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\textsuperscript{\textcopyright 1} and VIDAR THOMÉE\textsuperscript{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.


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

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

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

\textsuperscript{1}School of Mathematics and Statistics, The University of New South Wales, Sydney 2052, Australia;
e-mail: w.mclean@unsw.edu.au.
\textsuperscript{2}Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, S-412 96
Gothenburg, Sweden; e-mail: thomee@chalmers.se.
\textcopyright Australian Mathematical Society 2012. Serial-fee code 1446-1811/2012 $16.00
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

\[
\partial_t u + Lu = f \text{ in } \Omega \quad \text{with } u = 0 \text{ on } \partial \Omega \text{ for } t > 0,
\]

\[
u = u_0 \text{ in } \Omega \quad \text{when } t = 0,
\]

where \( \partial_t u = \partial u / \partial t \). For simplicity, we assume that the diffusivity \( a \) is a positive constant, and that \( \Omega \) 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 \( H^1_0(\Omega) \) and satisfying

\[
(\partial_t u, v) + a(\nabla u, \nabla v) = (f, v) \quad \text{for all } v \in H^1_0(\Omega) \text{ and } t > 0,
\]

with \( u(0) = u_0 \), where \( (v, w) = \int_\Omega v w \, dx \) is the inner product in \( L^2(\Omega) \). To discretize in space, let \( \{V_h\} \subset H^1_0(\Omega) \) be a family of continuous, piecewise linear finite-element spaces, based on a family of regular partitions \( T_h = \{\tau\} \) of \( \Omega \). The standard Galerkin spatially semidiscrete solution \( u_h : [0, \infty) \rightarrow V_h \) 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 \( L_h : V_h \rightarrow V_h \) 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

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

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

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

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 \( \Gamma \) 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

\[(zI + L_h)w_h(z) = P_h g(z) \quad \text{for } z \in \Gamma,\]

(1.4)
or, in weak form,

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

(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 \(\Gamma\) 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

\[U_{q,h}(t) := \frac{k}{2\pi i} \sum_{j=-q}^{q} e^{z(\xi_j)t} w_h(z_j)z'_j,\]

(1.6)

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

\[\xi_j := jk, \quad z_j := z(\xi_j), \quad z'_j := z'(\xi_j) \quad \text{for } |j| \leq q.\]

(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 \(zMw + Sw = g\), where \(M\) and \(S\) are the mass and stiffness matrices, \(w\) is the vector of nodal values of \(w_h(z)\), and \(g\) is the load vector for the source term \(g(z)\). Equivalently, \(zw + M^{-1}Sw = M^{-1}g\), which again has the form (1.1), and the operator \(A = M^{-1}S\) is Hermitian and positive definite with respect to the inner product induced by \(M\), 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 \( \tilde{U}_{q,h} \), 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:

\[
\| \tilde{U}_{q,h}(t) - u(t) \| \leq \| \tilde{U}_{q,h}(t) - U_{q,h}(t) \| + \| U_{q,h}(t) - u_h(t) \| + \| u_h(t) - u(t) \|.
\]

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

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

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

\[
\mathcal{E}(t) := \| \tilde{U}_{q,h}(t) - U_{q,h}(t) \| \leq \frac{k}{2\pi} \sum_{j=1}^{q} \varepsilon_j e^{\Re(z_j)u|z'_j|}. \tag{1.9}
\]

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 \( \varepsilon_j \) in such a way that \( \mathcal{E}(t) \leq C(h^2 + e^{-cq/\log q}) \) for some constant \( C \). The presence of the factor \( e^{\Re(z_j)u|z'_j|} \) allows \( \varepsilon_j \) to increase with \( |j| \); see (4.1) below and remember that \( \Re(z_j) < 0 \).

