\cos(1{\cdot}175t) - 1{\cdot}07\,\mathrm{e}^{-4{\cdot}848t}\sin(1{\cdot}175t)\\ & +0{\cdot}0328\,\mathrm{e}^{-7{\cdot}378t}\cos(-4{\cdot}917t) + 0{\cdot}023\,\mathrm{e}^{-7{\cdot}378t}\sin(4{\cdot}917t)\\ & +1{\cdot}12\,\mathrm{e}^{-1{\cdot}731t} + 0{\cdot}531\,\mathrm{e}^{-0{\cdot}802t} + 0{\cdot}168\,10^{-3}\,\mathrm{e}^{-0{\cdot}789t} + 0{\cdot}093\,\mathrm{e}^{-0{\cdot}236t}.\end{aligned}$$

Table 4 shows $L_1(0,4)$ approximation errors, together with the $F_{\Gamma,k}$, $k = 0, \ldots, 4$ defined in (5.1) We note that the first six terms have small coefficients, and can be neglected without affecting the quality of the approximation in the table. In the applications in the next section it is important either to omit such terms, or to take account of the inconvenient complex poles of $\widehat{F}(z)$ at $z = 0{\cdot}77 \pm 7{\cdot}96\mathrm{i}$ when selecting the deformed contour Γ.

We can see after the event that the Prony approximation to $1/(1 + t^2)$, although not guaranteed to perform satisfactorily by the available theory, does in fact yield an approximation belonging to the class $\mathcal{F}$, in that all the exponents have positive real part. The coefficients in the first two lines are again small enough to allow us to drop these terms, without affecting the accuracy in Table 4.

In spite of the success of Prony's method for the above particular example, it is clear that, even when it yields a function of the form $\sum_{j=1}^{J} P_j(t)\mathrm{e}^{-\lambda_j t}$ this may not satisfy $\mathrm{Re}\,\lambda_j \geqslant 0$ for all j, and thus may not necessarily belong to $\mathcal{F}$. A simple example is provided by the pure increasing exponential $f(t) = \mathrm{e}^t$, for which Prony's method with $J = 1$ will reproduce f which is not in $\mathcal{F}$.

6. Numerical examples

In this section we give several examples of application of our method to parabolic partial differential equations. We begin with two examples of scalar equations, so that the time discretization can be illustrated without any pollution from discretization in a spatial variable. In the first example the forcing term f belongs to $\mathcal{F}$ so that no error from approximation of f is present. In the second example f is approximated by $F \in \mathcal{F}$ using the least-squares method. We then consider a spatially one-dimensional example with $f \in \mathcal{F}(\Omega)$, and finally two examples of spatially two-dimensional examples, with Ω a square, the first with a $f(x, \cdot) \notin \mathcal{F}(\Omega)$, but which is completely monotonic for each $x \in \Omega$, the second not having this property.

In the examples one must choose the parameters σ, γ and ν which specify the contour Γ in (1.5) and (1.7). As stated earlier, the contour has to be chosen so that both the spectrum of $-A$ and all singularities of $\widehat{f}$ are to the left of Γ, and not close to Γ· (If f is approximated by F it is the singularities of $\widehat{F}$ that must be to the left of Γ)· In all our examples σ (the magnitude of the asymptotic slope of the arms of Γ) was chosen equal to 1. We also usually use $\tau = 0{\cdot}5$, thus (by the discussion at the end of Section 3) the method can be expected to yield the order of convergence $t/\tau = 2t$ for $t > 0{\cdot}5$.

EXAMPLE 1 As our first example we consider the ordinary differential equation

$$u' + u = \mathrm{e}^{-2t}, \quad \text{for } t > 0, \quad \text{with } u(0) = 0,$$

which has the exact solution $u(t) = \mathrm{e}^{-t} - \mathrm{e}^{-2t}$. The Laplace-transformed elliptic problem is $(z+1)w(z) = (z+2)^{-1}$, thus $w(z)$ is given by $w(z) = ((z+1)(z+2))^{-1}$· In this model problem $A = 1$ and $\sigma(A) = \{1\}$·

In Table 5, ϵ_N denotes the error at selected t-values, where N is the positive integer used in (1.11), and $\rho_N = \log_2(\epsilon_{N/2}/\epsilon_N)$· The parameters used for our contour in (1.5) and (1.7) were $\gamma = 0{\cdot}5$ and $\nu = 0{\cdot}5$, so that the contour crosses the real axis at the origin, and the threshold τ is chosen to be $\tau = 0{\cdot}5$.

The numerical results in Table 5 provide confirmation of the behavior predicted by Theorem 3.2, in that the apparent rates of convergence ρ_N, while erratic, are generally at least as large as the order $t/\tau = 2t$ predicted by the theorem, and sometimes very much larger. Increasing t increases the rate of convergence, and indeed for $t = 4$ the error is remarkably small if $N \geqslant 40$·

In Table 6 we illustrate the effect of different choices of τ by listing the results for $\tau = 0{\cdot}25$ and $\tau = 1$, with $N = 40$ and 80. For this example the benefit from the smaller τ value in increasing the order of convergence dominates the negative effects of small τ predicted by Theorem 3.2.

