IN-64-CR
217923

On Conjugate Gradient Type Methods and Polynomial Preconditioners for a Class of Complex Non-Hermitian Matrices

Roland Freund

December 1988

Research Institute for Advanced Computer Science
NASA Ames Research Center

RIACS Technical Report 88.44

NASA Cooperative Agreement Number NCC 2-387

RIACS
Research Institute for Advanced Computer Science

On Conjugate Gradient Type Methods
and Polynomial Preconditioners
for a Class of Complex Non-Hermitian Matrices

Roland Freund

Institut für Angewandte Mathematik und Statistik
Universität Würzburg
D – 8700 Würzburg
Federal Republic of Germany
and
RIACS, Mail Stop 230-5
NASA Ames Research Center
Moffett Field, CA 94035, USA

Summary. We consider conjugate gradient type methods for the solution of large linear systems $Ax = b$ with complex coefficient matrices of the type $A = T + i\sigma I$ where T is Hermitian and σ a real scalar. Three different conjugate gradient type approaches with iterates defined by a minimal residual property, a Galerkin type condition, and an Euclidian error minimization, respectively, are investigated. In particular, we propose numerically stable implementations based on the ideas behind Paige and Saunders's SYMMLQ and MINRES for real symmetric matrices and derive error bounds for all three methods. It is shown how the special shift structure of A can be preserved by using polynomial preconditioning, and results on the optimal choice of the polynomial preconditioner are given. Also, we report on some numerical experiments for matrices arising from finite difference approximations to the complex Helmholtz equation.

Subject Classification: AMS(MOS): 65F10, 65N20, 41A50; CR: G1.3.

Running Title: Conjugate Gradient Methods for a Class of Complex Matrices

This work was supported in part by Cooperative Agreement NCC 2-387 between the National Aeronautics and Space Administration (NASA) and the Universities Space Research Association (USRA) and by National Science Foundation Grant DCR-8412314.

1. Introduction

The classical conjugate gradient method (CG) of Hestenes and Stiefel [23] — considered as an iterative method and used in conjunction with preconditioning (see e.g. [7]) — is one of the most powerful techniques for solving large linear systems

$$Ax = b \tag{1}$$

with Hermitian positive definite coefficient matrix A. A considerable part of the research in numerical linear algebra since the mid-seventies has been devoted to generalizations of CG to indefinite and non-Hermitian matrices (see e.g. [39,36,35] for a survey). Almost exclusively, the focus was on real matrices, and the case of complex coefficient matrices A has received surprisingly little attention. Among the few exceptions, where complex matrices are included, are two recent papers by Faber and Manteuffel [13,14] (see also Joubert and Young [26]) in which they develop a theory on the existence of conjugate gradient type methods with certain desirable features adopted from classical CG. They showed that "nice" extensions of CG — with iterates defined by an error minimization or a Galerkin type condition over Krylov subspaces generated by A and computable by an s-term recursion — exist essentially only for matrices of the form

$$A = e^{i\theta}(T + i\sigma I) \quad \text{where} \quad T = T^H \quad \text{is Hermitian} \quad , \quad \sigma, \theta \in \mathbb{R} \quad . \tag{2}$$

Here I denotes the identity matrix. Actual implementations of such CG-type methods are well known for the real matrices in the class (2). Paige and Saunders [30] were the first to devise numerically stable algorithms (SYMMLQ and MINRES) for real symmetric, but in general indefinite A. Concus, Golub [6], and Widlund [44] found a Galerkin type method for the subclass of real nonsymmetric matrices

$$A = I - N \quad \text{where} \quad N = -N^T \quad \text{is real and skew-symmetric} \quad . \tag{3}$$

The first minimal residual type algorithm for (3) was proposed by Rapoport [33] (see also [11,18] for different implementations).