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 $\alpha$ in Section 2, extending and improving upon the results of Sheen et al. [17]. Let $\lambda_j$ denote the $j$th eigenvalue of $A$, labelled so that $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$. From a knowledge of the extremal eigenvalues $\lambda_1$ and $\lambda_N$, we can determine the optimal $\alpha$, 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 $\lambda_1$ 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 $\mu_z$, and find that $\mu_z \approx |z + \lambda_1| - \lambda_1$ for large $\lambda_N$. 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

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

(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 $|\eta_z| \leq 1 - ch$, 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 \( \mu_z \) 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 \([v, w] = (B_z^{-1/2} v, w)\). 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 \( n \)th 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 \( \lambda_1 \) and \( \lambda_N \), 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 \( A_z w = g \):

\[
\begin{align*}
  w^{n+1} &= (I - \alpha A_z)w^n + \alpha g &\text{for } n \geq 0 \text{ with } w^0 \text{ given.} \\
\end{align*}
\]

(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 \( V \),

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

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

\[
\| I - \alpha A \| = \frac{\kappa(A) - 1}{\kappa(A) + 1} \quad \text{where } \kappa(A) := \frac{\lambda_N}{\lambda_1}
\] (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

\[
\| I - \alpha A \| \leq 1 - ch^2 \quad \text{with } c > 0.
\] (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

\[
\varepsilon_z := \min_{\alpha} \| I - \alpha A_z \| \leq 1 - c\lambda_N^{-1} \quad \text{with } c = c(z, \lambda_1) > 0.
\]

**Proof.** We begin by demonstrating that, for \(\lambda_N \leq \Lambda < \infty\), there exists an \(\alpha\) such that

\[
\| I - \alpha A_z \| \leq c(z, \lambda_1, \Lambda) < 1.
\]

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 \(\lambda_N\) 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

\[
\varepsilon_z = \max_{\alpha} \| 1 - \alpha(z + \lambda_1) \|, \quad \| 1 - \alpha(z + \lambda_N) \|,
\]

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 \(y \neq 0\) 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 \(y < 0\),
the Richardson iteration algorithm becomes
\[ w_{n+1} = (I - \alpha G_z)w_n + \alpha \tilde{g}_z. \]  
(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

$$||I - \alpha G_z(A, \mu_z)|| = \max_{\lambda \leq \mu_z} |1 - \alpha G_z(\lambda, \mu_z)|$$

where $G_z(\lambda, \mu_z) = \frac{z + \lambda}{\mu_z + \lambda}$, and we want to choose $\alpha$ and $\mu_z$ so that this quantity is as small as possible. In practice we are interested only in $z = z_j$ with $\text{Re } z_j \geq \text{Re } z_q \approx -q/2$ 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 $\lambda_N$,

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

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

$$\mu_z = -\lambda_1 + \frac{\lambda_N - \lambda_1}{|\kappa_z| - 1} > -\lambda_1 \quad \text{where } \kappa_z := \frac{z + \lambda_N}{z + \lambda_1}.$$  \hspace{1cm} (2.8)

For this choice of $\mu_z$,

$$\tilde{\varepsilon}_z = \sin |\tilde{\varphi}_z| \quad \text{where } \tilde{\varphi}_z = \frac{\varphi_{z,1} - \varphi_{z,N}}{2} \quad \text{and } \varphi_{z,j} = \arg (z + \lambda_j).$$

**Proof.** For $\mu_z$ given, let $a_j = G_z(\lambda_j, \mu_z)$. We note that, since $G_z(\infty, \mu_z) = 1$ and $(\mu_z + \lambda)^{-1}$ is positive and decreasing in $\lambda$,

$$G_z(\lambda, \mu_z) = 1 + (\mu_z + \lambda)^{-1}(\bar{z} - \mu_z) \in [a_1, a_N] \subseteq [a_1, 1] \quad \text{for } \lambda \in [\lambda_1, \lambda_N],$$

and

$$F(\alpha) := ||I - \alpha G_z(A, \mu_z)|| = \max(||1 - \alpha a_1||, ||1 - \alpha a_N||).$$

We first show that $\alpha$ may be chosen so that the line segment $\alpha[a_1, 1]$ is inside the disk $|z - 1| < 1$. For this, we first determine $\arg \alpha$ by rotating the line segment $[a_1, 1]$ 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 $\alpha$, $F(\alpha) < 1$ uniformly in $\lambda_N$, and thus (2.7) holds uniformly in $\lambda_N$.

