ANZIAM J. 53(2011), 134–155
doi:10.1017/S1446181112000107

ITERATIVE SOLUTION OF SHIFTED POSITIVE-DEFINITE LINEAR SYSTEMS ARISING IN A NUMERICAL METHOD FOR THE HEAT EQUATION BASED ON LAPLACE TRANSFORMATION AND QUADRATURE

WILLIAM MCLEAN[✉1] and VIDAR THOMÉE[2]

(Received 14 February, 2012; revised 8 May, 2012)

Abstract

In earlier work we have studied a method for discretization in time of a parabolic problem, which consists of representing the exact solution as an integral in the complex plane and then applying a quadrature formula to this integral. In application to a spatially semidiscrete finite-element version of the parabolic problem, at each quadrature point one then needs to solve a linear algebraic system having a positive-definite matrix with a complex shift. We study iterative methods for such systems, considering the basic and preconditioned versions of first the Richardson algorithm and then a conjugate gradient method.

2010 *Mathematics subject classification*: primary 65F10; secondary 65M22, 65M60, 65R10.

Keywords and phrases: Laplace transform, finite elements, quadrature, Richardson iteration, conjugate gradient method, preconditioning.

1. Introduction

Let V be a complex finite-dimensional inner product space, and let A be a positive-definite Hermitian linear operator in V. We consider iterative methods for linear equations of the form

$$zw + Aw = g \quad \text{with } |\arg z| < \pi. \tag{1.1}$$

Such equations, with a complex shift z of the positive-definite operator A, need to be solved in a method for discretization in time of parabolic equations, based on Laplace transformation and quadrature, which has been studied recently, as is made

[1]School of Mathematics and Statistics, The University of New South Wales, Sydney 2052, Australia; e-mail: w.mclean@unsw.edu.au.
[2]Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, S-412 96 Gothenburg, Sweden; e-mail: thomee@chalmers.se.

more specific below. We consider a basic Richardson iteration and a conjugate gradient (CG) method for (1.1), as well as preconditioned versions of these methods. An earlier eprint [13] is a preliminary, somewhat more extensive and complete version of the present work, and is sometimes referred to here for additional details.

We now sketch the numerical approximation method [9] for the heat equation that leads to (1.1). Given a spatial domain $\Omega \subset \mathbb{R}^d$, an elliptic operator $Lu = -\nabla \cdot (a\nabla u)$, initial data $u_0 = u_0(x)$ and an inhomogeneous term $f = f(x,t)$, let $u = u(x,t)$ be the solution of the parabolic initial boundary-value problem

$$\begin{aligned} \partial_t u + Lu &= f \text{ in } \Omega \quad \text{with } u = 0 \text{ on } \partial\Omega \text{ for } t > 0, \\ u &= u_0 \text{ in } \Omega \quad \text{when } t = 0, \end{aligned} \tag{1.2}$$

where $\partial_t u = \partial u/\partial t$. For simplicity, we assume that the diffusivity a is a positive constant, and that Ω is a convex polygonal (if $d = 2$) or polyhedral (if $d \geq 3$) domain. In the usual weak formulation, we view the solution u as a function of t taking values in the Sobolev space $H_0^1(\Omega)$ and satisfying

$$(\partial_t u, v) + a(\nabla u, \nabla v) = (f, v) \quad \text{for all } v \in H_0^1(\Omega) \text{ and } t > 0,$$

with $u(0) = u_0$, where $(v, w) = \int_\Omega v\bar{w}\, dx$ is the inner product in $L_2(\Omega)$. To discretize in space, let $\{V_h\} \subset H_0^1(\Omega)$ be a family of continuous, piecewise linear finite-element spaces, based on a family of regular partitions $\mathcal{T}_h = \{\tau\}$ of Ω. The standard Galerkin spatially semidiscrete solution $u_h : [0,\infty) \to V_h$ satisfies

$$(\partial_t u_h, \chi) + a(\nabla u_h, \nabla \chi) = (f, \chi) \quad \text{for all } \chi \in V_h \text{ and } t > 0,$$

with $u_h(0) = u_{0h}$, where $u_{0h} \in V_h$ is a suitable approximation of the initial data.

Introducing the discrete elliptic operator $L_h : V_h \to V_h$ defined by

$$(L_h\psi, \chi) = a(\nabla\psi, \nabla\chi) \quad \text{for all } \psi, \chi \in V_h,$$

the spatially semidiscrete initial-value problem may also be written

$$\partial_t u_h + L_h u_h = P_h f \quad \text{for } t > 0 \text{ with } u_h(0) = u_{0h}, \tag{1.3}$$

where P_h is the L_2-projection onto V_h, and we may choose, for instance, $u_{0h} = P_h u_0$.

We denote the Laplace transform of a function $v(t)$ by

$$\hat{v}(z) = \mathcal{L}v(z) := \int_0^\infty e^{-zt} v(t)\, dt,$$

and recall the inversion formula

$$v(t) = \mathcal{L}^{-1}\hat{v}(t) = \frac{1}{2\pi i} \int_\Gamma e^{zt} \hat{v}(z)\, dz,$$

where Γ is any contour homotopic to the imaginary axis and to the right of all singularities of $\hat{v}$. Putting $w_h := \hat{u}_h$ and $g(z) := u_{0h} + \hat{f}(z)$, we find by application of

the Laplace transform to (1.3) that

$$(zI + L_h)w_h(z) = P_h g(z) \quad \text{for } z \in \Gamma, \tag{1.4}$$

or, in weak form,

$$z(w_h(z), \chi) + a(\nabla w_h(z), \nabla \chi) = (g(z), \chi) \quad \text{for all } \chi \in V_h. \tag{1.5}$$

By the inversion formula above, the solution of (1.3) may now be represented by

$$u_h(t) = \frac{1}{2\pi i} \int_\Gamma e^{zt} w_h(z)\, dz \quad \text{where } w_h(z) = (zI + L_h)^{-1} P_h g(z).$$

In our presentation, we use for Γ the curve parameterized by

$$z(\xi) := 1 - \cosh \xi + i \sinh \xi \quad \text{for } -\infty < \xi < \infty,$$

which is the left branch of the hyperbola $(x-1)^2 - y^2 = 1$, for $z = x + iy$, and thereby represent $u_h(t)$ as an integral along the real axis,

$$u_h(t) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} e^{z(\xi)t} w_h(z(\xi)) z'(\xi)\, d\xi.$$

This integrand decays rapidly if $t > 0$, and the time discretization is effected by applying an equal-weight quadrature rule to obtain our fully discrete numerical solution

$$U_{q,h}(t) := \frac{k}{2\pi i} \sum_{j=-q}^{q} e^{z_j t} w_h(z_j) z_j', \tag{1.6}$$

where, for an appropriate $k > 0$ that depends on the time scale of interest,

$$\xi_j := jk, \quad z_j := z(\xi_j), \quad z_j' := z'(\xi_j) \quad \text{for } |j| \le q. \tag{1.7}$$

Here we choose $k = q^{-1} \log q$, which is suitable for t of order 1, implying that

$$z_{\pm q} = 1 - \frac{q + q^{-1}}{2} \pm i \frac{q - q^{-1}}{2} \approx \frac{-1 \pm i}{2} q.$$

In this approach, the main computational task is to solve, at $z = z_j$ for each j, the discrete elliptic equation (1.4), which is of the form (1.1) with $z = z_j$, with $V = V_h$ equipped with the L_2 inner product, and $A = L_h$. Fortunately, these linear systems are independent for different j, so the solutions $w_h(z_j)$ may be computed in parallel.

Employing a standard nodal basis, the weak form (1.5) leads to a linear system $z\mathcal{M}\boldsymbol{w} + \mathcal{S}\boldsymbol{w} = \boldsymbol{g}$, where $\mathcal{M}$ and $\mathcal{S}$ are the mass and stiffness matrices, $\boldsymbol{w}$ is the vector of nodal values of $w_h(z)$, and $\boldsymbol{g}$ is the load vector for the source term $g(z)$. Equivalently, $z\boldsymbol{w} + \mathcal{M}^{-1}\mathcal{S}\boldsymbol{w} = \mathcal{M}^{-1}\boldsymbol{g}$, which again has the form (1.1), and the operator $A = \mathcal{M}^{-1}\mathcal{S}$ is Hermitian and positive definite with respect to the inner product induced by $\mathcal{M}$, which corresponds to the L_2 inner product in S_h.