In this paper, we present a detailed study, with the emphasis on practical aspects, of CG-type methods for arbitrary complex matrices of the form (2). Besides the two standard approaches based on a minimal residual property and a Galerkin condition, also a third less conventional method with iterates defined by an Euclidian error minimization is considered. In particular, it is shown how SYMMLQ and MINRES can be extended to numerically stable implementations of all three approaches, and we derive error bounds. For the practical use of CG-type methods it is crucial that they can be combined with efficient preconditioners. Unfortunately, the more classical techniques, such as incomplete factorization, lead to preconditioned matrices which in general are no longer in the class (2). We show that this problem can be resolved and the special structure of the matrices (2) preserved by using polynomial preconditioning, and results on the optimal choice of the preconditioner are given. Note that polynomial preconditioning is an attractive approach for vector and parallel computers and, because of that, has become very popular in recent years (see [35] for a survey).

Finally, we remark that large linear systems with complex coefficient matrices of type (2) or of the more general form

$$A = e^{i\theta}(T + i\sigma D) \quad \text{where} \quad T = T^H \quad \text{is Hermitian} \quad , \quad \sigma, \theta \in \mathbb{R} \quad , \tag{4}$$

where now D is a real positive semi-definite diagonal matrix, arise in important applications. Partial differential equations which model dissipative processes (e.g. [32, Chapter 10],[28]) usually involve complex coefficient functions and/or complex boundary conditions, and finite difference approximations lead to complex linear systems. A typical example is the complex Helmholtz equation

$$-\Delta u - \sigma_1 u + i\sigma_2 u = f \tag{5}$$

which describes the propagation of damped time-harmonic waves as e.g. electromagnetic waves in conducting media (e.g. [12, Chapter 8]). Further applications, which give rise to complex linear systems, include underwater acoustics [3,20,37], discretization of the time-dependent Schrödinger equation using implicit difference schemes [8], electromagnetic scattering problems [31], numerical computations in quantum chromodynamics [2], and numerical conformal mapping [43]. In these examples, the resulting matrices usually are of the form (4). Moreover, in almost all applications the case (2), $D = I$, occurs or D is at least close to I. Furthermore, T is typically indefinite, but has often, as for the discretized Helmholtz equation (5), only relatively few negative eigenvalues.

The paper is organized as follows. In Section 2, we define the three different CG-type approaches for matrices (2) via certain optimality poperties of their iterates and establish some connections with the Hermitian Lanczos algorithm. Actual implementations of all three methods are then presented in Section 3. We also point out how these algorithms can be adapted to matrices of the family (4) for the case that D is a positive definite diagonal matrix. In Section 4, other implementations which have been proposed in the literature are briefly reviewed and operation counts for the various algorithms are given. In Section 5, we derive error bounds and present some new results on related constrained approximation problems. In Section 6, polynomial preconditioning is considered. Finally, in Section 7, we report on some numerical experiments for matrices arising from finite difference approximations to the complex Helmholtz equation (2) with constant coefficients σ_1, σ_2.

Throughout the paper, all vectors and matrices, unless stated otherwise, are assumed to be complex. We use the notation

$$K_k(c, B) := \text{span}\,\{c, Bc, \ldots, B^{k-1}c\}$$

for the kth Krylov subspace of $\mathbf{C}^n$ generated by $c \in \mathbf{C}^n$ and the $n \times n$ matrix B. Moreover, $(x, y) = y^H x$ is the Euclidian inner product and $||x|| = \sqrt{x^H x}$ the associated norm. For a Hermitian positive definite matrix D, $||x||_D = \sqrt{x^H D x}$ denotes the D-norm of x. Finally, $\lambda_{\min}(T)$ (resp. $\lambda_{\max}(T)$) is the smallest (resp. largest) eigenvalue of the Hermitian matrix T.