For $\mu_z$ given, the optimal $\alpha$ satisfies $|1 - \alpha a_1| = |1 - \alpha a_N|$, and thus also $|1/\alpha - a_1| = |1/\alpha - a_N|$. Therefore, $1/\alpha$ has to be chosen on the line in $\mathbb{C}$ through the midpoint $c = (a_1 + a_N)/2$ which is perpendicular to $a_N - a_1$, or has the direction of $d = i(a_N - a_1)$, so that $\alpha = \alpha(s) = 1/(c + sd)$. Then for $\alpha$ to be optimal, $s \in \mathbb{R}$ has to minimize the rational function $F(\alpha(s))^2$. 


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

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

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 $\varphi_{z,1} - \varphi = \tilde{\varphi}$ and so

$$\tilde{\varphi} = \text{dist}(1, \ell(\varphi_1)) = \sin |\varphi_1|.$$  

Let $a_\pm = e^{\pm i\tilde{\varphi}} \cos \tilde{\varphi}$ be the points on $\ell(\varphi_1)$ and $\ell(-\varphi_1)$ closest to 1. Thus $|a_\pm - 1| = \sin \tilde{\varphi}$. Now choose $\mu_\varphi$ so that $|a_\pm| = |a_\varphi|$, or in other words so that $(\mu_\varphi + \lambda_N)/(\mu_\varphi + \lambda_1) = |\kappa_z|$, which is equivalent to (2.8), and set $\tilde{\alpha}_z = \rho_z e^{-i\tilde{\varphi}_z}$, where $\rho_z = \cos \tilde{\varphi}_z/|a_1|$. We find that $\tilde{\alpha}_z a_1 = a_\pm$ and $\tilde{\alpha}_z a_N = a_\varphi$, so

$$\tilde{\varphi}_z \leq \max\{|1 - a_\pm|, |1 - a_\varphi|\} = \sin |\varphi_1|$$  

and hence this $\mu_\varphi$ minimizes $\tilde{\varphi}(\mu_\varphi)$, which completes the proof. \hfill \Box

Note that $\mu_\varphi$ defined in (2.8) tends to $|\varphi + \lambda_1| - \lambda_1$ 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 $\tilde{\varphi}_z \leq \sin(3\pi/8) \approx 0.9239$.

On the right-hand side of Table 1, we see the dramatic effect of the preconditioner $B_z = (\mu_\varphi I + A)^{-1}$ 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 $\tilde{\alpha}_z = \rho_z e^{-i\tilde{\varphi}_z}$, computed as in the proof above. Notice that $\tilde{\varphi}_z = \|1 - \tilde{\alpha}_z A\|$ increases with $j$, whereas $\varepsilon_\varphi$ decreases.

Since computing the action of $(\mu_\varphi I + A)^{-1}$ 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

$$G_z = B_z A_z = \hat{\varphi} B_z + B_z (\mu_\varphi I + A) \quad \text{where } \hat{\varphi} := z - \mu_\varphi.$$  

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

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

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 $H_z := B_z (\mu_\varphi I + A)$ 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 $\mu_\varphi = 0$, 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 $\text{Im } z \geq 0$. 