EXAMPLE 2 As our next example we consider the ordinary differential equation

$$u' + u = f_0(t) := \frac{(1-t)^2}{(1+t^2)^2}, \quad \text{for } t > 0, \quad \text{with } u(0) = 1,$$

which has the exact solution $u(t) = 1/(1+t^2)$. We recall the discussion in Section 5 of the least-squares approximation of $f(t) = f_0(t)$ by the function $F(t)$ of the form (5.1). Table 7 shows the error and apparent order of convergence for Example 2 with that approximation

TABLE 5 *Error and apparent order of convergence for Example 1 with* $N = 10, 20, 40,$ *and* 80 *for* $\tau = 0{\cdot}5$

t	ϵ_{10}	ϵ_{20}	ρ_{20}	ϵ_{40}	ρ_{40}	ϵ_{80}	ρ_{80}
0·2	0·372E−03	0·202E−03	0·88	0·357E−03	− 0·82	0·345E−03	0·05
0·4	0·481E−03	0·165E−03	1·54	0·397E−04	2·06	0·171E−05	4·54
0·6	0·170E−04	0·259E−04	−0·60	0·955E−05	1·44	0·170E−05	2·49
0·8	0·144E−04	0·184E−06	6·29	0·117E−05	−2·68	0·313E−06	1·91
1·0	0·909E−05	0·116E−05	2·97	0·430E−07	4·75	0·445E−07	−0·05
1·2	0·148E−05	0·263E−06	2·49	0·278E−07	3·24	0·507E−08	2·46
1·4	0·521E−05	0·242E−07	7·75	0·936E−08	1·37	0·287E−09	5·03
1·6	0·958E−05	0·253E−07	8·56	0·125E−08	4·34	0·619E−10	4·33
1·8	0·155E−04	0·778E−08	10·96	0·167E−09	5·54	0·265E−10	2·66
2·0	0·230E−04	0·789E−08	11·51	0·122E−09	6·01	0·544E−11	4·49
3·0	0·912E−04	0·219E−07	12·02	0·125E−12	17·42	0·631E−15	7·63
4·0	0·264E−03	0·654E−07	11·98	0·578E−14	23·43	0·104E−16	9·12

TABLE 6 *Errors and apparent orders of convergence for Example 1 with* $N = 40$ *and* 80 *for* $\tau = 0{\cdot}25$ *and* 1

	$\tau = 0{\cdot}25$			$\tau = 1$		
t	ϵ_{40}	ϵ_{80}	ρ_{80}	ϵ_{40}	ϵ_{80}	ρ_{80}
0·2	0·195E−04	0·651E−06	4·91	0·494E−02	0·279E−02	0·82
0·4	0·415E−06	0·126E−06	1·72	0·134E−02	0·106E−02	0·34
0·6	0·946E−08	0·161E−08	2·55	0·721E−03	0·321E−03	1·17
0·8	0·771E−09	0·251E−10	4·94	0·637E−04	0·248E−04	1·36
1·0	0·439E−09	0·150E−11	8·19	0·636E−04	0·318E−04	1·00
1·2	0·434E−09	0·218E−13	14·28	0·268E−04	0·332E−05	3·01
1·4	0·145E−08	0·189E−14	19·56	0·853E−06	0·277E−05	−1·70
1·6	0·270E−08	0·361E−15	22·84	0·440E−05	0·896E−06	2·30
1·8	0·422E−08	0·389E−15	23·37	0·128E−05	0·180E−06	2·83
2·0	0·607E−08	0·569E−15	23·35	0·368E−06	0·160E−06	1·20
3·0	0·235E−07	0·235E−14	23·26	0·347E−08	0·181E−08	0·93
4·0	0·701E−07	0·701E−14	23·25	0·614E−09	0·322E−10	4·25

and with $\gamma = 1, \tau = 0{\cdot}5, \nu = 1$ and $N = 10, 20, 40,$ and 80. Since by Table 1 $\|F - f_0\|_{L_1(0,4)} \approx 10^{-4}$, the error in F has a surprisingly small effect on the final error.

EXAMPLE 3 We now consider the spatially one-dimensional problem

$$u_t - u_{xx} = f(x,t), \quad \text{for } 0 < x < \pi,\ t > 0,$$
$$u(x,t) = 0, \quad \text{for } x = 0 \text{ and } \pi,\ t > 0, \quad \text{with } u(x,0) = u_0(x), \quad \text{for } 0 < x < \pi.$$