2. Three conjugate gradient type approaches for shifted Hermitian matrices

We are concerned with the solution of complex linear systems (1) with $n \times n$ coefficient matrices of the form

$$A = T + i\sigma I \quad \text{where} \quad T = T^H \quad \text{is Hermitian} \quad , \quad \sigma \in \mathbb{R} \quad . \tag{6}$$

Clearly, by multiplication of the right-hand side b or the unknown vector x by $e^{-i\theta}$ the more general case (2) can always be reduced to (6). Although our main interest is in non-Hermitian A, we include the case $\sigma = 0$ and assume that $A = T$ is nonsingular then. This guarantees that A is always nonsingular, and the exact solution of (1) is denoted by $x_\star := A^{-1}b$.

We consider three different CG-type appropaches for solving (1). For any given starting vector $x_0 \in \mathbf{C}^n$, all three methods compute a sequence of approximations x_k, $k = 1, 2, \ldots$, to $x_\star$ which — in exact arithmetic — would terminate after at most n steps with $x_\star$. The first two approaches are based on the usual Krylov subspaces $K_k := K_k(r_0, A)$, $k = 1, 2, \ldots$, generated by A and the starting residual $r_0 := b - Ax_0$. The iterates of the minimal residual (MR) method are characterized by

$$||b - Ax_k|| = \min_{x \in x_0 + K_k} ||b - Ax|| \quad , \quad x_k \in x_0 + K_k \quad . \tag{7}$$

The Galerkin (GAL) (or orthogonal error [14]) approach aims at computing approximations x_k satisfying

$$(b - Ax_k, y) = 0 \quad \text{for all} \quad y \in K_k \quad , \quad x_k \in x_0 + K_k \quad . \tag{8}$$

Note that, for Hermitian positive definite A, this method is equivalent to the classical CG algorithm (see e.g. [30]). While MR and GAL are standard approaches for non-Hermitian matrices, the third method we propose is less conventional. Its iterates are defined by the minimal Euclidian error (ME) property

$$||x_\star - x_k|| = \min_{x \in x_0 + K_k^\star} ||x_\star - x|| \quad , \quad x_k \in x_0 + K_k^\star \tag{9}$$

where now

$$K_k^\star := K_k(A^H r_0, A) = \operatorname{span}\{A^H r_0, A\ A^H r_0, \ldots, A^{k-1} A^H r_0\} = A^H K_k \tag{10}$$

denotes the Krylov subspace generated by A and $A^H r_0$. Note that the last identity in (10) is true since matrices (6) are normal and thus

$$A\ A^H = A^H A \quad . \tag{11}$$

In the sequel, if it is not evident from the context which method we are considering, the superscripts MR, GAL, and ME will be used to distinguish iterates x_k and the corresponding residual vectors $r_k := b - Ax_k$ of the different approaches.

Remark 1. MINRES and SYMMLQ [30] are numerically stable implementations of the MR and GAL methods, respectively, for real symmetric matrices A. If A is indefinite, a Galerkin iterate satisfying (8) need not exist for every k, Paige and Saunders resolve this problem in SYMMLQ by actually working with a sequence of well-defined auxiliary vectors from which the existing Galerkin iterates can then be computed in a stable manner.

Remark 2. The ME approach (9) is a generalization of Fridman's method [17] for real symmetric matrices A. However, the algorithm he proposed is numerically unstable (see [18,40] for an explanation of the instability and a simple remedy). Fletcher [16] showed that the sequences of the Fridman iterates and the auxiliary vectors generated by SYMMLQ are mathematically equivalent. Therefore, as a by-product, SYMMLQ also yields a numerically stable implementation of Fridman's method.