[10] Iterative solution of shifted positive-definite linear systems 143

Downloaded from https://www.cambridge.org/core. IP address: 54.70.40.11, on 14 Jun 2019 at 15:40:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181112000107
Theorem 2.3. Assume that \( \zeta := \arg \hat{\zeta} \in [0, \pi) \), define \( \hat{\varphi}_z \) as the number in the interval \( J := (\zeta - \pi/2, \pi/2) \) that minimizes

\[
\hat{v}_z(\varphi) := \frac{m_z \cos^2 \varphi \cos(\zeta - \varphi)}{M_z \cos(\zeta - \varphi) + \Lambda_z \cos \varphi}
\]

where \( \Lambda_z := |\hat{\zeta}| \|B_z\| \),

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

\[
||I - \hat{\alpha}_z G_z|| \leq \hat{\varepsilon}_z := (1 - \hat{v}_z(\hat{\varphi}_z))^{1/2}
\]

where \( \hat{\alpha}_z := \hat{\rho}_z e^{-i\hat{\varphi}_z} \). (2.10)

If, in addition, there is a \( \gamma_z \geq 0 \) such that

\[
\text{Re}(\hat{\zeta} [B_z v, B_z(\mu_z I + A)v]) \leq -\gamma_z [B_z v, v] \quad \text{for all } v \in V,
\]

define \( \hat{\varphi}_z \) as the value in \( J \) minimizing

\[
\tilde{v}_z(\varphi) := \frac{m_z \cos^2 \varphi \cos(\zeta - \varphi)}{\max[M_z \cos(\zeta - \varphi), \Lambda_z \cos \varphi]}
\]

where \( \tilde{\Lambda}_z := \Lambda_z - \frac{2\gamma_z}{|\hat{\zeta}|} \),

and put \( \tilde{\rho}_z := \tilde{v}_z(\tilde{\varphi}_z) / (m_z \cos \tilde{\varphi}_z) \). We then have the sharper estimate

\[
||I - \tilde{\alpha}_z G_z|| \leq \tilde{\varepsilon}_z := (1 - \tilde{v}_z(\tilde{\varphi}_z))^{1/2}
\]

where \( \tilde{\alpha}_z := \tilde{\rho}_z e^{-i\tilde{\varphi}_z} \). (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

\[
||(I - \alpha G_z)v||^2 = ||v||^2 - 2 \text{Re}(\alpha[G_zv, v]) + |\alpha|^2 ||G_zv||^2 \leq (1 - \tilde{v}_z(\varphi))||v||^2.
\]

Here we have used (2.9) to obtain

\[
\text{Re}(\alpha[G_zv, v]) = \text{Re}(\alpha)[B_z(\mu_z I + A)v, v] + \text{Re}(\hat{\alpha} \hat{\zeta})[B_zv, v],
\]

where \( \alpha \) is chosen so that \( \text{Re}(\alpha) \) and \( \text{Re}(\hat{\alpha} \hat{\zeta}) \) are positive, and

\[
||G_zv|| \leq ||B_z(\mu_z I + A)v|| + ||\hat{\zeta}||B_zv||.
\]

If instead of the latter inequality one uses

\[
||G_zv||^2 = ||[B_z(\mu_z I + A)v]||^2 + ||\hat{\zeta}||B_zv||^2 + 2 \text{Re}(\hat{\zeta}[B_zv, B_z(\mu_z I + A)v]),
\]

then the estimate (2.11) makes it possible to replace \( \hat{v}_z(\varphi) \) by the smaller quantity \( \tilde{v}_z(\varphi) \); 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 \( \xi_j \), the value of \( k \) shown is the smallest such that (2.11) holds. The choice of \( \mu_z \) is the same as in Table 1.
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

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

Given $w_0$, a preliminary guess for the solution $w$, we define the residual $r_0 := g - A_zw_0$ and the associated Krylov subspace of order $n \geq 1$,

$$V_n := \text{span}\{r_0, A_zr_0, \ldots, A_z^{n-1}r_0\} = \text{span}\{r_0, Ar_0, \ldots, A^{n-1}r_0\},$$

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

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

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

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

and find that $v_n = w_n - w_0$ satisfies

$$(A_zv_n, \varphi) = (A_z(w_n - w_0), \varphi) = (g, \varphi) - (A_zw_0, \varphi) = (r_0, \varphi) \quad \text{for all } \varphi \in V_n.$$  

The homogeneous equation, $(A_zv, \varphi) = 0$ 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

$$(A_ze_n, \varphi) = 0 \quad \text{for all } \varphi \in V_n. \quad (3.3)$$
To study the convergence of $w_n$, we introduce the norm
\[ ||v||^2 := |z| ||v||^2 + (Av, v), \]
and begin with the following lemma.