We choose f and u_0 so that the exact solution is

$$u(x,t) = (1+t)\mathrm{e}^{-t}\sin x + \cos t\,\mathrm{e}^{-2t}\sin 2x,$$

giving a forcing term in $\mathcal{F}(\Omega)$,

$$f(x,t) = \mathrm{e}^{-t}\sin x + \mathrm{e}^{-2t}(2\cos t - \sin t)\sin 2x$$

TABLE 7 *Error and apparent order of convergence for Example 2 by fitting* $f_0(t)$ *on* $[0, 4]$ *by functions of the form (5.1) with* $N = 10, 20, 40,$ *and* 80

t	ϵ_{10}	ϵ_{20}	ρ_{20}	ϵ_{40}	ρ_{40}	ϵ_{80}	ρ_{80}
0·2	0·324E−01	0·157E−01	1·04	0·672E−02	1·23	0·165E−02	2·03
0·4	0·238E−02	0·275E−02	− 0·21	0·146E−02	0·91	0·565E−03	1·37
0·6	0·626E−04	0·257E−05	4·61	0·133E−03	−5·69	0·683E−04	0·96
0·8	0·995E−03	0·905E−04	3·46	0·117E−05	6·28	0·722E−05	−2·63
1·0	0·571E−03	0·147E−04	5·28	0·313E−06	5·55	0·415E−05	−3·73
1·2	0·314E−03	0·342E−05	6·52	0·600E−06	2·51	0·158E−05	−1·40
1·4	0·215E−03	0·180E−05	6·90	0·383E−05	−1·09	0·375E−05	0·03
1·6	0·108E−03	0·691E−06	7·28	0·962E−06	−0·48	0·995E−06	−0·05
1·8	0·159E−04	0·392E−05	2·02	0·414E−05	−0·08	0·412E−05	0·01
2·0	0·146E−03	0·131E−05	6·79	0·130E−05	0·01	0·130E−05	0·00
3·0	0·128E−02	0·309E−05	8·70	0·340E−05	−0·14	0·340E−05	0·00
4·0	0·572E−02	0·147E−05	11·92	0·407E−07	5·18	0·407E−07	0·00

and $u_0(x) = \sin x + \sin 2x\cdot$ Thus

$$\widehat{f}(x, z) = \frac{1}{1+z}\sin x + \frac{2z+3}{(z+2)^2+1}\sin 2x,$$

with singularities at $z = -1$ and $-2 \pm \mathrm{i}$.

For the spatial problem in this case we solve (4.5) using piecewise linear finite-elements on a uniform mesh with spacing $h = \pi/M$, and with $M = 10, 20, 40,$ and 80. The fully discrete approximation $U_{N,h,\tau}$ is given by the expression (4.4), and the contour Γ is the same as in Example 1, i.e. with $\gamma = \tau = \nu = 0{\cdot}5\cdot$

Table 8 shows the maximum-norm of the errors ϵ_M with $N = 20$ for the discretization in time, and $M = 10, 20, 40$ and 80 for the finite-element discretization, together with apparent rates of convergence $\rho_M := \log_2(\epsilon_{M/2}/\epsilon_M)\cdot$ We note that the apparent order of convergence is very close to $O(h^2)$ even up to $M = 80$, once t is slightly greater than $\tau = 0{\cdot}5$. Since this is the theoretical rate of convergence for piecewise linear finite-elements, we observe, as expected from the results in Example 1, that the error for the time discretization with $N = 20$ is negligible compared with the error from the spatial approximation.

EXAMPLE 4 We now consider the spatially two-dimensional problem

$$\begin{aligned} u_t - \Delta u &= f(x,t), \quad \text{for } x \in \Omega = (0,\pi)\times(0,\pi),\ t \in (0,4], \\ u &= 0, \quad \text{for } x \in \partial\Omega,\ t \in [0,4], \quad \text{with } u(\cdot,0) = u_0, \quad \text{in } \Omega. \end{aligned} \tag{6.1}$$

We choose the data of the problem so that the exact solution is

$$u(x,t) = u(x_1,x_2,t) = \sin x_1\ \sin x_2\ (1 + x_1x_2 + t)^{-1},$$

which with $p(x) = 1 + x_1x_2$ corresponds to

$$f(x,t) = \sum_{l=1}^{3} \varphi_l(x)(p(x)+t)^{-l} \quad \text{and} \quad u_0(x) = \sin x_1\ \sin x_2\ p(x)^{-1}, \tag{6.2}$$

TABLE 8 *Error and apparent order of convergence for Example 3 with* $N = 20$ *and* $M = 10, 20, 40$ *and* 80

t	ϵ_{10}	ϵ_{20}	ρ_{20}	ϵ_{40}	ρ_{40}	ϵ_{80}	ρ_{80}
0·2	0·531E−01	0·318E−01	0·74	0·265E−01	0·26	0·252E−01	0·07
0·4	0·231E−01	0·468E−02	2·30	0·428E−02	0·13	0·418E−02	0·03
0·6	0·217E−01	0·554E−02	1·97	0·141E−02	1·97	0·378E−03	1·90
0·8	0·165E−01	0·432E−02	1·93	0·116E−02	1·89	0·371E−03	1·65
1·0	0·123E−01	0·316E−02	1·96	0·780E−03	2·02	0·180E−03	2·12
1·2	0·980E−02	0·249E−02	1·98	0·622E−03	2·00	0·155E−03	2·01
1·4	0·806E−02	0·204E−02	1·98	0·513E−03	1·99	0·130E−03	1·99
1·6	0·710E−02	0·179E−02	1·99	0·447E−03	2·00	0·112E−03	2·00
1·8	0·652E−02	0·164E−02	1·99	0·410E−03	2·00	0·103E−03	2·00
2·0	0·607E−02	0·153E−02	1·99	0·383E−03	2·00	0·957E−04	2·00
3·0	0·387E−02	0·974E−03	1·99	0·244E−03	2·00	0·609E−04	2·00
4·0	0·214E−02	0·544E−03	1·98	0·136E−03	2·00	0·338E−04	2·01

where

$$\varphi_1(x) = 2\sin x_1 \sin x_2,$$
$$\varphi_2(x) = -\sin x_1 \sin x_2 + 2(x_2 \cos x_1 \sin x_2 + x_1 \sin x_1 \cos x_2),$$
$$\varphi_3(x) = -2(x_1^2 + x_2^2)\sin x_1 \sin x_2.$$