Applying an iterative solver to (1.1) yields an approximate finite-element solution $\tilde{w}_h$. Replacing w_h with $\tilde{w}_h$ in the quadrature sum (1.6), we arrive at the computed approximation $\widetilde{U}_{q,h}$, and using the triangle inequality we can estimate the overall error in $L_2(\Omega)$ as the sum of the solver error, the quadrature error and the finite-element error:

$$\|\widetilde{U}_{q,h}(t) - u(t)\| \le \|\widetilde{U}_{q,h}(t) - U_{q,h}(t)\| + \|U_{q,h}(t) - u_h(t)\| + \|u_h(t) - u(t)\|.$$

Under appropriate assumptions on the data [9], there is a positive constant c such that

$$\|U_{q,h}(t) - u_h(t)\| = O(e^{-cq/\log q}) \quad \text{and} \quad \|u_h(t) - u(t)\| = O(h^2), \tag{1.8}$$

for $0 < t_0 \le t \le T < \infty$, and if $\|\tilde{w}_h(z_j) - w_h(z_j)\| \le \varepsilon_j$ then

$$\mathcal{E}(t) := \|\tilde{U}_{q,h}(t) - U_{q,h}(t)\| \le \frac{k}{2\pi} \sum_{j=-q}^{q} \varepsilon_j e^{\mathrm{Re}(z_j)t} |z_j'|. \tag{1.9}$$

We may use these estimates as the basis for a stopping criterion: in view of (1.8), it is desirable to choose the solver tolerance ε_j in such a way that $\mathcal{E}(t) \le C(h^2 + e^{-cq/\log q})$ for some constant C. The presence of the factor $e^{\mathrm{Re}(z_j)t}|z_j'|$ allows ε_j to increase with $|j|$; see (4.1) below and remember that $\mathrm{Re}(z_j) < 0$.

The program for time discretization of parabolic equations sketched above was initiated by Sheen et al. [16, 17], and continued by Gavrilyuk and Makarov [5], McLean et al. [9] and McLean and Thomée [10–12], and the error in (1.6) was analysed in both $L_2(\Omega)$ and $L_\infty(\Omega)$, under various assumptions on the data of the problem. In the latter papers [10–12], fractional-order diffusion equations were also treated.

In our earlier papers [9–12], the analysis was illustrated by numerical examples. These were carried out in simple cases, in one space dimension and also in the case of a square spatial domain in two dimensions, and direct solvers were used for the linear system (1.5). However, even though powerful direct solvers are available, for large-size problems in more complicated geometries, particularly if the spatial dimension $d = 3$, it becomes natural to apply iterative methods, and our purpose in this paper is therefore to study such methods for equations of the form (1.1), with application to our above method for the heat equation in mind. Some preliminary results on this problem were sketched by Sheen et al. [17] using the Richardson iteration algorithm, and also for a preconditioned form of this method.

Equations of the form (1.1) have been widely studied in the numerical linear algebra literature. For instance, in connection with the spatial discretization of the Helmholtz equation, Freund [4] analysed Krylov methods for (1.1) that generalize some iterative techniques for real and symmetric, but indefinite, systems. The GAL method from Freund's paper is a generalization of the SYMMLQ method of Paige and Saunders [15], and computes in a different way the same sequence of iterates as our basic CG method. Our algorithm is somewhat simpler and cheaper, but we require

that arg z is bounded away from $\pm\pi$, a restriction that holds for our application but not typically for the Helmholtz equation.

In a number of applications, one wants to solve (1.1) for many values of z with the same right-hand side g. For instance, in our problem (1.5), g is independent of z when $f \equiv 0$ in (1.2). Several iterative methods take advantage of the fact that in this case the Krylov subspaces are independent of z [14, 18, 19]. Recently, in 't Hout and Weideman [7] used a Laplace transform technique, similar to the one considered here, for the time discretization of the Black–Scholes and Heston equations, in combination with a spatial discretization by finite differences. They described an iterative method for (1.1) with a real, nonsymmetric A, in which the Arnoldi algorithm is applied once to an operator of the form $(\mu I + A)^{-1}$ for a suitable $\mu > 0$. The solution w for each z can then be found by solving a small, upper-Hessenberg system. This approach is efficient provided the cost of obtaining a sparse factorization of $\mu I + A$ is acceptable.

Another approach, not discussed here, is to reformulate the complex linear system as an equivalent real one with twice as many equations and unknowns. See, for example, the paper by Benzi and Bertaccini [2] and the list of references therein.

We study Richardson iteration with a complex acceleration parameter α in Section 2, extending and improving upon the results of Sheen et al. [17]. Let λ_j denote the jth eigenvalue of A, labelled so that $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_N$. From a knowledge of the extremal eigenvalues λ_1 and λ_N, we can determine the optimal α, in the sense of minimizing the associated error reduction factor. For the finite-element problem on quasiuniform partitions with maximal mesh size h, recall that λ_1 is bounded below by the minimum eigenvalue of the continuous elliptic operator L, whereas $\lambda_N = O(h^{-2})$. As a consequence, the basic Richardson method converges slowly, with an error reduction factor $1 - ch^2$, for $c > 0$, depending on z and growing with $|z|$. We extend the analysis to a preconditioned Richardson iteration, using the special preconditioner $B_z = (\mu_z I + A)^{-1}$, where $\mu_z > -\lambda_1$, and show that the error reduction factor is now bounded away from 1 as $\lambda_N \to \infty$. We also determine the optimal μ_z, and find that $\mu_z \approx |z + \lambda_1| - \lambda_1$ for large λ_N. In the case of a more general preconditioner B_z, we quote a result from our aforementioned eprint [13] showing geometric convergence in the norm $|[v]| := (B_z^{-1} v, v)^{1/2}$, where the acceleration parameter is defined in terms of bounds for the spectrum of $B_z(\mu_z I + A)$.

In Section 3 we analyse a CG method, which does not involve choosing an acceleration parameter. Generalizing the usual convergence analysis to allow the complex shift of A in (1.1), we show geometric convergence of the iterates w_n, which follows from the error bound

$$|||w_n - w||| \le \frac{\sec(\frac{1}{2}\arg z)}{|T_n(-s_z)|}\, |||w_0 - w||| \quad \text{with } s_z := \frac{\lambda_1 + \lambda_N + 2z}{\lambda_N - \lambda_1}, \tag{1.10}$$

where $|||v|||^2 := |z|\,\|v\|^2 + (Av, v)$ and T_n is the Chebyshev polynomial of degree n. Since $T_n(-s_z) = (\eta_z^n + \eta_z^{-n})/2$ with $|\eta_z| < 1$, this indicates geometric convergence with rate $|\eta_z|^n$. For the finite-element problem discussed above, we find that $|\eta_z| \le 1 - ch$, with $c > 0$ depending on z, giving a better convergence rate than Richardson iteration.

In the case of the special preconditioner $B_z = (\mu_z I + A)^{-1}$, by letting $\tilde{z} := (z - \mu_z)^{-1}$ we can write the preconditioned equation as $\tilde{z}w + B_z w = \tilde{z}B_z g$, which has the same form as the original equation (1.1). The error reduction factor is again bounded away from 1 as $\lambda_N \to \infty$, and is now smaller than for the corresponding preconditioned Richardson iteration. The optimal choice of μ_z turns out to be the same for both of these methods.

It is natural to also consider more general preconditioners for the CG iteration. The preconditioned equation $zB_z w + B_z A w = B_z g$ is again equivalent to an equation of the form (1.1), namely $zv + B_z^{1/2} A B_z^{-1/2} v = B_z^{1/2} g$, where $v = B_z^{1/2} w$, and the transformed operator $B_z^{1/2} A B_z^{-1/2}$ is Hermitian and positive definite with respect to the inner product $[v, w] = (B_z^{-1} v, w)$. However, computing the action of $B_z^{\pm 1/2}$ is usually costly, so we instead work with the preconditioned equation in its original form. Although the error is still optimal in a certain sense, we are not able to show a precise error bound of the type (1.10).

Section 3 concludes with a short discussion of the algorithmic implementation of the CG method. Both the basic method and the method using the special preconditioner $B_z = (\mu_z I + A)^{-1}$ admit a simple and inexpensive recursion for the successive iterates. However, for a more general preconditioner B_z, the nth step of the iteration requires us to compute a sum of n vectors, which becomes expensive for large n. For this reason, we recommend choosing a high-quality preconditioner that leads to convergence after only a few iterations.

We illustrate our error analysis in Section 4 by numerical calculations in a concrete case of (1.2), and discuss how to choose the parameters to balance the contributions to the error from the discretizations in space and time and in the iterative procedure.

2. Iteration algorithms of Richardson type

We now assume, as in (1.1), that A is a positive-definite Hermitian operator in a finite-dimensional complex inner product space V, with extremal eigenvalues λ_1 and λ_N, and for brevity put $A_z := zI + A$. In this section, following Sheen et al. [17], we consider first the basic Richardson iteration with acceleration parameter $\alpha \in \mathbb{C}$, applied to $A_z w = g$:

$$w^{n+1} = (I - \alpha A_z) w^n + \alpha g \quad \text{for } n \geq 0 \text{ with } w^0 \text{ given.} \tag{2.1}$$

The error reduction in each time step is then described by the inequality

$$\|w^{n+1} - w\| \leq \|I - \alpha A_z\| \, \|w^n - w\|,$$

and since A_z is a normal operator in V,

$$\|I - \alpha A_z\| = \max_{1 \leq j \leq N} |1 - \alpha(z + \lambda_j)|. \tag{2.2}$$

In (2.1), in addition to choosing w^0, the issue is to select $\alpha \in \mathbb{C}$ so that the norm in (2.2) is as small as possible. For $z = 0$, as is well known, the optimal choice of α

is $2/(\lambda_1+\lambda_N)$, which gives

$$\|I-\alpha A\| = \frac{\kappa(A)-1}{\kappa(A)+1} \quad \text{where } \kappa(A) := \frac{\lambda_N}{\lambda_1}. \tag{2.3}$$

Recall that in the case of the finite-element equations (1.5) based on a quasiuniform family of triangulations $\mathcal{T}_h$, we have $A = L_h$ and $\kappa(A) = O(\lambda_N) = O(h^{-2})$, so

$$\|I-\alpha A\| \le 1 - ch^2 \quad \text{with } c>0. \tag{2.4}$$

The following was shown by Sheen et al. [17]; a proof is included for completeness.