**Lemma 3.1.** For all $v, w \in V$,
\[ |(A_z v, w)| \leq ||v|| ||w|| \quad \text{and} \quad |(A_z v, v)| \geq \cos(\phi/2) ||v||^2 \quad \text{where} \ \phi = \arg z. \]

**Proof.** The first part follows by the Cauchy–Schwarz inequality, since
\[
|A_z v, w| \leq |z| |(v, w)| + |(Av, w)| \leq (|z|^{1/2}||v||)(|z|^{1/2}||w||) + (Av, v)^{1/2}(Aw, w)^{1/2} \\
\leq (|z| ||v||^2 + (Av, v))^{1/2}(|z| ||w||^2 + (Aw, w))^{1/2}.
\]
Setting $\beta := e^{-i\phi/2}$, the second part now results from
\[
\text{Re}(\beta(A_z v, v)) = \text{Re}(\beta z)||v||^2 + \text{Re}\beta(Av, v) \\
\geq |z| \cos(\phi/2)||v||^2 + \cos(\phi/2) (Av, v) = \cos(\phi/2)||v||^2. \quad \Box
\]

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),
\[ ||w_n - w|| \leq \sec(\phi/2) \inf_{v \in w_0 + V_n} ||v - w|| \quad \text{with} \ \phi = \arg z. \]

**Proof.** Lemma 3.1 and (3.3) show that, for any $v \in w_0 + V_n$,
\[
\cos(\phi/2)||w_n - w||^2 \leq |(A_z(w_n - w), w_n - w)| = |(A_z(w_n - w), v - w)| \\
\leq ||w_n - w|| ||v - w||,
\]
which implies the stated result. \quad \Box

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
\[ ||e_n|| \leq \sec(\phi/2) \max_{1 \leq j \leq N} |Q_n(z + \lambda_j)| ||e_0|| \quad \text{with} \ \phi = \arg z. \]

**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
\[
\cos(\phi/2)\|e_n\| \leq \|v - w\| = \|Q_n(A_z)e_0\|.
\]
Since \( A_z \) is a normal operator,
\[
\|Q_n(A_z)e_0\| \leq \max_{1 \leq j \leq N} |Q_n(z + \lambda_j)| \|e_0\|.
\]
Similarly,
\[
(AQ_n(A_z)e_0, Q_n(A_z)e_0) \leq \max_{1 \leq j \leq N} |Q_n(z + \lambda_j)|^2 (Ae_0, e_0),
\]
and we conclude that
\[
\|||Q_n(A_z)e_0||| \leq \max_{1 \leq j \leq N} |Q_n(z + \lambda_j)| \|e_0|||
\]
which completes the proof. \( \square \)

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,
\[
\|||e_n||| \leq \sec(\phi/2)|T_n(-s_z)|^{-1}|||e_0||| \quad \text{where } s_z := \frac{\lambda_N + \lambda_1 + 2z}{\lambda_N - \lambda_1}.
\]

Here, with \( \kappa_z = (\lambda_N + z)/(\lambda_1 + z) \),
\[
T_n(-s_z) = \frac{\eta^n_z + \eta^{-n}_z}{2} \quad \text{where } \eta_z := -\frac{\sqrt{\kappa_z} - 1}{\sqrt{\kappa_z} + 1}, \text{ and } |\arg \sqrt{\kappa_z}| < \pi/2.
\]