We now turn to the derivation of algorithms, modelled after SYMMLQ and MINRES, for the actual computation of the iterates defined by (7)-(9). The main ingredient is the Lanczos algorithm [27] applied to the Hermitian part T of (6) and with r_0 as starting vector:

Algorithm 1 (Lanczos).

```
0) Set $v_0 = 0$, $\beta_1 = ||r_0||$, $v = r_0$.
For $k = 1, 2, \ldots$
1) If $\beta_k = 0$, stop.
   Otherwise, compute
2) $v_k = v/\beta_k$,
   $v = Tv_k - \alpha_k v_k - \beta_k v_{k-1}$  with  $\alpha_k = (Tv_k, v_k)$ ,
   $\beta_{k+1} = ||v||$ .
```

We refer to [21, pp. 325] for a detailed discussion of the Lanczos algorithm; in particular, proofs of the properties collected in the following proposition can be found there. We set

$$V_k := (v_1, v_2, \ldots, v_k) \quad \text{and} \quad T_k := \begin{pmatrix} \alpha_1 & \beta_2 & 0 & \cdots & 0 \\ \beta_2 & \alpha_2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \beta_k \\ 0 & \cdots & 0 & \beta_k & \alpha_k \end{pmatrix} .$$

Proposition 1. a) *In exact arithmetic, Algorithm 1 stops for $k = m + 1$ where $m := \dim K_n(r_0, T)$.*
b) For $k = 1, 2, \ldots, m$:

$$V_k^H V_k = I_k \ , \ \textit{the } k \times k \textit{ identity matrix} \quad , \tag{12}$$

$$TV_k = V_k T_k + \beta_{k+1} v_{k+1} e_k^T \quad , \quad e_k := (0, \ldots, 0, 1)^T \in \mathbb{R}^k \quad , \quad v_{m+1} := 0 \quad , \tag{13}$$

$$K_k(r_0, T) = \operatorname{span}\{v_1, v_2, \ldots, v_k\} \quad . \tag{14}$$

Remark 3. The termination index m is just the minimal number of components in any expansion of r_0 into orthonormal eigenvectors of T, i.e.

$$r_0 = \sum_{l=1}^{m} \rho_l u_l \quad , \tag{15}$$

$$\rho_l \neq 0 \quad , \quad T u_l = \lambda_l u_l \quad , \quad (u_j, u_l) = \delta_{jl} \quad , \quad \lambda_j \neq \lambda_l \quad , \quad l, j, = 1, \ldots, m \quad ,$$

and no such representation with a smaller m exists.

From now on , we assume that $k \in \{1, 2, \ldots, m\}$. Note that (14) is equivalent to

$$K_k = \operatorname{span}\{v_1, v_2, \ldots v_k\} = \{V_k z \mid z \in \mathbf{C}^k\} \quad , \tag{14'}$$

since A and T differ only by a scalar multiple of the identity matrix. Moreover, with

$$S_k = \begin{pmatrix} T_k + i\sigma I_k \\ \beta_{k+1} e_k^T \end{pmatrix} \tag{16}$$

and by adding $i\sigma V_k$ to both sides of (13), we obtain

$$A V_k = V_{k+1} S_k \quad . \tag{13'}$$

Remark that, since A is nonsingular and V_k has full column rank, (13′) implies

$$\operatorname{rank} S_k \; = \; k \quad . \tag{17}$$

Next, we rewrite (7)-(9) in terms of V_k and S_k. $e_1 := (1, 0, \ldots, 0)^T$ denotes the first unit vector; its actual length will be clear from the context.

Proposition 2. *a) $x_k^{MR} = x_0 + V_k z_k^{MR}$ where z_k^{MR} is the solution of*

$$\beta_1 S_k^H e_1 = S_k^H S_k z \quad . \tag{18}$$

b) $x_k^{ME} = x_0 + A^H V_k z_k^{ME}$ where z_k^{ME} is the solution of

$$\beta_1 e_1 = S_k^H S_k z \quad . \tag{19}$$

c) $x_k^{GAL} = x_0 + V_k z_k^{GAL}$ where z_k^{GAL} is the solution of

$$\beta_1 e_1 = (T_k + i\sigma I_k) z \quad . \tag{20}$$

Moreover, if $\sigma = 0$ and T_k is singular, then no Galerkin iterate satisfying (8) exists.

d) $x_m^{MR} = x_m^{ME} = x_m^{GAL} = x_\star$.