THEOREM 2.1. *Let* $|\arg z| < \pi$. *Then the optimal error reduction factor in* (2.1) *satisfies*

$$\varepsilon_z := \min_\alpha \|I-\alpha A_z\| \le 1 - c\lambda_N^{-1} \quad \textit{with } c = c(z,\lambda_1) > 0.$$

PROOF. We begin by demonstrating that, for $\lambda_N \le \Lambda < \infty$, there exists an α such that $\|I-\alpha A_z\| \le c(z,\lambda_1,\Lambda) < 1$. In fact, we first rotate the line segment $[z+\lambda_1, z+\lambda_N]$ by $\pm\pi$ so that it becomes parallel and to the right of the imaginary axis, and then multiply by a suitable number $\rho>0$ so that the resulting line segment is inside the disk $|z-1|<1$. In view of (2.2), this shows our initial claim with $\alpha = \pm i\rho$.

To handle the case when λ_N is large, write $z = x+iy$ and put

$$\sigma := x + \frac{\lambda_1+\lambda_N}{2} \quad \text{and} \quad \tau := (x+\lambda_1)(x+\lambda_N) - y^2.$$

Using (2.2), we find that

$$\varepsilon_z = \max\{|1-\alpha(z+\lambda_1)|, |1-\alpha(z+\lambda_N)|\},$$

and this expression is minimized when the two terms in the "max" are equal, implying that $1/\alpha = \sigma + is$. Since

$$\varepsilon_z^2 = \left|1 - \frac{z+\lambda_1}{\sigma+is}\right|^2 = 1 - R(s) \quad \text{where } R(s) := \frac{2sy+\tau}{s^2+\sigma^2},$$

we seek $s\in\mathbb{R}$ such that the rational function $R(s)$ is as large as possible. For $y=0$ and $x>0$, the maximum occurs when $s=0$, giving

$$\varepsilon_x = \frac{\kappa_x - 1}{\kappa_x + 1}, \quad \kappa_x = \frac{x+\lambda_N}{x+\lambda_1}, \quad \alpha_x = \frac{2}{(x+\lambda_N)+(x+\lambda_1)},$$

consistent with (2.3). If $y\ne 0$ then $R'(s)=0$ at the roots

$$s_\pm := \frac{-\tau \pm \sqrt{\tau^2+4y^2\sigma^2}}{2y}$$

of the quadratic equation $ys^2 + \tau s - y\sigma^2 = 0$. Since $R(s)\to 0$ as $|s|\to\infty$, by considering the sign of $R(s)$ for $s\to\infty$ and $s\to-\infty$, we find that the maximum occurs at $s=s_+$ (and the minimum at s_-), both when $y>0$ and when $y<0$,

TABLE 1. Error reduction by Theorems 2.1 and 2.2.

j	x_j	y_j	ρ_z	φ_z	ε_z	$\tilde{\rho}_z$	$\tilde{\varphi}_z$	μ_z	$\tilde{\varepsilon}_z$
0	0.00	0.00	4.99×10^{-4}	0.00	0.9995	1.000	0.00	0.00	0.000
2	−0.05	0.30	4.93×10^{-4}	0.15	0.9995	0.988	0.15	0.00	0.152
4	−0.18	0.64	4.73×10^{-4}	0.33	0.9995	0.947	0.33	0.03	0.321
6	−0.43	1.02	4.31×10^{-4}	0.53	0.9996	0.864	0.53	0.16	0.503
8	−0.81	1.51	3.76×10^{-4}	0.72	0.9996	0.754	0.72	0.51	0.658
10	−1.35	2.12	3.24×10^{-4}	0.86	0.9995	0.650	0.86	1.14	0.760
12	−2.10	2.93	2.85×10^{-4}	0.96	0.9995	0.572	0.96	2.12	0.821
14	−3.13	4.01	2.58×10^{-4}	1.03	0.9994	0.517	1.03	3.53	0.856
16	−4.54	5.45	2.39×10^{-4}	1.07	0.9993	0.478	1.07	5.49	0.878
18	−6.45	7.38	2.25×10^{-4}	1.10	0.9991	0.452	1.10	8.18	0.892
20	−9.02	9.97	2.16×10^{-4}	1.12	0.9988	0.433	1.12	11.85	0.902

giving $\alpha = \alpha_z := 1/(\sigma + is_+)$. If λ_N is large then $2s_+y \approx b\lambda_N$, where

$$b = \sqrt{(x+\lambda_1)^2 + y^2} - (x+\lambda_1) = |z+\lambda_1| - (x+\lambda_1) > 0,$$

and hence

$$\varepsilon_z^2 = 1 - R(s_+) \approx 1 - \frac{|z+\lambda_1|}{b^2+y^2}\,\frac{4y^2}{\lambda_N}. \qquad \square$$

When, as above, $A = L_h$ and $\lambda_N = O(h^{-2})$, the error bound has the same form as in (2.4), except that now the constant c depends on z.

The rate of convergence shown in Theorem 2.1 is too slow for the iteration (2.1) to be of practical use. On the left-hand side of Table 1, we show the values of ρ_z and φ_z for the optimal parameter $\alpha_z = \rho_z e^{-i\varphi_z}$, computed as in the proof of Theorem 2.1, and the error reduction factor $\varepsilon_z = \|I - \alpha_z A\|$ with $z = z_j = x_j + iy_j \in \Gamma$ as in (1.7) and $k = q^{-1}\log q$, for even j in the range $0 \le j \le q = 20$. Here, the operator $A = L_h$ is from the model problem described in Section 4, for which $\lambda_1 \approx 1$ and $\lambda_N \approx 4000$.

One way to improve the convergence of the iterative method (2.1), considered briefly by Sheen et al. [17], is to precondition the linear system by multiplication by a positive-definite Hermitian operator B_z, which, in contrast to the choice in that paper, we here allow to depend on z. Rewriting (1.1) as

$$G_z w = \tilde{g}_z := B_z g, \quad \text{where } G_z := B_z A_z, \tag{2.5}$$

the Richardson iteration algorithm becomes

$$w^{n+1} = (I - \alpha G_z)w^n + \alpha \tilde{g}_z. \tag{2.6}$$

We first consider the special preconditioner $B_z = (\mu_z I + A)^{-1}$, where $\mu_z > -\lambda_1$; for $\mu_z = 0$ we have $B_z = A^{-1}$, independently of z. Since

$$G_z = G_z(A, \mu_z) = (\mu_z I + A)^{-1}(zI + A)$$

is a normal operator, the error reduction is now measured by

$$\|I - \alpha G_z(A, \mu_z)\| = \max_{1\le j\le N} |1 - \alpha G_z(\lambda_j, \mu_z)| \quad \text{where } G_z(\lambda, \mu_z) = \frac{z+\lambda}{\mu_z+\lambda},$$

and we want to choose α and μ_z so that this quantity is as small as possible. In practice we are interested only in $z = z_j$ with $\operatorname{Re} z_j \ge \operatorname{Re} z_q \approx -q/2$ and $q \ll \lambda_N$, so the assumption $|z + \lambda_N| > |z + \lambda_1|$ is not restrictive. We show the following.

THEOREM 2.2. *Let* $|\arg z| < \pi$ *and* $|z + \lambda_N| > |z + \lambda_1|$. *Then, for the optimal error reduction factor in* (2.6)*, uniformly in* λ_N*,*

$$\tilde\varepsilon_z = \tilde\varepsilon_z(\mu_z) := \min_{\alpha\in\mathbb{C}} \|I - \alpha G_z(A, \mu_z)\| < 1. \tag{2.7}$$

This quantity is minimized in μ_z *when* $|G_z(\lambda_1, \mu_z)| = |G_z(\lambda_N, \mu_z)|$*, or for*

$$\mu_z = -\lambda_1 + \frac{\lambda_N - \lambda_1}{|\kappa_z| - 1} > -\lambda_1 \quad \textit{where } \kappa_z := \frac{z+\lambda_N}{z+\lambda_1}. \tag{2.8}$$

For this choice of μ_z*,*

$$\tilde\varepsilon_z = \sin|\tilde\varphi_z| \quad \textit{where } \tilde\varphi_z = \frac{\varphi_{z,1} - \varphi_{z,N}}{2} \textit{ and } \varphi_{z,j} = \arg(z + \lambda_j).$$

PROOF. For μ_z given, let $\mathsf{a}_j = G_z(\lambda_j, \mu_z)$. We note that, since $G_z(\infty, \mu_z) = 1$ and $(\mu_z + \lambda)^{-1}$ is positive and decreasing in λ,

$$G_z(\lambda, \mu_z) = 1 + (\mu_z + \lambda)^{-1}(z - \mu_z) \in [\mathsf{a}_1, \mathsf{a}_N] \subseteq [\mathsf{a}_1, 1] \quad \text{for } \lambda \in [\lambda_1, \lambda_N],$$

and

$$F(\alpha) := \|I - \alpha G_z(A, \mu_z)\| = \max\{|1 - \alpha\mathsf{a}_1|, |1 - \alpha\mathsf{a}_N|\}.$$

We first show that α may be chosen so that the line segment $\alpha[\mathsf{a}_1, 1]$ is inside the disk $|z - 1| < 1$. For this, we first determine $\arg\alpha$ by rotating the line segment $[\mathsf{a}_1, 1]$ around the origin so that it becomes parallel to and to the right of the imaginary axis. We then shrink the line segment thus obtained so that it comes inside the indicated disk, which gives $|\alpha|$. With this choice of α, $F(\alpha) < 1$ uniformly in λ_N, and thus (2.7) holds uniformly in λ_N.