Furthermore, \( |\eta_z| \leq 1 - c\lambda_N^{-1/2} \) 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
\[
\max_{\lambda \in [\lambda_1, \lambda_N]} |Q_n(\lambda + z)| = \max_{\tau \in [\lambda_1 + z, \lambda_N + z]} |Q_n(\tau)| = \max_{-1 \leq s \leq 1} \frac{|T_n(s)|}{|T_n(-s_z)|} = \frac{1}{|T_n(-s_z)|},
\]
and hence the first statement of the theorem follows by Theorem 3.3.
Defining $\theta$ by $\cosh \theta \equiv (e^\theta + e^{-\theta})/2 = -s_{\zeta}$ and letting $\eta_{\zeta} = e^\theta$,

$$T_n(-s_{\zeta}) = T_n(\cosh \theta) = \cosh(n\theta) = \frac{\eta_{\zeta}^n + \eta_{\zeta}^{-n}}{2}. $$

Here $\eta_{\zeta}$ satisfies the quadratic equation $\eta_{\zeta} + \eta_{\zeta}^{-1} = -2s_{\zeta}$, with roots

$$\eta_{\zeta} = s_{\zeta} \pm \sqrt{s_{\zeta}^2 - 1} = -\frac{(\sqrt{s_{\zeta} + 1} \mp \sqrt{s_{\zeta} - 1})^2}{2}. $$

Let $\eta_{\zeta} = \eta_{\zeta,+}$. Since $(s_{\zeta} + 1)/(s_{\zeta} - 1) = (\lambda_N + z)/(\lambda_1 + z) = \kappa_{\zeta}$, we find that

$$-\eta_{\zeta} = \sqrt{s_{\zeta} + 1} - \sqrt{s_{\zeta} - 1} = \frac{\kappa_{\zeta} - 1}{\sqrt{\kappa_{\zeta} + 1}} = \frac{1 - \kappa_{\zeta}^{-1/2}}{1 + \kappa_{\zeta}^{-1/2}}, $$

and since $|\arg(\sqrt{\kappa_{\zeta}})| < \pi/2$ we have $\text{Re} \sqrt{\kappa_{\zeta}} > 0$. Hence $|\eta_{\zeta}| < 1$.

For the final conclusion, we note that $\kappa_{\zeta}^{-1} = (z + \lambda_1)\lambda_N^{-1} + O(\lambda_N^{-4})$ and thus, because $|\arg(z + \lambda_1)| < |\phi| < \pi$, we have, for $\lambda_N$ large,

$$\text{Re} \, \kappa_{\zeta}^{-1/2} \geq \cos(\phi/2)\sqrt{|z + \lambda_1|\lambda_N^{-1/2}} + O(\lambda_N^{-2}) \geq c(z, \lambda_1)\lambda_N^{-1/2}, $$

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

$$|\eta_{\zeta}|^2 = 1 - 4 \text{Re} \, \kappa_{\zeta}^{-1/2} + O(\lambda_N^{-1}) \leq 1 - 2c(z, \lambda_1)\lambda_N^{-1/2} \quad \text{for } \lambda_N \text{ large,}$$

which completes the proof. \qed

Since $|\eta_{\zeta}| < 1$, it follows that $|T_n(-s_{\zeta})| \approx 2|\eta_{\zeta}|^n$, and so Theorem 3.4 shows linear convergence with approximately this rate. When $A = T_h$ and $\lambda_N = O(h^{-2})$, the error bound is thus of order $(1 - ch)^n$, with $c > 0$. The values of $|\eta_{\zeta}|$ 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 $|\eta_{\zeta}|$ with the corresponding values of $\varepsilon_{\zeta}$ 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 $B_{\zeta} = (\mu_{\zeta}I + A)^{-1}$, with $\mu_{\zeta} > -\lambda_1$. We multiply (3.1) by $(z - \mu_{\zeta})^{-1}B_{\zeta}$ to write the equation in the form

$$\tilde{z}w + B_{\zeta}w = \tilde{z}B_{\zeta}g \quad \text{where } \tilde{z} := (z - \mu_{\zeta})^{-1},$$