Proof. First, recall that $r_0 = \beta_1 v_1$ and thus, by (12)

$$V_j^H r_0 = \beta_1 e_1 \quad , \quad j = 1, 2, \ldots, m+1 \quad . \tag{21}$$

Using (13′) and (12), we obtain

$$V_k^H A^H A\ V_k = S_k^H S_k \quad . \tag{22}$$

a) In view of (14′), $x_k = x_0 + V_k z_k$ where z_k is the solution of

$$\min_{z \in \mathbf{C}^k} \ \|r_0 - AV_k z\| \quad \text{or, equivalently,} \quad V_k^H A^H r_0 = V_k^H A^H A\ V_k z \quad . \tag{7′}$$

By (22), (13′), and (21) (for $j = k+1$), the linear system in (7′) is the same as (18).
b) (10), (14′), and (11) imply that

$$x_k = x_0 + A^H V_k y_k \quad , \quad r_k = r_0 - A^H A\ V_k y_k \quad \text{with} \quad y_k \in \mathbf{C}^k \quad .$$

The minimization property (9) is equivalent to

$$0 = (x_\star - x_k, A^H v) = (r_k, v) \quad \text{for all} \quad v \in K_k \quad .$$

By (14′), it suffices to consider these equations for $v = v_j$, $j = 1, \ldots, k$, and it follows that y_k is the solution of

$$V_k^H r_0 = V_k^H A^H A\ V_k y$$

which, by (21) (for $j = k$) and (22), is just the linear system (19).
For c), we similarly obtain that $x_k = x_0 + V_k y_k$ satisfies (8) iff y_k solves the linear system

$$\beta_1 e_1 = V_k^H V_{k+1} S_k y$$

whose coefficient matrix, by (12) and (16), is $T_k + i\sigma I_k$. If $\sigma = 0$ and T_k is singular, the linear system (20) could have a solution only if it was consistent. Using the fact that T_{k-1} is nonsingular then and $\beta_k > 0$, one easily verifies that this can not be the case.
d) Part a) of Proposition 1 implies that K_m is an invariant subspace for T and, since $r_0 \in K_m$, we conclude that

$$x_\star - x_0 = A^{-1} r_0 \in A^{-1} K_m = K_m = A^H K_m \quad .$$

On the other hand, $x_\star$ trivially satisfies (7)-(9) and it follows that $x_m = x_\star$ for all three methods. ∎

3. Practical implementations

For the special case of real symmetric matrices A, the result of Proposition 2 is due to Paige and Saunders [30], and their derivation of SYMMLQ and MINRES is based on solving (18)-(20) by an LQ factorization of T_k. Using similar ideas, it is possible to obtain practical implementations for the MR, ME, and GAL methods for the class of non-Hermitian matrices (6). In the following, we sketch such derivations. Details which are completely analogous to the real symmetric case will be omitted, and we refer the reader to [30] instead.

For the solution of the linear systems (18)-(20) a QR factorization of the $k+1 \times k$ matrix (16) S_k of the type

$$Q_k S_k = \begin{pmatrix} R_k \\ 0 \end{pmatrix} \tag{23}$$

is used, where Q_k is a $k+1 \times k+1$ unitary matrix and R_k a nonsingular matrix of the form

$$R_k = \begin{pmatrix} \gamma_1 & \delta_2 & \epsilon_3 & 0 & \cdots & 0 \\ 0 & \gamma_2 & \delta_3 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \gamma_3 & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \epsilon_k \\ \vdots & & & \ddots & \ddots & \delta_k \\ 0 & \cdots & \cdots & \cdots & 0 & \gamma_k \end{pmatrix} .$$