For μ_z given, the optimal α satisfies $|1 - \alpha\mathsf{a}_1| = |1 - \alpha\mathsf{a}_N|$, and thus also $|1/\alpha - \mathsf{a}_1| = |1/\alpha - \mathsf{a}_N|$. Therefore, $1/\alpha$ has to be chosen on the line in $\mathbb{C}$ through the midpoint $\mathsf{c} = (\mathsf{a}_1 + \mathsf{a}_N)/2$ which is perpendicular to $\mathsf{a}_N - \mathsf{a}_1$, or has the direction of $\mathsf{d} = i(\mathsf{a}_N - \mathsf{a}_1)$, so that $\alpha = \alpha(s) = 1/(\mathsf{c} + s\mathsf{d})$. Then for α to be optimal, $s \in \mathbb{R}$ has to minimize the rational function $F(\alpha(s))^2$.

Let $\alpha = \rho e^{-i\varphi}$ be the optimal α, and let $\ell(\phi)$ be the line through the origin and $e^{i\phi}$. Then α, $\mathsf{a}_j \in \ell(\varphi_{z,j} - \varphi)$, and hence

$$F(\alpha) \geq \max\{\operatorname{dist}(1, \ell(\varphi_{z,1} - \varphi)), \operatorname{dist}(1, \ell(\varphi_{z,N} - \varphi))\}.$$

The latter quantity is smallest when $\varphi_{z,N} - \varphi = -(\varphi_{z,1} - \varphi)$, or when $\varphi = (\varphi_{z,1} + \varphi_{z,N})/2$, in which case $\varphi_{z,1} - \varphi = \tilde{\varphi}_z$ and so

$$\tilde{\varepsilon}_z \geq \operatorname{dist}(1, \ell(\tilde{\varphi}_z)) = \sin|\tilde{\varphi}_z|.$$

Let $\mathsf{a}_\pm = e^{\pm i\tilde{\varphi}_z} \cos\tilde{\varphi}_z$ be the points on $\ell(\tilde{\varphi}_z)$ and $\ell(-\tilde{\varphi}_z)$ closest to 1. Thus $|a_\pm - 1| = \sin\tilde{\varphi}_z$. Now choose μ_z so that $|\mathsf{a}_1| = |\mathsf{a}_N|$, or in other words so that $(\mu_z + \lambda_N)/(\mu_z + \lambda_1) = |\kappa_z|$, which is equivalent to (2.8), and set $\tilde{\alpha}_z = \tilde{\rho}_z e^{-i\tilde{\varphi}_z}$, where $\tilde{\rho}_z = \cos\tilde{\varphi}_z/|\mathsf{a}_1|$. We find that $\tilde{\alpha}_z\,\mathsf{a}_1 = \mathsf{a}_+$ and $\tilde{\alpha}_z\,\mathsf{a}_N = \mathsf{a}_-$, so

$$\tilde{\varepsilon}_z \leq \max\{|1 - \mathsf{a}_+|, |1 - \mathsf{a}_-|\} = \sin|\tilde{\varphi}_z|$$

and hence this μ_z minimizes $\tilde{\varepsilon}_z(\mu_z)$, which completes the proof. □

Note that μ_z defined in (2.8) tends to $|z + \lambda_1| - \lambda_1$ as $\lambda_N \to \infty$. Also, since

$$|\varphi_{z,1} - \varphi_{z,N}| \leq 3\pi/4 \quad \text{for } z \in \Gamma,$$

we then have $\tilde{\varepsilon}_z \leq \sin(3\pi/8) \approx 0.9239$.

On the right-hand side of Table 1, we see the dramatic effect of the preconditioner $B_z = (\mu_z I + A)^{-1}$ on the error reduction factor, in the case of the model problem from Section 4, with $z = z_j$. Here we denote the optimal choice of the acceleration parameter by $\tilde{\alpha}_z = \tilde{\rho}_z e^{-i\tilde{\varphi}_z}$, computed as in the proof above. Notice that $\tilde{\varepsilon}_z = \|zI - \tilde{\alpha}_z A\|$ increases with j, whereas ε_z decreases.

Since computing the action of $(\mu_z I + A)^{-1}$ is expensive, we now want to consider a more general preconditioner B_z (still assumed to be positive definite and Hermitian). We then write G_z in (2.6) in the form

$$G_z = B_z A_z = \hat{z}\, B_z + B_z(\mu_z I + A) \quad \text{where } \hat{z} := z - \mu_z. \tag{2.9}$$

We take B_z^{-1} to be spectrally equivalent to $\mu_z I + A$, or

$$m_z(B_z^{-1}v, v) \leq ((\mu_z I + A)v, v) \leq M_z(B_z^{-1}v, v) \quad \text{for all } v \in V,$$

and define the associated z-dependent inner product and norm,

$$[v, w] := (B_z^{-1}v, w), \quad |[v]| := [v, v]^{1/2}.$$

The operator $H_z := B_z(\mu_z I + A)$ is then Hermitian with respect to $[\cdot, \cdot]$, and the largest and smallest choices of m_z and M_z are the minimum and maximum eigenvalues of H_z, respectively.

Sheen et al. [17] briefly discussed the preconditioning of (2.1) by using an operator B independent of z, corresponding to $\mu_z = 0$, and this turned out to be advantageous for small $|z|$. In our general case we have the following theorem, in which, for simplicity, we assume that $\operatorname{Im} z \geq 0$.

THEOREM 2.3. *Assume that* $\zeta := \arg \hat{z} \in [0, \pi)$, *define* $\hat{\varphi}_z$ *as the number in the interval* $J := (\zeta - \pi/2, \pi/2)$ *that minimizes*

$$\hat{\nu}_z(\varphi) := \frac{m_z \cos^2 \varphi \cos(\zeta - \varphi)}{M_z \cos(\zeta - \varphi) + \Lambda_z \cos \varphi} \quad \text{where } \Lambda_z := |\hat{z}|\, \|B_z\|,$$

and set $\hat{\rho}_z := \hat{\nu}_z(\hat{\varphi}_z)/(m_z \cos \hat{\varphi}_z)$. *Then*

$$|[I - \hat{\alpha}_z G_z]| \le \hat{\varepsilon}_z := (1 - \hat{\nu}_z(\hat{\varphi}_z))^{1/2} \quad \text{where } \hat{\alpha}_z := \hat{\rho}_z e^{-i\hat{\varphi}_z}. \tag{2.10}$$

If, in addition, there is a $\gamma_z \ge 0$ *such that*

$$\mathrm{Re}(\hat{z}\,[B_z v, B_z(\mu_z I + A)v]) \le -\gamma_z [B_z v, v] \quad \text{for all } v \in V, \tag{2.11}$$

define $\check{\varphi}_z$ *as the value in* J *minimizing*

$$\check{\nu}_z(\varphi) := \frac{m_z \cos^2 \varphi \cos(\zeta - \varphi)}{\max\{M_z \cos(\zeta - \varphi), \check{\Lambda}_z \cos \varphi\}} \quad \text{where } \check{\Lambda}_z := \Lambda_z - \frac{2\gamma_z}{|\hat{z}|},$$

and put $\check{\rho}_z := \check{\nu}_z(\check{\varphi}_z)/(m_z \cos \check{\varphi}_z)$. *We then have the sharper estimate*

$$|[I - \check{\alpha}_z G_z]| \le \check{\varepsilon}_z := (1 - \check{\nu}_z(\check{\varphi}_z))^{1/2} \quad \text{where } \check{\alpha}_z := \check{\rho}_z e^{-i\check{\varphi}_z}. \tag{2.12}$$

Since the operators involved are now not normal, the spectral characterization of the norms used above does not apply, and the rather lengthy and technical proof is based on energy methods. The argument begins by using standard estimates to show, for appropriate values of $\varphi := -\arg \alpha$ and $\rho := |\alpha|$, that

$$|[(I - \alpha G_z)v]|^2 = |[v]|^2 - 2\,\mathrm{Re}(\alpha [G_z v, v]) + |\alpha|^2 |[G_z v]|^2 \le (1 - \tilde{\nu}_z(\varphi))|[v]|^2.$$

Here we have used (2.9) to obtain

$$\mathrm{Re}(\alpha [G_z v, v]) = \mathrm{Re}(\alpha)[B_z(\mu_z I + A)v, v] + \mathrm{Re}(\alpha \hat{z})[B_z v, v],$$

where α is chosen so that $\mathrm{Re}(\alpha)$ and $\mathrm{Re}(\alpha \hat{z})$ are positive, and

$$|[G_z v]| \le |[B_z(\mu_z I + A)v]| + |\tilde{z}|\, |[B_z v]|.$$

If instead of the latter inequality one uses

$$|[G_z v]|^2 = |[B_z(\mu_z I + A)v]|^2 + |\hat{z}|^2 |[B_z v]|^2 + 2\,\mathrm{Re}(\hat{z}[B_z v, B_z(\mu_z I + A)v]),$$

then the estimate (2.11) makes it possible to replace $\hat{\nu}_z(\varphi)$ by the smaller quantity $\check{\nu}_z(\varphi)$; we refer to our aforementioned eprint [13] for details.