For fixed $x \in [0, \pi] \times [0, \pi]$ the factors $(p(x) + t)^{-l}$ are completely monotonic. To allow comparison, we approximate these factors in two different ways, corresponding to two different approximations studied in Section 5. In the first (Gauss–Laguerre) approach we represent $(p(x) + t)^{-l}$, analogously to $f_l(t)$ in Section 5.2, as

$$(p(x) + t)^{-l} = (p(x) + \xi)^{-l} \frac{1}{(l-1)!} \int_0^\infty s^{l-1} \mathrm{e}^{-(t-\xi)s/(p(x)+\xi)} \mathrm{e}^{-s}\, \mathrm{d}s,$$

and then use the J-point Gauss–Laguerre approximation (5.6), with $\xi = 2$, to obtain

$$F(x,t) = \sum_{l=1}^{3} \varphi_l(x) \sum_{j=1}^{J} \mu_{l,j}(x) \mathrm{e}^{-\lambda_j(x)t} = \sum_{j=1}^{J} \psi_j(x) \mathrm{e}^{-\lambda_j(x)t},$$

$$\psi_j(x) = \sum_{l=1}^{3} \mu_{l,j}(x) \varphi_l(x),$$

where

$$\lambda_j(x) = s_j(p(x) + 2)^{-1}, \quad \mu_{l,j}(x) = w_j s_j^{l-1} (p(x) + 2)^{-l} \mathrm{e}^{2s_j/(p(x)+2)}/(l-1)!,$$

with s_j for $j = 1, \ldots, J$ being the zeros of the Laguerre polynomial of degree J and w_j the corresponding Gauss weights. The corresponding Laplace transform is

$$\widehat{F}(x,z) = \sum_{j=1}^{J} \psi_j(x)(z + \lambda_j(x))^{-1},$$

TABLE 9 *Errors and estimated orders of convergence for Example 4, using Gauss–Laguerre quadrature with* $J = 12$ *to approximate* $f(x, t)$, *and with* $N = 20$ *and* $M =$ 20, 40, 80, 160

t	ϵ_{20}	ϵ_{40}	ρ_{40}	ϵ_{80}	ρ_{80}	ϵ_{160}	ρ_{160}
0·20	0·644E−02	0·577E−02	0·16	0·562E−02	0·04	0·559E−02	0·01
0·40	0·418E−02	0·127E−02	1·71	0·109E−02	0·23	0·104E−02	0·06
0·60	0·346E−02	0·967E−03	1·84	0·257E−03	1·91	0·640E−04	2·01
0·80	0·283E−02	0·780E−03	1·86	0·206E−03	1·92	0·513E−04	2·01
1·00	0·235E−02	0·638E−03	1·88	0·167E−03	1·94	0·410E−04	2·02
1·20	0·199E−02	0·534E−03	1·90	0·139E−03	1·94	0·341E−04	2·03
1·40	0·170E−02	0·454E−03	1·91	0·117E−03	1·95	0·287E−04	2·03
1·60	0·148E−02	0·390E−03	1·92	0·101E−03	1·96	0·245E−04	2·04
1·80	0·130E−02	0·339E−03	1·93	0·871E−04	1·96	0·212E−04	2·04
2·00	0·115E−02	0·298E−03	1·94	0·761E−04	1·97	0·185E−04	2·04
3·00	0·752E−03	0·188E−03	2·00	0·470E−04	2·00	0·110E−04	2·10
4·00	0·626E−03	0·156E−03	2·00	0·387E−04	2·02	0·868E−05	2·15

and from (4.5) the discrete elliptic problems we need to solve are

$$(\nabla w_h, \nabla \chi) + z_j(w_h, \chi) = (u_0 + \widehat{F}(\cdot, z_j), \chi) \quad \forall \chi \in V_h,\ j = 0, \ldots, N-1. \tag{6.3}$$

Note that $\lambda_j(x) > 0$, thus all poles of $\widehat{F}(x, z)$ are on the negative real axis.

The spatial part of problem (6.3) is solved with bilinear elements of size $h \times h$, where $h = \pi/M$. This time the contour Γ is given by (1.5) and (1.7) with $\gamma = 1$ and $\nu = 0{\cdot}5$, so that Γ crosses the real axis at 0·5, ensuring that the contour passes well to the right of all singularities of $\widehat{F}\cdot$ Again the value of N in (4.4) is 20, that of τ is 0.5, and for the Gauss–Laguerre formula (5.6) we chose $J = 12$.