Note that S_k is tridiagonal and, by (17), has full rank, and a decomposition (23) clearly exists. Moreover, since

$$S_k^H S_k = T_k^2 + \sigma^2 I_k + \beta_{k+1}^2 e_k e_k^T$$

is a real matrix, we can choose Q_k such that R_k is real, and this will help to reduce complex arithmetic in the final algorithms. Using standard matrix calculus, one verifies that a factorization (23) with real R_k can be achieved with a unitary matrix Q_k of the form

$$Q_k = Q_{k,k+1}\, D_k\, Q_{k-1,k}\, D_{k-1} \cdots D_3\, Q_{2,3}\, D_2\, Q_{1,2}\, D_1$$

with complex diagonal matrices

$$D_j = \mathrm{diag}(1,\ldots,1,\underset{\overset{\uparrow}{j}}{e^{i\varphi_j}},1,\ldots,1) \quad , \quad \varphi_j \in \mathbb{R} \quad ,$$

and real Givens rotations

$$Q_{j,j+1} = \begin{pmatrix} 1 &&&&&&& \\ & \ddots &&&&&& \\ && 1 &&&&& \\ &&& c_j & s_j &&& \\ &&& -s_j & c_j &&& \\ &&&&& 1 && \\ &&&&&& \ddots & \\ &&&&&&& 1 \end{pmatrix} \begin{array}{l} \\ \\ \\ \leftarrow j \\ \leftarrow j+1 \\ \\ \\ \\ \end{array} \quad , \quad c_j, s_j \in \mathbb{R},\ c_j^2 + s_j^2 = 1 \quad .$$

Moreover, the factorization is easily updated from the one of the previous step $k-1$ by simply setting

$$\begin{aligned} &\epsilon_k = s_{k-2}\beta_k \quad , \quad \delta_k = s_{k-1}\alpha_k + c_{k-1}c_{k-2}\beta_k \cos\varphi_{k-1} \quad , \\ &h_k = -s_{k-1}c_{k-2}\beta_k e^{-i\varphi_{k-1}} + c_{k-1}(\alpha_k - i\sigma) \quad , \quad \tilde{\gamma}_k = |h_k| \quad , \\ &e^{i\varphi_k} = \begin{cases} h_k/|h_k| & \text{if} \quad h_k \neq 0 \\ 0 & \text{if} \quad h_k = 0 \end{cases} \end{aligned} \tag{24}$$

and

$$\gamma_k = \sqrt{\tilde{\gamma}_k^2 + \beta_{k+1}^2} \quad , \quad c_k = \tilde{\gamma}_k/\gamma_k \quad , \quad s_k = \beta_{k+1}/\gamma_k \quad . \tag{25}$$

Note that

$$Q_k = Q_{k,k+1} D_k \begin{pmatrix} Q_{k-1} & 0 \\ 0 & 1 \end{pmatrix} \quad . \tag{26}$$

Based on the factorization (23), the solutions of (18)-(20) and therefore, in view of Proposition 2, the actual iterates x_k of the MR, ME, and GAL methods can now be computed in a numerically stable manner. Moreover, by arranging the calculations in analogy to the real symmetric case [30] it is possible to obtain simple recursions for the x_k.

First, we consider the MR method. With (18) and (23), it follows that

$$x_k = x_0 + V_k R_k^{-1} t_k \quad \text{where} \quad t_k := \beta_1 (I_k \ 0) Q_k e_1 \in \mathbf{C}^k \quad .$$

(26) shows that t_k differs from t_{k-1} only by its kth entry η_k and with p_k denoting the last column of $V_k R_k^{-1}$ the recursion

$$x_k = x_{k-1} + \eta_k p_k$$

results (cf. [30]). In combination with Algorithm 1, this leads to the following implementation.