One example of a preconditioner $B_z = B_{z,k}$ satisfying the assumptions of Theorem 2.3 is obtained by applying k V-cycles of a symmetric algebraic multigrid algorithm [1] to approximately invert the operator $\mu_z I + A$. Since $B_{z,k} \to (\mu_z I + A)^{-1}$ as $k \to \infty$, it may be seen that (2.11) is valid for k sufficiently large. Table 2 shows results obtained with such a B_z first, on the left, using (2.10), and second, on the right, using (2.12), where, for each quadrature point z_j, the value of k shown is the smallest such that (2.11) holds. The choice of μ_z is the same as in Table 1.

TABLE 2. Error reduction by Theorem 2.3.

j	$\hat{\rho}_z$	$\hat{\varphi}_z$	$\hat{\varepsilon}_z$	k	γ_z	$\check{\rho}_z$	$\check{\varphi}_z$	$\check{\varepsilon}_z$
0	1.000	0.00	0.643	1	0.000	1.000	0.00	0.643
2	0.517	0.54	0.767	3	0.003	0.919	0.41	0.464
4	0.341	0.73	0.873	3	0.086	0.811	0.63	0.623
6	0.238	0.88	0.934	2	0.043	0.541	1.00	0.869
8	0.166	1.02	0.962	2	0.301	0.429	1.13	0.918
10	0.125	1.12	0.976	2	0.738	0.341	1.22	0.947
12	0.102	1.19	0.983	2	1.351	0.277	1.27	0.963
14	0.093	1.23	0.989	1	0.387	0.185	1.30	0.983
16	0.083	1.26	0.991	1	1.112	0.175	1.32	0.985
18	0.077	1.28	0.992	1	2.388	0.169	1.34	0.985
20	0.072	1.29	0.992	1	3.426	0.158	1.35	0.987

3. The conjugate gradient method

Once again, assume that A is a positive-definite Hermitian operator in a finite-dimensional complex inner product space V, and consider the equation

$$A_z w = g \quad \text{where } A_z := zI + A \text{ and } |\arg z| < \pi. \tag{3.1}$$

Given w_0, a preliminary guess for the solution w, we define the residual $r_0 := g - A_z w_0$ and the associated Krylov subspace of order $n \geq 1$,

$$V_n := \operatorname{span}\{r_0, A_z r_0, \ldots, A_z^{n-1} r_0\} = \operatorname{span}\{r_0, A r_0, \ldots, A^{n-1} r_0\},$$

with $V_0 := \{0\}$. Note that V_n depends on z through r_0. The exact solution of (3.1) satisfies

$$(A_z w, \varphi) = (g, \varphi) \quad \text{for all } \varphi \in V.$$

As in the classical CG method, we define the approximate solution $w_n = w_0 + v_n$, with $v_n \in V_n$, by Galerkin's method, or

$$(A_z w_n, \varphi) = (g, \varphi) \quad \text{for all } \varphi \in V_n, \tag{3.2}$$

and find that $v_n = w_n - w_0$ satisfies

$$(A_z v_n, \varphi) = (A_z(w_n - w_0), \varphi) = (g, \varphi) - (A_z w_0, \varphi) = (r_0, \varphi) \quad \text{for all } \varphi \in V_n.$$

The homogeneous equation, $(A_z v, \varphi) = 0$ for all $\varphi \in V_n$, admits only the trivial solution $v = 0$ in V_n, because taking $\varphi = v$ gives $z\|v\|^2 + (Av, v) = 0$. Hence there exists a unique solution of the finite-dimensional problem (3.2). The error $e_n := w_n - w$ satisfies

$$(A_z e_n, \varphi) = 0 \quad \text{for all } \varphi \in V_n. \tag{3.3}$$

To study the convergence of w_n, we introduce the norm

$$|||v|||^2 := |z|\,\|v\|^2 + (Av, v), \tag{3.4}$$

and begin with the following lemma.

LEMMA 3.1. *For all v, $w \in V$,*

$$|(A_z v, w)| \le |||v|||\,|||w||| \quad \textit{and} \quad |(A_z v, v)| \ge \cos(\phi/2)\,|||v|||^2 \quad \textit{where } \phi = \arg z.$$

PROOF. The first part follows by the Cauchy–Schwarz inequality, since

$$\begin{aligned} |(A_z v, w)| &\le |z|\,|(v, w)| + |(Av, w)| \le (|z|^{1/2}\|v\|)(|z|^{1/2}\|w\|) + (Av, v)^{1/2}(Aw, w)^{1/2} \\ &\le (|z|\,\|v\|^2 + (Av, v))^{1/2}(|z|\,\|w\|^2 + (Aw, w))^{1/2}. \end{aligned}$$

Setting $\beta := e^{-i\phi/2}$, the second part now results from

$$\begin{aligned} \mathrm{Re}(\beta(A_z v, v)) &= \mathrm{Re}(\beta z)\|v\|^2 + \mathrm{Re}\,\beta(Av, v) \\ &\ge |z|\cos(\phi/2)\|v\|^2 + \cos(\phi/2)\,(Av, v) = \cos(\phi/2)|||v|||^2. \end{aligned}$$

□

Using this lemma, we obtain the following quasioptimality result.

PROPOSITION 3.2. *With w and w_n the solutions of (3.1) and (3.2),*

$$|||w_n - w||| \le \sec(\phi/2) \inf_{v \in w_0 + V_n} |||v - w||| \quad \textit{with } \phi = \arg z.$$

PROOF. Lemma 3.1 and (3.3) show that, for any $v \in w_0 + V_n$,

$$\begin{aligned} \cos(\phi/2)|||w_n - w|||^2 &\le |(A_z(w_n - w), w_n - w)| = |(A_z(w_n - w), v - w)| \\ &\le |||w_n - w|||\,|||v - w|||, \end{aligned}$$

which implies the stated result. □

We now proceed to generalize the classical convergence analysis of the CG method by allowing for the complex shift in A_z. Let $\mathbb{P}_n$ denote the space of polynomials of degree at most n, with complex coefficients.

THEOREM 3.3. *Let w and w_n be the solutions of (3.1) and (3.2), and $e_n := w_n - w$. If $Q_n \in \mathbb{P}_n$ and $Q_n(0) = 1$ then*

$$|||e_n||| \le \sec(\phi/2) \max_{1 \le j \le N} |Q_n(z + \lambda_j)|\,|||e_0||| \quad \textit{with } \phi = \arg z.$$

PROOF. Let $v := w + Q_n(A_z)e_0$. Since

$$Q_n(\lambda) = 1 + \lambda P_{n-1}(\lambda) \quad \text{with } P_{n-1} \in \mathbb{P}_{n-1}$$

and

$$r_0 = g - A_z w_0 = -A_z(w_0 - w) = -A_z e_0,$$

we have $Q_n(A_z)e_0 = e_0 - P_{n-1}(A_z)r_0$. Hence $v = w_0 - P_{n-1}(A_z)r_0 \in w_0 + V_n$, and we conclude by Proposition 3.2 that

$$\cos(\phi/2)|||e_n||| \le |||v - w||| = |||Q_n(A_z)e_0|||.$$

Since A_z is a normal operator,

$$\|Q_n(A_z)e_0\| \le \max_{1\le j\le N} |Q_n(z+\lambda_j)|\,\|e_0\|.$$

Similarly,

$$(AQ_n(A_z)e_0, Q_n(A_z)e_0) \le \max_{1\le j\le N} |Q_n(z+\lambda_j)|^2\,(Ae_0, e_0),$$

and we conclude that

$$|||Q_n(A_z)e_0||| \le \max_{1\le j\le N} |Q_n(z+\lambda_j)|\,|||e_0|||,$$

which completes the proof. □

With the Chebyshev polynomial $T_n \in \mathbb{P}_n$ defined by $T_n(\cos\theta) = \cos(n\theta)$ for $\theta \in \mathbb{C}$, or equivalently (since $\cos(i\theta) = \cosh\theta$) by $T_n(\cosh\theta) = \cosh(n\theta)$, we show the following consequence of Theorem 3.3.

THEOREM 3.4. *With the above notation,*

$$|||e_n||| \le \sec(\phi/2)|T_n(-s_z)|^{-1}|||e_0||| \quad \textit{where } s_z := \frac{\lambda_N + \lambda_1 + 2z}{\lambda_N - \lambda_1}.$$

Here, with $\kappa_z = (\lambda_N + z)/(\lambda_1 + z)$,

$$T_n(-s_z) = \frac{\eta_z^n + \eta_z^{-n}}{2} \quad \textit{where } \eta_z := -\frac{\sqrt{\kappa_z} - 1}{\sqrt{\kappa_z} + 1} \textit{ and } |\arg\sqrt{\kappa_z}| < \pi/2.$$

Furthermore, $|\eta_z| \le 1 - c\lambda_N^{-1/2}$ *with* $c = c(z, \lambda_1) > 0$.