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

$$V_n = \text{span}\{r_0, B_{\zeta}r_0, \ldots, B_{\zeta}^{n-1}r_0\} \quad \text{with } r_0 = \tilde{z}B_{\zeta}g - (\tilde{z}I + B_{\zeta})w_0, \quad (3.6)$$
and the iterates \( w_n = w_0 + v_n \), with \( v_n \in V_n \), are defined by

\[
((\tilde{z}I + Bz)w_n, \varphi) = (\tilde{z} Bz \varphi, \varphi) \quad \text{for all } \varphi \in V_n.
\] (3.7)

The earlier analysis remains valid, with \( s_z \) now replaced by

\[
\tilde{s}_z := \frac{\lambda_N + \lambda_1 + 2\tilde{z}}{\lambda_N - \lambda_1} \quad \text{where } \lambda_j := (\mu_z + \lambda_{N+1-j})^{-1} \quad \text{for } j = 1, N,
\]

and correspondingly for \( \eta_z \). Since \( \lambda_N \) is bounded, Theorem 3.4 then shows that the error reduction factor is bounded away from 1, independently of \( \lambda_N \). We also show that the optimal choice of \( \mu_z \) 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 \( ||v||^2 = ||z||^2 + (B_z v, v) \),

\[
||e_n|| \leq \sec(\tilde{\phi}/2) |T_n(-\tilde{s}_z)|^{-1} ||e_0|| \quad \text{with } T_n(-\tilde{s}_z) = \frac{\tilde{\eta}_z^n + \tilde{\eta}_z^{-n}}{2}.
\]

Here

\[
\tilde{\eta}_z := \frac{\sqrt{\kappa_z} - 1}{\sqrt{\kappa_z} + 1} \quad \text{where } \tilde{\kappa}_z := \frac{\lambda_N + \tilde{z}}{\lambda_1 + \tilde{z}} = \frac{G_z(\lambda_1, \mu_z)}{G_z(\lambda_N, \mu_z)},
\] (3.8)

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

\[
|\tilde{\eta}_z| = \tan(|\tilde{\varphi}_z|/2) \quad \text{where } \tilde{\varphi}_z := \frac{\varphi_{z,1} - \varphi_{z,N}}{2} \quad \text{and } \varphi_{z,j} := \arg (z + \lambda_j).
\]
**Proof.** With \( a_j = G_z(\lambda_j, \mu_z) \) we have \( \tilde{k}_z = a_1/a_N \), and so it follows from (3.8) that

\[
\tilde{\eta}_z = -\frac{\sqrt{a_1/a_N} - 1}{\sqrt{a_1/a_N} + 1}.
\]

Since \( \arg a_j = \varphi_{z,j} \), we may write \( \sqrt{a_1/a_N} = \tau e^{i\varphi_z} \), with \( \tau > 0 \) and \( |\varphi_z| < \pi/2 \), and find that

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

The ratio is minimized at \( \tau = 1 \), that is, when \( |a_1| = |a_N| \), or when (2.8) holds. For \( \tau = 1 \), (3.9) becomes \( |\tilde{\eta}_z|^2 = \tan^2(\varphi_z/2) \), and since \( |\varphi_z| < \pi/2 \), this completes the proof.

Note that since \( |\varphi_z| < \pi/2 \),

\[
|\tilde{\eta}_z| = \tan(|\varphi_z|/2) < \sin |\varphi_z| = \tilde{e}_z,
\]

that is, the error reduction factor for CG is smaller than that for Richardson iteration. If \( z \in \Gamma \) then \( |\varphi_z| < 3\pi/8 \), 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 \( \mu_z \) 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

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

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

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

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

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

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 \( 1 \leq n < N^* \),

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

Putting \( p_0 := r_0 \), one may show that

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

making it possible to compute the \( w_n \) recursively by

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

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 \( \alpha_n \) 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 \( \alpha_n \) and \( \beta_n \), 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