The errors in the maximum-norm and apparent rates of convergence are shown in Table 9. The results are broadly similar to those in Table 8, except that now by the time M is 80 we are beginning to see departures from the $O(h^2)$ convergence of the finite-element method. This indicates that either or both of the time discretization error with $N = 20$ and the Gauss–Laguerre quadrature error with $J = 12$ are becoming significant. Of course it is easy to reduce both of these errors to insignificance by increasing N and J moderately.

We next solve Example 4 by Prony's method. Again we use the representation (6.2), so that it is enough to have representations for the factors $(p(x) + t)^{-1}$, for $l = 1, 2, 3$. At first sight this seems to require a separate Prony calculation for each x, but in this special case the functional form of f is simple enough to allow us to do the Prony calculation only once for each l, with $p(x)$ replaced by its smallest value 1. To see this, for each l consider the scalar function

$$f_p(t) = (p + t)^{-l} = p^{-l} f_1(t/p), \quad \text{with } p \geqslant 1. \tag{6.4}$$

An approximation to this function on [0, 4] may be obtained once one is available for $f_1(t)$. For suppose that, with $F \in \mathcal{F}$, we have

$$\| f_1 - F \|_{L_1(0,4)} < \epsilon. \tag{6.5}$$

TABLE 10 *Error and apparent order of convergence for Example 4 using Prony's method with* $J = 8$ *and with* $N = 20$ *and* $M = 20, 40, 80,$ *and* 160

t	ϵ_{20}	ϵ_{40}	ρ_{40}	ϵ_{80}	ρ_{80}	ϵ_{160}	ρ_{160}
0·2	0·644E−02	0·577E−02	0·16	0·562E−02	0·04	0·559E−02	0·01
0·4	0·418E−02	0·127E−02	1·71	0·109E−02	0·23	0·104E−02	0·06
0·6	0·346E−02	0·967E−03	1·84	0·257E−03	1·91	0·640E−04	2·01
0·8	0·283E−02	0·780E−03	1·86	0·206E−03	1·92	0·513E−04	2·01
1·0	0·235E−02	0·638E−03	1·88	0·167E−03	1·94	0·410E−04	2·02
1·2	0·199E−02	0·534E−03	1·89	0·139E−03	1·94	0·341E−04	2·03
1·4	0·170E−02	0·454E−03	1·91	0·117E−03	1·95	0·287E−04	2·03
1·6	0·148E−02	0·390E−03	1·92	0·101E−03	1·96	0·245E−04	2·04
1·8	0·130E−02	0·339E−03	1·93	0·871E−04	1·96	0·212E−04	2·04
2·0	0·115E−02	0·298E−03	1·94	0·761E−04	1·97	0·185E−04	2·04
3·0	0·752E−03	0·188E−03	2·00	0·470E−04	2·00	0·110E−04	2·09
4·0	0·626E−03	0·156E−03	2·00	0·387E−04	2·02	0·868E−05	2·15

Then $p^{-l}F(t/p) \in \mathcal{F}$, and in view of (6.4) we have

$$\| f_p - p^{-l}F(\cdot/p)\|_{L_1(0,4)} = p^{-l+1}\| f_1 - F\|_{L_1(0,4/p)} \leqslant \| f_1 - F\|_{L_1(0,4)} < \epsilon.$$

We chose $J = 8$ for the Prony calculation of $(1+t)^{-l}$ for $l = 1, 2, 3$, and again we fixed $N = 20$. Table 10 shows the error behaviour as M doubles progressively from 20 to 160. The results are almost identical to those in Table 9, giving a strong indication that in both tables the contribution to tbe errors from the respective approximations to f are negligible compared to the errors from the finite-element discretization.

EXAMPLE 5 We now consider a two-dimensional problem in which the forcing term is not completely monotonic, and cannot be expressed as a finite sum of products of functions of space and time. We take the initial-value problem (6.1) with

$$f(x,t) = \frac{1}{1 + x_1x_2 + t^2}, \quad \text{for } t \in [0,4], \quad u(0) = 0.$$

For this example we use spatial discretization as in Example 4, with piecewise bilinear elements on squares of size $h \times h$, with $h = \pi/M$. The parameters of Γ are again chosen as $\tau = 0{\cdot}5, \gamma = 1, \nu = 0{\cdot}5$. Because the space and time dependence of $f(x,t)$ are in this example not separated, we apply the procedure described at the end of Section 4, using least squares to approximate $f_\xi(t) = f(\xi,t)$ by $F_\xi(t)$ of the form (5.1) for each vertex ξ of the finite-element discretization, with the Laplace transform given at the vertices by (5.2). For the additional contribution to the error bound from the first term of (4.11) we obtain $\int_0^4 \max_\xi |F_\xi(s) - f_\xi(s)|\,\mathrm{d}s = 1{\cdot}103 \times 10^{-5}$ with the maximum being taken over the 80×80 mesh points for $M = 80$. Table 11 shows the actual errors for the method. Since the exact solution for this problem is not known, we use the numerical solution with $M = 160$ and $N = 80$ as a reference solution. (This is thought to be the reason for the small departures from 2 in the columns for ρ_{40} and ρ_{80}.) The errors in Table 11 for this our most difficult example are clearly still dominated by the errors from the spatial discretization.