PROOF. The affine change of variables $s \mapsto \tau$ in the complex plane,

$$\tau = \frac{(1-s)(\lambda_1 + z) + (1+s)(\lambda_N + z)}{2},$$

takes the real interval $[-1, 1]$ onto the segment $[\lambda_1 + z, \lambda_N + z]$, parallel to the real axis. We note that $\tau = 0$ when $s = -s_z$, so that if we define

$$Q_n(\tau) := \frac{T_n(s)}{T_n(-s_z)} \quad \text{with } s = -\frac{\lambda_N + \lambda_1 + 2(z-\tau)}{\lambda_N - \lambda_1},$$

then $Q_n(\tau) \in \mathbb{P}_n$ and $Q_n(0) = 1$. Thus

$$\max_{\lambda\in[\lambda_1,\lambda_N]} |Q_n(\lambda + z)| = \max_{\tau\in[\lambda_1+z,\lambda_N+z]} |Q_n(\tau)| = \max_{-1\le s\le 1} \frac{|T_n(s)|}{|T_n(-s_z)|} = \frac{1}{|T_n(-s_z)|},$$

and hence the first statement of the theorem follows by Theorem 3.3.

Defining θ by $\cosh\theta \equiv (e^{\theta}+e^{-\theta})/2 = -s_z$ and letting $\eta_z = e^{\theta}$,

$$T_n(-s_z) = T_n(\cosh\theta) = \cosh(n\theta) = \frac{\eta_z^n + \eta_z^{-n}}{2}.$$

Here η_z satisfies the quadratic equation $\eta_z + \eta_z^{-1} = -2s_z$, with roots

$$\eta_{z,\pm} = -s_z \pm \sqrt{s_z^2-1} = -\frac{(\sqrt{s_z+1} \mp \sqrt{s_z-1})^2}{2}.$$

Let $\eta_z = \eta_{z,+}$. Since $(s_z+1)/(s_z-1) = (\lambda_N+z)/(\lambda_1+z) = \kappa_z$, we find that

$$-\eta_z = \frac{\sqrt{s_z+1}-\sqrt{s_z-1}}{\sqrt{s_z+1}+\sqrt{s_z-1}} = \frac{\sqrt{\kappa_z}-1}{\sqrt{\kappa_z}+1} = \frac{1-\kappa_z^{-1/2}}{1+\kappa_z^{-1/2}},$$

and since $|\arg\sqrt{\kappa_z}| < \pi/2$ we have $\operatorname{Re}\sqrt{\kappa_z} > 0$. Hence $|\eta_z| < 1$.

For the final conclusion, we note that $\kappa_z^{-1} = (z+\lambda_1)\lambda_N^{-1} + O(\lambda_N^{-2})$ and thus, because $|\arg(z+\lambda_1)| < |\phi| < \pi$, we have, for λ_N large,

$$\operatorname{Re}\kappa_z^{-1/2} \ge \cos(\phi/2)\sqrt{|z+\lambda_1|}\lambda_N^{-1/2} + O(\lambda_N^{-1}) \ge c(z,\lambda_1)\lambda_N^{-1/2},$$

with $c(z,\lambda_1) > 0$. Since $-\eta_z = 1 - 2\kappa_z^{-1/2} + O(\lambda_N^{-1})$, it follows that

$$|\eta_z|^2 = 1 - 4\operatorname{Re}\kappa_z^{-1/2} + O(\lambda_N^{-1}) \le 1 - 2c(z,\lambda_1)\lambda_N^{-1/2} \quad \text{for } \lambda_N \text{ large},$$

which completes the proof. □

Since $|\eta_z| < 1$, it follows that $|T_n(-s_z)|^{-1} \approx 2|\eta_z|^n$, and so Theorem 3.4 shows linear convergence with approximately this rate. When $A = L_h$ and $\lambda_N = O(h^{-2})$, the error bound is thus of order $(1-ch)^n$, with $c > 0$. The values of $|\eta_z|$ shown in Table 3 refer to the model problem from Section 4, for which $\lambda_1 \approx 1$ and $\lambda_N \approx 4000$. Comparing the $|\eta_z|$ with the corresponding values of ε_z in Table 1 confirms the superiority of the CG method over the Richardson iteration (without preconditioning).

Freund [4, Theorem 4(c)] proved a similar error bound in the norm $\|v\|_T := \sqrt{(Tv,v)}$, where $T = A - xI$, under the restriction $x > -\lambda_1(A)$ so that T is positive definite.

We now seek to precondition the CG method applied to (3.1), and consider first the special preconditioner $B_z = (\mu_z I + A)^{-1}$, with $\mu_z > -\lambda_1$. We multiply (3.1) by $(z-\mu_z)^{-1}B_z$ to write the equation in the form

$$\tilde{z}w + B_z w = \tilde{z}\, B_z g \quad \text{where } \tilde{z} := (z-\mu_z)^{-1}, \tag{3.5}$$

in which thus $\tilde{z}$ and B_z play the roles previously taken by z and A. Note that the angle $\tilde{\phi} := \arg\tilde{z} = -\arg(z-\mu_z)$ satisfies $|\tilde{\phi}| < \pi$. In particular, the Krylov subspaces are now

$$V_n = \operatorname{span}\{r_0, B_z r_0, \dots, B_z^{n-1} r_0\} \quad \text{with } r_0 = \tilde{z}\, B_z g - (\tilde{z}I + B_z)w_0, \tag{3.6}$$

TABLE 3. Error reduction by CG iteration.

j	x_j	y_j	$\lvert\eta_z\rvert$	$\lvert\tilde{\eta}_z\rvert$	μ_z	$\lvert\tilde{\eta}_z\rvert$ for $\mu_z = 0$
0	0.00	0.00	0.9687	0.0000	0.000	0.0000
2	−0.05	0.30	0.9690	0.0762	0.002	0.0762
4	−0.18	0.64	0.9699	0.1650	0.031	0.1652
6	−0.43	1.02	0.9708	0.2698	0.165	0.2724
8	−0.81	1.51	0.9711	0.3749	0.507	0.3880
10	−1.35	2.12	0.9703	0.4605	1.138	0.4948
12	−2.10	2.93	0.9686	0.5221	2.119	0.5839
14	−3.13	4.01	0.9659	0.5646	3.530	0.6553
16	−4.54	5.45	0.9622	0.5939	5.492	0.7121
18	−6.45	7.38	0.9577	0.6143	8.183	0.7577
20	−9.02	9.97	0.9523	0.6287	11.850	0.7946

and the iterates $w_n = w_0 + v_n$, with $v_n \in V_n$, are defined by

$$((\tilde{z}I + B_z)w_n, \varphi) = (\tilde{z}\, B_z g, \varphi) \quad \text{for all } \varphi \in V_n. \tag{3.7}$$

The earlier analysis remains valid, with s_z now replaced by

$$\tilde{s}_z := \frac{\tilde{\lambda}_N + \tilde{\lambda}_1 + 2\tilde{z}}{\tilde{\lambda}_N - \tilde{\lambda}_1} \quad \text{where } \tilde{\lambda}_j := (\mu_z + \lambda_{N+1-j})^{-1} \text{ for } j = 1, N,$$

and correspondingly for η_z. Since $\tilde{\lambda}_N$ is bounded, Theorem 3.4 then shows that the error reduction factor is bounded away from 1, independently of λ_N. We also show that the optimal choice of μ_z is the same as for Richardson iteration in Theorem 2.2. Recall that $G_z(\lambda, \mu_z) = (z + \lambda)/(\mu_z + \lambda)$.

THEOREM 3.5. *For the CG method* (3.7) *applied to equation* (3.5)*, and for the norm* $|||v|||^2 = |\tilde{z}|\,\|v\|^2 + (B_z v, v)$*,*

$$|||e_n||| \le \sec(\tilde{\phi}/2)\,|T_n(-\tilde{s}_z)|^{-1}|||e_0||| \quad \textit{with } T_n(-\tilde{s}_z) = \frac{\tilde{\eta}_z^n + \tilde{\eta}_z^{-n}}{2}.$$

Here

$$\tilde{\eta}_z := -\frac{\sqrt{\tilde{\kappa}_z} - 1}{\sqrt{\tilde{\kappa}_z} + 1} \quad \textit{where } \tilde{\kappa}_z := \frac{\tilde{\lambda}_N + \tilde{z}}{\tilde{\lambda}_1 + \tilde{z}} = \frac{G_z(\lambda_1, \mu_z)}{G_z(\lambda_N, \mu_z)}, \tag{3.8}$$

and hence $|\tilde{\eta}_z| \le c(z, \lambda_1, \mu_z) < 1$*. If* $|z + \lambda_N| > |z + \lambda_1|$ *then the smallest value of* $|\tilde{\eta}_z|$ *is attained for* μ_z *as in* (2.8)*, and for this choice*

$$|\tilde{\eta}_z| = \tan(|\tilde{\varphi}_z|/2) \quad \textit{where } \tilde{\varphi}_z := \frac{\varphi_{z,1} - \varphi_{z,N}}{2} \textit{ and } \varphi_{z,j} := \arg(z + \lambda_j).$$