Algorithm 2 (MR method).

```
0) Choose $x_0 \in C^n$ and set $v = b - Ax_0$, $v_0 = p_0 = p_{-1} = 0$,
   $\beta_1 = \tilde{\eta}_1 = ||v||$, $c_0 = c_{-1} = 1$, $s_0 = s_{-1} = \varphi_0 = 0$.
For $k = 1, 2, \ldots$
1) If $\beta_k = 0$, stop: $x_{k-1}$ solves $Ax = b$.
   Otherwise, compute
2) $v_k = v/\beta_k$, $\alpha_k = (Tv_k, v_k)$,
   $v = Tv_k - \alpha_k v_k - \beta_k v_{k-1}$, $\beta_{k+1} = ||v||$,
   and then $\epsilon_k$, $\delta_k$, $\gamma_k$, $\varphi_k$, $c_k$, $s_k$ using formulae (24), (25),
3) $p_k = (v_k - \delta_k p_{k-1} - \epsilon_k p_{k-2})/\gamma_k$,
   $x_k = x_{k-1} + \eta_k p_k$ with $\eta_k = c_k \tilde{\eta}_k e^{i\varphi_k}$,
   $\tilde{\eta}_{k+1} = -s_k \tilde{\eta}_k e^{i\varphi_k}$.
```

We now turn to the ME and GAL methods. With (23) and by setting

$$W_k = (w_1, w_2, \ldots, w_k) := A^H V_k R_k^{-1} \quad ,$$

it follows from part b) of Proposition 2 that

$$x_k^{ME} = x_0 + W_k y_k \quad \text{where } y_k \text{ is the solution of } \beta_1 e_1 = R_k^T y \quad .$$

Similarly, using that, by (16),

$$T_k + i\sigma I_k = S_k^T \begin{pmatrix} I_k \\ 0 \end{pmatrix}$$

and with (23), (26), one deduces from (20) that x_k^{GAL} exists iff $c_k \neq 0$ and then

$$x_k^{GAL} = x_0 + \tilde{W}_k \tilde{y}_k \quad , \quad \tilde{W}_k := V_k Q_{k-1}^T \operatorname{diag}(1, 1, \ldots, 1, e^{i\varphi_k}) \quad ,$$

where $\tilde{y}_k$ is the solution of

$$\beta_1 e_1 = R_k^T \operatorname{diag}(1, \ldots, 1, c_k)\, \tilde{y} \quad .$$

Clearly, y_k and $\tilde{y}_k$ differ only in their last elements η_k and $\tilde{\eta}_k$. Moreover, with (13′), (23), and (26), one easily verifies that $\tilde{W}_k$ is identical to W_k up to its last column $\tilde{w}_k$. Hence, we obtain the recursions

$$x_k^{ME} = x_{k-1}^{ME} + \eta_k w_k \quad \text{and, if } c_k \neq 0, \quad x_k^{GAL} = x_{k-1}^{ME} + \tilde{\eta}_k \tilde{w}_k \tag{27}$$

(cf. [30]). The resulting implementations can be summarized as follows:

Algorithm 3 (ME/GAL method).

```
0) Choose $x_0 \in C^n$ and set $x_0^{ME} = x_0^{GAL} = x_0$, $v = b - Ax_0$, $v_0 = \tilde{w}_0 = 0$,
   $\beta_1 = ||v||$, $\eta_0 = -1$, $c_0 = c_{-1} = 1$, $s_0 = s_{-1} = \varphi_0 = \eta_{-1} = 0$.
   If $\beta_1 > 0$, set $v_1 = v/\beta_1$.
For $k = 1, 2, \ldots$
1) If $\beta_k = 0$, stop: $x_{k-1}^{ME} = x_{k-1}^{GAL} = x_\star$ solves $Ax = b$.
   Otherwise, compute
2) $\alpha_k = (Tv_k, v_k)$,
   $v = Tv_k - \alpha_k v_k - \beta_k v_{k-1}$, $\beta_{k+1} = ||v||$,
   and then $\epsilon_k$, $\delta_k$, $\tilde{\gamma}_k$, $\gamma_k$, $\varphi_k$, $c_k$, $s_k$ using formulae (24), (25),
3) $\tilde{w}_k = e^{i\varphi_k}\ (-s_{k-1}\tilde{w}_{k-1} + c_{k-1}v_k)$
   and, if $\tilde{\gamma}_k \neq 0$ and the Galerkin iterate is desired,
   $x_k^{GAL} = x_{k-1}^{ME} + \tilde{\eta}_k \tilde{w}_k$ with $\tilde{\eta}_k = -(\delta_k \eta_{k-1} + \epsilon_k \eta_{k-2})/\tilde{\gamma}_k$.
4) Set $v_{k+1} = v/\beta_{k+1}$, if $\beta_{k+1} > 0$, and $v_{k+1} = 0$ otherwise,
   $w_k = c_k \tilde{w}_k + s_k v_{k+1}$,
   $x_k^{ME} = x_{k-1}^{ME} + \eta_k w_k$ with $\eta_k = -(\delta_k \eta_{k-1} + \epsilon_k \eta_{k-2})/\gamma_k$.
```