TABLE 11 *Error and apparent order of convergence for Example 5 using least-squares method with* $N = 20$ *and* $M = 10, 20, 40,$ *and* 80

t	ϵ_{10}	ϵ_{20}	ρ_{20}	ϵ_{40}	ρ_{40}	ϵ_{80}	ρ_{80}
0·2	0·169E−01	0·605E−02	1·48	0·188E−02	1·68	0·976E−03	0·949
0·4	0·183E−01	0·543E−02	1·76	0·138E−02	1·98	0·344E−03	2·00
0·6	0·173E−01	0·477E−02	1·86	0·119E−02	2·00	0·270E−03	2·14
0·8	0·154E−01	0·406E−02	1·93	0·102E−02	1·99	0·233E−03	2·14
1·0	0·134E−01	0·342E−02	1·96	0·853E−03	2·00	0·192E−03	2·15
1·2	0·114E−01	0·288E−02	1·98	0·710E−03	2·02	0·158E−03	2·17
1·4	0·967E−02	0·244E−02	1·99	0·594E−03	2·04	0·132E−03	2·17
1·6	0·829E−02	0·207E−02	2·00	0·505E−03	2·04	0·113E−03	2·16
1·8	0·717E−02	0·179E−02	2·00	0·437E−03	2·03	0·976E−04	2·16
2·0	0·625E−02	0·156E−02	2·00	0·381E−03	2·04	0·852E−04	2·16
3·0	0·337E−02	0·840E−03	2·00	0·205E−03	2·03	0·460E−04	2·16
4·0	0·194E−02	0·486E−03	2·00	0·118E−03	2·04	0·264E−04	2·16

7. Remarks on the numerical solution of the linear system

In our numerical examples in Section 6 in two space dimensions we used the direct Gauss elimination method to solve the discrete elliptic problems, as is appropriate because these have a very simple structure, given the square spatial domain. However, for problems in more complicated geometries it is more natural to apply iterative methods. We will leave a thorough discussion of the application of such methods in the present context to future work, but we want to make a few comments on the questions that arise. We now assume that A is a positive definite self-adjoint operator in a finite-dimensional inner product space V, with $\sigma(A) = \{\lambda_j\}_{j=1}^N$ ordered as an increasing positive sequence. (In applications of the type presented in Section 4 we would have $A = -\Delta_h$ and $V = V_h$, equipped with the norm in $L_2(\Omega)$.)

Let us first recall that the most basic iterative method for the equation

$$Ax = g, \tag{7.1}$$

is the Richardson iteration algorithm

$$x^{n+1} = x^n - \alpha(Ax^n - g) = (I - \alpha A)x^n + \alpha g, \tag{7.2}$$

for which the error reduction is described by

$$\|x^{n+1} - x\| \leqslant \|I - \alpha A\| \, \|x^n - x\|.$$

Here $\|I - \alpha A\| = \max_{\lambda \in \sigma(A)} |1 - \alpha\lambda|$, and the optimal choice of α is $2/(\lambda_1 + \lambda_N)$ which gives $\|I - \alpha A\| \leqslant (\kappa(A) - 1)/(\kappa(A) + 1)$, where $\kappa(A) = \lambda_N/\lambda_1$ is the spectral condition number of A. When $A = -\Delta_h$ is based on a quasi-uniform family of triangulations and linear finite-elements we have $\kappa(A) = O(h^{-2})$ and hence this shows that convergence factor satisfies

$$\|I - \alpha A\| \leqslant 1 - ch^2, \quad \text{with } c > 0.$$

In our time stepping method we want to solve instead the complex-valued equation

$$(z\,I + A)x = g, \quad \text{with } z = \varphi(y) + \mathrm{i}y \in \Gamma \subset \rho(-A), \tag{7.3}$$

where for simplicity we have taken the slope parameter $\sigma = 1$. The operator $z\,I + A$ is not Hermitian positive definite, and neither is its symmetric real part $\mathrm{Re}\, z\,I + A$ for $z \in \Gamma$, but $z\,I + A$ is still normal and nonsingular.

One way of solving (7.3) by iteration would be to first transform it to an equation with a positive definite operator by multiplying (7.3) by $(z\,I + A)^* = \bar{z}I + A$ to obtain

$$(|z|^2 I + 2\varphi(y)A + A^2)x = (\bar{z}\,I + A)g, \tag{7.4}$$

and then to apply Richardson iteration. For $z = \varphi(y) + \mathrm{i}y \in \rho(-A)$, the operator on the left in (7.4) is obviously positive definite, having eigenvalues $y^2 + (\varphi(y) + \lambda_k)^2 > 0$ for $k = 1, \ldots, N$, and so is positive definite. However, even though the distance between Γ and $\sigma(A)$ is bounded below, the condition number of the operator is of order $O(\lambda_N^2) = O(h^{-4})$, and we have for the basic iteration method with optimal choice of parameters

$$\|x^{n+1} - x\| \leqslant (1 - ch^4)\|x^n - x\|,$$

which indicates very slow convergence.