PROOF. With $\mathsf{a}_j = G_z(\lambda_j, \mu_z)$ we have $\tilde{\kappa}_z = \mathsf{a}_1/\mathsf{a}_N$, and so it follows from (3.8) that

$$\tilde{\eta}_z = -\frac{\sqrt{\mathsf{a}_1/\mathsf{a}_N} - 1}{\sqrt{\mathsf{a}_1/\mathsf{a}_N} + 1}.$$

Since $\arg \mathsf{a}_j = \varphi_{z,j}$, we may write $\sqrt{\mathsf{a}_1/\mathsf{a}_N} = \tau e^{i\tilde{\varphi}_z}$, with $\tau > 0$ and $|\tilde{\varphi}_z| < \pi/2$, and find that

$$|\tilde{\eta}_z|^2 = \frac{(\tau \cos \tilde{\varphi}_z - 1)^2 + \tau^2 \sin^2 \tilde{\varphi}_z}{(\tau \cos \tilde{\varphi}_z + 1)^2 + \tau^2 \sin^2 \tilde{\varphi}_z} < 1. \tag{3.9}$$

The ratio is minimized at $\tau = 1$, that is, when $|\mathsf{a}_1| = |\mathsf{a}_N|$, or when (2.8) holds. For $\tau = 1$, (3.9) becomes $|\tilde{\eta}_z|^2 = \tan^2(\tilde{\varphi}_z/2)$, and since $|\tilde{\varphi}_z| < \pi/2$, this completes the proof. □

Note that since $|\tilde{\varphi}_z| < \pi/2$,

$$|\tilde{\eta}_z| = \tan(|\tilde{\varphi}_z|/2) < \sin |\tilde{\varphi}_z| = \tilde{\varepsilon}_z,$$

that is, the error reduction factor for CG is smaller than that for Richardson iteration. If $z \in \Gamma$ then $|\tilde{\varphi}_z| < 3\pi/8$, implying that $|\tilde{\eta}_z| < \tan(3\pi/16) \approx 0.6682$. In Table 3 we give some values of $|\tilde{\eta}_z|$, first for the optimal μ_z determined by Theorem 3.5, and then (in the final column) for $\mu_z = 0$. As for Richardson iteration, the preconditioning becomes less effective with increasing j.

We now consider the preconditioned form (2.5) of (1.1), where B_z is a more general Hermitian positive-definite operator than $(\mu_z I + A)^{-1}$, or

$$G_z w = \tilde{g}_z := B_z g \quad \text{where } G_z = B_z A_z = zB_z + B_z A. \tag{3.10}$$

Note that both B_z and $B_z A$ are Hermitian and positive definite with respect to the inner product $[v, w] := (B_z^{-1} v, w)$. We now define the Krylov subspaces by

$$\tilde{V}_n := \operatorname{span}\{\tilde{r}_0, G_z \tilde{r}_0, \ldots, G_z^{n-1} \tilde{r}_0\} \quad \text{where } \tilde{r}_0 := \tilde{g}_z - G_z w_0 = B_z r_0,$$

and the CG iterates $w_n = w_0 + v_n$ with $v_n \in \tilde{V}_n$ by

$$(A_z w_n, \varphi) = (g, \varphi) \quad \text{for all } \varphi \in \tilde{V}_n,$$

or equivalently $[G_z w_n, \varphi] = [\tilde{g}_z, \varphi]$ for all $\varphi \in \tilde{V}_n$.

The existence and uniqueness of w_n follow as before, and the inequalities in Lemma 3.1 remain valid, with $|||\cdot|||$ defined in (3.4). The error $e_n = w_n - w$ again satisfies an orthogonality property,

$$(A_z e_n, \varphi) = 0 \quad \text{for all } \varphi \in \tilde{V}_n,$$

and the proof of the quasioptimality result of Proposition 3.2 carries over verbatim. However, the proof of the error bound of Theorem 3.3 does not remain valid, in general, owing to the presence of the operator B_z in the definition of the Krylov subspaces $\tilde{V}_n$.

We discuss briefly the implementation of the CG method, referring again to our aforementioned eprint [13] for further details. Starting with the basic method (3.2), one shows that, assuming $r_0 \neq 0$, there is an $N^* \leq N = \dim(V)$ such that $r_n \neq 0$ for $0 \leq n < N^*$, but $r_n = 0$ (and thus $w_n = w$) for $n \geq N^*$. The residuals satisfy

$$V_n = \operatorname{span}\{r_0, r_1, \ldots, r_{n-1}\} \quad \text{and} \quad (r_n, \varphi) = 0 \quad \text{for all } \varphi \in V_n.$$

In order to obtain a recursive algorithm for computing the CG iterates w_n, one seeks a second sequence of vectors p_n, such that, if $1 \leq n < N^*$,

$$V_n = \operatorname{span}\{p_0, p_1, \ldots, p_{n-1}\} \quad \text{and} \quad (A_z p_n, \varphi) = 0 \quad \text{for all } \varphi \in V_n. \tag{3.11}$$

Putting $p_0 := r_0$, one may show that

$$p_{n+1} = r_{n+1} + \beta_n p_n \quad \text{where } \beta_n := -\frac{(r_{n+1}, A_z p_n)}{(A_z p_n, p_n)},$$

making it possible to compute the w_n recursively by

$$w_{n+1} = w_n + \alpha_n p_n \quad \text{where } \alpha_n := \frac{\|r_n\|^2}{(A_z p_n, p_n)} \text{ for } 0 \leq n < N^*.$$

This implies that

$$r_{n+1} = r_n - A_z(w_{n+1} - w_n) = r_n - \alpha_n A_z p_n,$$

so that the r_n may also be computed recursively. For real $z > -\lambda_1(A)$ the scalar α_n is real, so

$$-\alpha_n(r_{n+1}, A_z p_n) = (r_{n+1}, r_n - \alpha_n A_z p_n) = \|r_{n+1}\|^2 \quad \text{and} \quad \beta_n = \|r_{n+1}\|^2 / \|r_n\|^2,$$

which is the formula used in the classical CG method.

The preceding analysis remains valid if we use the special preconditioner $B_z = (\mu_z I + A)^{-1}$, reformulating (3.1) as (3.5), with iterates defined by (3.6) and (3.7). We now have $r_n = \tilde{z} B_z g - (\tilde{z}\, I + B_z) w_n$, and in the computation of α_n and β_n, the inner product $(A_z v, w)$ is replaced by $((\tilde{z}\, I + B_z)v, w)$.

Finally, consider a general preconditioned equation of the form (3.10). The preconditioned residuals $\tilde{r}_n := \tilde{g}_z - G_z w_n = B_z r_n$ satisfy

$$\tilde{V}_n = \operatorname{span}\{\tilde{r}_0, \tilde{r}_1, \ldots, \tilde{r}_{n-1}\} \quad \text{and} \quad [\tilde{r}_n, \varphi] = (r_n, \varphi) = 0 \quad \text{for all } \varphi \in V_n,$$

and for $1 \leq n < N^*$ we seek p_n satisfying (3.11) with V_n replaced by $\tilde{V}_n$. With $p_0 := \tilde{r}_0$, we find that

$$p_{n+1} = \tilde{r}_{n+1} + \sum_{k=0}^{n} \beta_{nk} p_k,$$

where the β_{nk} satisfy the nonsingular lower-triangular linear system

$$\sum_{k=0}^{j} (A_z p_k, p_j)\beta_{nk} = -(A_z \tilde{r}_{n+1}, p_j) \quad \text{for } 0 \leq j \leq n < N^*.$$

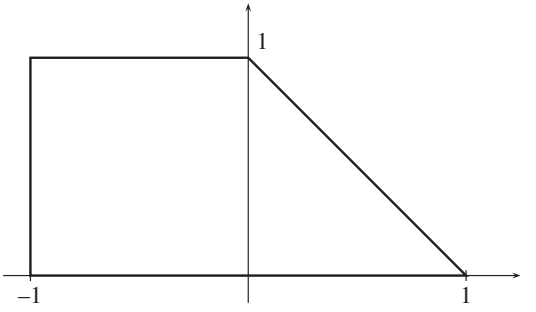

FIGURE 1. The domain Ω.

Unfortunately, in contrast to the situation earlier, $\beta_{nk} \neq 0$ is possible for $k < n-1$, and consequently all the p_j must be stored.

Given w_n, r_n and p_n, we can compute

$$w_{n+1} = w_n + \alpha_n p_n \quad \text{where } \alpha_n := \frac{|[\tilde{r}_n]|^2}{(A_z p_n, p_n)} = \frac{(r_n, \tilde{r}_n)}{(A_z p_n, p_n)}.$$

The residuals again satisfy $r_{n+1} = r_n - \alpha_n A_z p_n$, implying that the preconditioned residuals satisfy $\tilde{r}_{n+1} = \tilde{r}_n - \alpha_n G_z p_n$. Each iteration is now more expensive than in the basic CG method, both in CPU time and memory requirements, so one may want to restart the iteration every m steps for some moderate choice of m. An investigation of such a restarted iteration is beyond the scope of the present work, however.

4. A model problem

We now describe a concrete initial boundary-value problem (1.2), used already in the numerical examples of Sections 2 and 3, and present some further illustrations of our results.