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

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 \( \beta_{nk} \) satisfy the nonsingular lower-triangular linear system

\[
\sum_{k=0}^{j} (A_z p_k, p_j) \beta_{nk} = -(A_z \tilde{r}_{n+1}, p_j) \quad \text{for} \quad 0 \leq j \leq n < N^*.
\]
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
\[ w_{n+1} = w_n + \alpha_n p_n \quad \text{where} \quad \alpha_n := \frac{||\tilde{r}_n||^2}{(Az_p, p_n)} = \frac{(r_n, \tilde{r}_n)}{(Az_p, p_n)}. \]
The residuals again satisfy $r_{n+1} = r_n - \alpha_n Az_p$, implying that the preconditioned residuals satisfy $\tilde{r}_{n+1} = \tilde{r}_n - \alpha_n Gz_p$. 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 $\Omega$ 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 $\Omega$ 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 $\Omega$, 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
for the Richardson iterations, we take the optimal
several versions of the Richardson and CG iterations. As the acceleration parameter
\( j \) optimal for the INV preconditioner. Except for
\( \alpha \) from Theorem 2.3 for AMG(3). In all
cases, we use the value of the parameter \( \mu_z \) 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 \( \mu_z \) for several nearby
quadrature points, but we have not investigated the trade-off between the cost saving
and a possibly slower convergence.

\[ \frac{20}{20} \text{ Iterative solution of shifted positive-definite linear systems 153} \]

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

<table>
<thead>
<tr>
<th>( t )</th>
<th>( q = 10 )</th>
<th>( q = 20 )</th>
<th>( q = 30 )</th>
<th>( |u(t)|_h )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>( 1.3436 \times 10^{-2} )</td>
<td>( 4.3778 \times 10^{-4} )</td>
<td>( 4.1747 \times 10^{-4} )</td>
<td>0.4452</td>
</tr>
<tr>
<td>0.50</td>
<td>( 6.1232 \times 10^{-4} )</td>
<td>( 1.6260 \times 10^{-4} )</td>
<td>( 1.7541 \times 10^{-4} )</td>
<td>0.4623</td>
</tr>
<tr>
<td>1.00</td>
<td>( 2.2024 \times 10^{-4} )</td>
<td>( 2.1088 \times 10^{-4} )</td>
<td>( 2.1114 \times 10^{-4} )</td>
<td>0.4206</td>
</tr>
<tr>
<td>2.00</td>
<td>( 1.9403 \times 10^{-4} )</td>
<td>( 1.9411 \times 10^{-4} )</td>
<td>( 1.9411 \times 10^{-4} )</td>
<td>0.2579</td>
</tr>
</tbody>
</table>

**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 \( B_c = (\mu I + A)^{-1} \), 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 \( \mu_c I + A \).

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

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 \( \alpha = \bar{\alpha}_z \) from Theorem 2.2 in the
case of the INV preconditioner, and \( \alpha = \bar{\alpha}_z \) from Theorem 2.3 for AMG(3). In all
cases, we use the value of the parameter \( \mu_z \) 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 \( \mu_z \) 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
\[
\|e_n\| \leq \epsilon_j \quad \text{where} \quad \epsilon_j := \delta \times \frac{2\pi e^{|x_j|t}}{(2q+1)k|z_j'|} \quad \text{for} \quad \delta = 10^{-5},
\] (4.1)
where $x_j = \Re z_j < 0$. This way, the estimate (1.9) ensures that the additional error in $U_{q,h}(t)$ due to the iterative solver is less than $\delta$. For $j = 0$, we start each iteration with the zero vector, but for $j \geq 1$ 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 $\epsilon_j$. The former decrease whereas the latter increase with increasing $j$. In particular, the growth in $\epsilon_j$ more than compensates for the deterioration in the error reduction factors of the iterative solvers, seen in Tables 1 and 3.

References