We now consider instead the basic Richardson iteration (7.2) applied directly to (7.3), i.e.

$$x^{n+1} = (I - \alpha(z\,I + A))x^n + \alpha g. \tag{7.5}$$

Here the issue is to choose α, which may now be complex, in such a way that

$$\|I - \alpha(z\,I + A)\| = \sup_{\lambda \in \sigma(A)} |1 - \alpha(z + \lambda)| < 1,$$

and as small as possible. Again $\sigma(A) \subset [\lambda_1, \lambda_N]$, and so $z + \lambda$ belongs to the line segment in the complex plane joining $z + \lambda_1$ and $z + \lambda_N$ for $\lambda \in \sigma(A)$. It is easily seen that $|1 - \alpha(z + \lambda)|$ is a convex function of λ, and hence

$$\sup_{\lambda \in \sigma(A)} |1 - \alpha(z + \lambda)| = \max(|1 - \alpha(z + \lambda_1)|,\ |1 - \alpha(z + \lambda_N)|). \tag{7.6}$$

We now take $z = \varphi(y) + \mathrm{i}y$ with $y \geqslant 0$, and write $\alpha = 1/(a + \mathrm{i}b)$. With α chosen to minimize the right-hand side of (7.6), it is easily established that the terms on the right-hand side of (7.6) are equal and that as a result $a = \varphi(y) + (\lambda_1 + \lambda_N)/2$, and b is the unique positive solution of

$$y\,b^2 - \Big(y^2 - (\varphi(y) + \lambda_1)(\varphi(y) + \lambda_N)\Big)b - y\left(\varphi(y) + \frac{\lambda_1 + \lambda_N}{2}\right)^2 = 0.$$

For λ_N large we find $b = O(\lambda_N)$, and hence

$$\|I - \alpha(z\,I + A)\| = \left|1 - \frac{z + \lambda_1}{a + \mathrm{i}b}\right| \leqslant 1 - \frac{c}{\lambda_N} \leqslant 1 - ch^2, \tag{7.7}$$

thus with a bound is of the same form as in the self-adjoint case, except that now the constant c depends on $z\cdot$

Even in the self-adjoint case the rate of convergence shown in (7.7) is relatively slow. One way to improve the convergence is to precondition the linear system by a positive definite self-adjoint operator B, such that the product BA has a more advantageous condition number. The preconditioned version of the basic equation (7.1) is

$$BAx = Bg,$$

and hence the corresponding iterative scheme

$$x^{n+1} = (I - \alpha BA)x^n + \alpha Bg.$$

It is known (see for example Bramble, 1993) that if, for instance, one assumes that

$$m(B^{-1}x, x) \leqslant (Ax, x) \leqslant M(B^{-1}x, x) \tag{7.8}$$

for some positive m and M, then for α chosen appropriately the iterative scheme converges geometrically with respect to a suitable energy norm. More precisely, setting

$$[x, y] = (B^{-1}x, y), \quad |||x||| = [x, x]^{1/2},$$

then BA is self-adjoint with respect to $[\cdot, \cdot]$, and by analogy with the above,

$$|||I - \alpha BA||| \leqslant (\kappa(BA) - 1)/(\kappa(BA) + 1) \leqslant (M - m)/(M + m),$$

if α is properly chosen. Note that by (7.8) $|||x|||$ is equivalent to the energy norm $(Ax, x)^{1/2}$.

We now discuss the application of the preconditioner B satisfying (7.8) to our complex system (7.3), which gives rise to the iterative scheme (see (7.5))

$$x^{n+1} = (I - \alpha G)x^n + \alpha Bg, \quad \text{where } G = G(z) := BA + zB.$$

Because $[\cdot, \cdot]$ is an inner product, we find

$$|||(I - \alpha G)x|||^2 = |||x|||^2 - 2\,\mathrm{Re}(\alpha[Gx, x]) + |\alpha|^2|||Gx|||^2.$$

Noting that both BA and B are self-adjoint with respect to $[\cdot, \cdot]$, it follows from (7.8) that $\lambda_{\max}(BA) \leqslant M$, and hence, if $\lambda_{\max}(B) = \Lambda$, that

$$\mathrm{Re}(\alpha[Gx, x]) = \mathrm{Re}\,\alpha\,[BAx, x] + \mathrm{Re}(\alpha z)\,[Bx, x] \quad \text{and} \quad |||Gx||| \leqslant (M + \Lambda|z|)|||x|||.$$

With Γ chosen as in (1.5) and (1.7), we now restrict attention to the case in which Γ crosses the real axis at or to the right of the origin. Then, for $z \in \Gamma$, let $z = |z|e^{i\psi}$ and $|\psi| < \pi$, and choose $\alpha = \rho e^{-i\zeta}$, with $\rho > 0$ and with $\zeta = \max(|\psi| - \frac{1}{2}\pi, 0)\,\mathrm{sign}\,\psi$. Then $|\arg(\alpha z)| = |\psi - \zeta| \leqslant \frac{1}{2}\pi$, so that $\mathrm{Re}(\alpha z)[Bx, x] \geqslant 0$. Using (7.8) we thus obtain

$$\mathrm{Re}(\alpha[Gx, x]) \geqslant \mathrm{Re}\,\alpha[BAx, x] \geqslant \rho m \cos\zeta\, |||x|||^2.$$