For the domain Ω we take the trapezium with vertices $(1, 0)$, $(0, 1)$, $(-1, 1)$ and $(-1, 0)$, shown in Figure 1. The minimum eigenvalue of $-\nabla^2$ on Ω is close to 15, so we choose the diffusivity $a = 1/15$ to give a time scale of order 1 for (1.2). We choose the data u_0 and f so that the exact solution is

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

and use continuous, piecewise linear finite elements on a quasiuniform, unstructured triangulation $\mathcal{T}_h$ of Ω, generated by the program Gmsh [6]. The dimension of the finite-element space V_h is $N = 2663$, and the maximum element diameter is $h = 0.035$. The extremal eigenvalues of the operator $A = L_h$ are $\lambda_1 \approx 1.014$ and $\lambda_N \approx 4006$.

In our numerical results, we employ the discrete L_2 norm $\|v\|_h = \|I_h v\|$, where I_h is the nodal interpolation operator for the finite-element space V_h. Table 4 shows this discrete L_2-norm of the error in $U_{q,h}(t)$ at four values of t, for three choices of q, as well

TABLE 4. Discretization error $\|U_{q,h}(t) - u(t)\|_h$.

t	$q = 10$	$q = 20$	$q = 30$	$\|u(t)\|_h$
0.25	1.3436×10^{-2}	4.3778×10^{-4}	4.1747×10^{-4}	0.4452
0.50	6.1232×10^{-4}	1.6260×10^{-4}	1.7541×10^{-4}	0.4623
1.00	2.2024×10^{-4}	2.1088×10^{-4}	2.1114×10^{-4}	0.4206
2.00	1.9403×10^{-4}	1.9411×10^{-4}	1.9411×10^{-4}	0.2579

TABLE 5. Iteration counts for $t = 1$ at different quadrature points. In the column headings, B indicates the basic CG method, INV the special preconditioner $B_z = (\mu_z I + A)^{-1}$, and AMG(k) the algebraic multigrid preconditioner [1] with k V-cycles, used previously in Table 2. The heading IC refers to preconditioning using an incomplete Cholesky factorization [8] of $\mu_z I + A$.

	Richardson			CG				
j	INV	AMG(3)	B	INV	IC	AMG(1)	$\|w(z_j)\|_h$	ε_j
0	1	5	250	1	52	7	1.14×10^{0}	3.18×10^{-6}
2	7	9	227	5	48	7	1.13×10^{0}	3.06×10^{-6}
4	10	15	235	6	50	8	1.03×10^{0}	2.84×10^{-6}
6	15	25	242	7	51	9	7.67×10^{-1}	2.78×10^{-6}
8	24	42	234	8	50	10	4.39×10^{-1}	3.03×10^{-6}
10	39	56	219	9	46	11	2.21×10^{-1}	3.86×10^{-6}
12	49	57	184	10	40	11	1.19×10^{-1}	6.08×10^{-6}
14	48	45	149	9	32	10	7.41×10^{-2}	1.27×10^{-5}
16	44	37	98	8	22	9	5.11×10^{-2}	3.83×10^{-5}
18	32	26	34	5	11	5	3.69×10^{-2}	1.91×10^{-4}
20	8	7	10	2	3	2	2.71×10^{-2}	1.87×10^{-3}

as the norm of the solution itself. We see that once q is about 20, the $O(h^2)$ error from the spatial discretization dominates the $O(e^{-cq/\log q})$ error from the time discretization; compare with (1.8).

In Table 5 we show iteration counts for $t = 1$ at alternate quadrature points for several versions of the Richardson and CG iterations. As the acceleration parameter for the Richardson iterations, we take the optimal $\alpha = \tilde{\alpha}_z$ from Theorem 2.2 in the case of the INV preconditioner, and $\alpha = \check{\alpha}_z$ from Theorem 2.3 for AMG(3). In all cases, we use the value of the parameter μ_z given in Theorems 2.2 and 3.5, which is optimal for the INV preconditioner. Except for $j = 0$, the AMG(1) preconditioner for CG is almost as effective as INV, requiring only 11 iterations in the worst case. One could also reduce the set-up cost for AMG by using the same μ_z for several nearby quadrature points, but we have not investigated the trade-off between the cost saving and a possibly slower convergence.

Let $e_n = \tilde{w}_h(z_j) - w_h(z_j)$ denote the solver error after n iterations. We compute $w_h(z_j)$ using a sparse direct solver [3], and stop the iterations once

$$\|e_n\| \le \epsilon_j \quad \text{where } \epsilon_j := \delta \times \frac{2\pi e^{|x_j|t}}{(2q+1)k|z_j'|} \text{ for } \delta = 10^{-5}, \tag{4.1}$$

where $x_j = \operatorname{Re} z_j < 0$. This way, the estimate (1.9) ensures that the additional error in $U_{q,h}(t)$ due to the iterative solver is less than δ. For $j = 0$, we start each iteration with the zero vector, but for $j \ge 1$ we use the final iterate at z_{j-1} as the starting iterate at z_j. The remaining columns of the table show the values of $\|w_h(z_j)\|$ and ϵ_j. The former decrease whereas the latter increase with increasing j. In particular, the growth in ϵ_j more than compensates for the deterioration in the error reduction factors of the iterative solvers, seen in Tables 1 and 3.

References

[1] W. N. Bell, L. N. Olson and J. B. Schroder, "PyAMG: algebraic multigrid solvers in Python v2.0", http://www.pyamg.org.

[2] M. Benzi and D. Bertaccini, "Block preconditioning of real-valued iterative algorithms for complex linear systems", *IMA J. Numer. Anal.* **28** (2008) 598–618; doi:10.1093/imanum/drm039.

[3] T. A. Davis, "Algorithm 832: UMFPACK V4.3—an unsymmetric-pattern multifrontal method", *ACM Trans. Math. Software* **30** (2004) 196–199; doi:10.1145/992200.992206.

[4] R. Freund, "On conjugate gradient type methods and polynomial preconditioners for a class of complex non-Hermitian matrices", *Numer. Math.* **57** (1990) 285–312; doi:10.1007/BF01386412.

[5] I. P. Gavrilyuk and V. L. Makarov, "Exponentially convergent algorithms for the operator exponential with applications to inhomogeneous problems in Banach spaces", *SIAM J. Numer. Anal.* **43** (2005) 2144–2171; doi:10.1137/040611045.

[6] C. Geuzaine and J.-F. Remacle, "Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities", http://www.geuz.org/gmsh.

[7] K. J. in 't Hout and J. A. C. Weideman, "A contour integral method for the Black–Scholes and Heston equations", *SIAM J. Sci. Comput.* **33** (2011) 763–785; doi:10.1137/090776081.

[8] M. T. Jones and P. E. Plassmann, "Algorithm 740: Fortran subroutines to compute improved incomplete Cholesky factorizations", *ACM Trans. Math. Software* **21** (1995) 18–19; doi:10.1145/200979.200986.

[9] W. McLean, I. H. Sloan and V. Thomée, "Time discretization via Laplace tranformation of an integro-differential equation of parabolic type", *Numer. Math.* **102** (2006) 497–522; doi:10.1007/s00211-005-0657-7.

[10] W. McLean and V. Thomée, "Time discretization of an evolution equation via Laplace transforms", *IMA J. Numer. Anal.* **24** (2004) 439–463; doi:10.1093/imanum/24.3.439.

[11] W. McLean and V. Thomée, "Maximum-norm error analysis of a numerical solution via Laplace transformation and quadrature of a fractional order evolution equation", *IMA J. Numer. Anal.* **30** (2010) 208–230; doi:10.1093/imanum/drp004.

[12] W. McLean and V. Thomée, "Numerical solution via Laplace transforms of a fractional-order evolution equation", *J. Integral Equations Appl.* **22** (2010) 57–94; doi:10.1216/JIE-2010-22-1-57.

[13] W. McLean and V. Thomée, "Iterative methods for shifted positive definite linear systems and time discretization of the heat equation", Preprint, 2011, http://arxiv.org/abs/1111.5105.

[14] K. Meerbergen, "The solution of parametrized symmetric linear systems", *SIAM J. Matrix Anal. Appl.* **24** (2003) 1038–1059; doi:10.1137/S0895479800380386.

[15] C. C. Paige and M. A. Saunders, "Solution of sparse indefinite systems of linear equations", *SIAM J. Numer. Anal.* **12** (1975) 617–629; doi:10.1137/0712047.

[16] D. Sheen, I. H. Sloan and V. Thomée, "A parallel method for time-discretization of parabolic problems based on contour integral representation and quadrature", *Math. Comp.* **69** (2000) 177–195; doi:10.1090/S0025-5718-99-01098-4.

[17] D. Sheen, I. H. Sloan and V. Thomée, "A parallel method for time-discretization of parabolic equations based on Laplace transformation and quadrature", *IMA J. Numer. Anal.* **23** (2003) 269–299; doi:10.1093/imanum/23.2.269.

[18] V. Simoncini, "Restarted full orthogonalization method for shifted linear systems", *BIT Numer. Math.* **43** (2003) 459–466; doi:10.1023/A:1026000105893.

[19] V. Simoncini and D. B. Szyld, "Recent computational developments in Krylov subspace methods for linear systems", *Numer. Linear Algebra Appl.* **14** (2007) 1–59; doi:10.1002/nla.499.