Note that if $|\psi| \leqslant \frac{1}{2}\pi$ (corresponding to $\mathrm{Re}\,z \geqslant 0$) we thus choose α real and positive, and if $\frac{1}{2}\pi < |\psi| < \pi$ (when $\mathrm{Re}\,z < 0$) we choose α so that $\mathrm{Re}(\alpha z) = 0$. Also note that in the

latter case $|\psi| \leqslant 3\pi/4$, giving $|\varphi| < \pi/4$. Altogether we thus obtain in operator norm, by optimizing in ρ,

$$\begin{aligned} |\|I - \alpha G|\|^2 &\leqslant 1 - 2\rho\, m \cos\zeta + (M + \Lambda|z|)^2\rho^2 \\ &= 1 - \Big(\frac{m\cos\zeta}{M + \Lambda|z|}\Big)^2, \quad \text{for } \rho = \frac{m\cos\zeta}{(M + \Lambda|z|)^2}, \end{aligned}$$

which again gives geometric convergence, if M and m are independent of N. The bound on $|\|I - \alpha G|\|^2$ approaches 1 as z approaches ∞, but this is not of significant concern, since the largest value of $|z|$ that occurs in the approximation (2.9) or (4.4), with $\zeta(y)$ chosen as in (1.7), is easily seen to be $|z_{N-1}| \approx \sqrt{2}\,\tau^{-1}\log(2N)$, growing only logarithmically with N. The convergence ratio $(1 - (m\cos\varphi/(M + \Lambda|z|))^2)^{1/2}$ also depends on Λ. In the ideal case in which the preconditioner is $B = A^{-1}$ we have $\Lambda = \lambda_1^{-1}$, where λ_1 is the smallest eigenvalue of A, which in turn is bounded below by the smallest eigenvalue of $-\Delta$. More realistic choices of B should be selected in such a way that it is still true that $\Lambda = \lambda_{\max}(B)$ is bounded above as $h \to 0$.

Acknowledgements

The support of the Australian Research Council is gratefully acknowledged. In particular, in 1999 D. Sheen was supported at the University of New South Wales by an International Research Fellowship of the Australian Research Council. He was also supported in part by the Korea Research Foundation (KRF-2000-DS0004).

REFERENCES

AGMON, S., DOUGLIS, A. & NIRENBERG, L. (1950) Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. *Comm. Pure Appl. Math.*, **12**, 623–727.

BAKAEV, YU N., THOMÉE, V. & WAHLBIN, L. B. (2001) Maximum-norm estimates for resolvents of elliptic finite element operators. *Math. Comp.*, to appear.

BRAMBLE, J. H. (1993) Multigrid Methods. *Pitman Research Notes in Mathematics, 294.*

DAVIS, P. J. & RABINOWITZ, P. (1975) *Methods of Numerical Integration*. New York: Academic.

DE PRONY, R. (1795) Essai expérimentale et analytique. *J. Écol. Polytech. (Paris)*, **1**, 24–76.

KAMMLER, D. W. (1977) Prony's method for completely monotonic functions. *J. Math. Analysis and Appl.*, **57**, 560–570.

LUBICH, C., SLOAN, I. H. & THOMÉE, V. (1996) Nonsmooth data error estimates for approximations of an evolution equation with a positive type memory term. *Math. Comp.*, **65**, 1–17.

SCHATZ, A. H. & WAHLBIN, L. B. (1982) On the quasi-optimality in l_∞ of the H^1-projection into finite-element spaces. *Math. Comp.*, **38**, 1–22.

SHEEN, D., SLOAN, I. H. & THOMÉE, V. (1999) A parallel method for time-discretization of parabolic problems based on contour integral representation and quadrature. *Math. Comp.*, **69**, 177–195.

STEWART, H. B. (1974) Generation of analytic semigroups by strongly elliptic operators. *Trans. Amer. Math. Soc.*, **199**, 141–161.

STOER, J. & BULIRSCH, R. (1992) *Introduction to Numerical Analysis*, volume 12 of *Texts in Applied Mathematics*, Second edition. Berlin: Springer

TALBOT, A. (1979) The accurate numerical inversion of Laplace transforms. *J. Inst. Maths. Applics.*, **23**, 97–120.

THOMÉE, V. (1997) *Galerkin Finite Element Methods for Parabolic Problems*, volume 25 of Springer Series in Computational Mathematics. Berlin: Springer.

THOMÉE, V. & WAHLBIN, L. (2000) Stability and analyticity in maximum norm for simplicial Lagrange finite-element semidiscretizations of parabolic equations with Dirichlet boundary conditions. *Numer. Math.*, **87**, 373–389.

WIDDER, D. V. (1941) *The Laplace Transform*. Princeton, NJ: Princeton University Press.