The finite termination property of the Lanczos algorithm does no longer hold in the presence of roundoff error (see e.g. [21, pp. 332]), and the stopping criterion stated in Algorithms 2 and 3 is not useful in practice. Instead, one should terminate the iteration as soon as $||r_k||$ is sufficiently reduced. Note that, similar to the real symmetric case [30], $||r_k||$ can be obtained without computing the vector r_k itself by using the following identities:

$$\begin{aligned} ||r_k^{ME}|| &= \sqrt{\eta_{k+1}^2 \gamma_{k+1}^2 + \eta_k^2 \epsilon_{k+2}^2} \quad , \\ ||r_k^{MR}|| &= \beta_1 s_1 s_2 \cdots s_k \quad , \\ ||r_k^{GAL}|| &= \beta_{k+1} |s_{k-1} \eta_{k-1} + c_{k-1} \tilde{\eta}_k e^{i\varphi_k}| \quad . \end{aligned}$$

Finally, consider linear systems $Ax = b$ with coefficient matrices A of the more general class (4) with D a positive definite diagonal matrix. Then, $Ax = b$ is equivalent to the linear system

$$A'x' = b' \quad \text{where} \quad A' = e^{-i\theta} D^{-1/2} A D^{-1/2} \ , \ x' = D^{1/2} x \ , \ b' = e^{-i\theta} D^{-1/2} b \quad ,$$

whose coefficient matrix A' is now of the form (6), so that we can use Algorithm 2 or 3 for its solution. Note that one never needs to form A' and b' explicitly, and it is straightforward to rewrite both Algorithms 2 and 3 in terms of the original linear system $Ax = b$. We omit the details and only state that the resulting MR, ME, and GAL algorithms generate iterates which are characterized by the properties (7) (with $||\ || = ||\ ||_{D^{-1}}$), (9) (with $||\ || = ||\ ||_D$), and (8), respectively, where now

$$K_k = K_k(D^{-1} r_0, D^{-1} A) \quad , \quad K_k^\star = K_k(D^{-1} A^H D^{-1} r_0, D^{-1} A) \quad .$$

4. Comparisons with other implementations. Operation counts

Several authors [26,38,1] have proposed algorithms for the computation of the MR resp. GAL iterates (7) resp. (8). However, most of these implementations (like Orthomin and Orthores in [26]) are modelled after variants of the conjugate residual or conjugate gradient algorithm for Hermitian positive definite matrices. It is well known [30,5,40] that, for Hermitian indefinite A, these approaches are numerically unstable and can even break down; e.g. for the GAL method this occurs whenever a Galerkin iterate does not exist (cf. [30] and part c) of Proposition 2.). The same stability problems can arise for the non-Hermitian matrices (6) if σ is small. Hence, all these algorithms derived directly from the positive definite case are stable only for matrices (6) which fulfill additional requirements such as T positive definite or $|\sigma|$ bounded away from 0. Note that these two conditions are not satisfied for most of the applications mentioned in the introduction.

Here, we consider only implementations which are numerically stable for the general class of matrices (6). Among the proposed algorithms in the literature merely the Orthodir approach [26,1] for the computation of the MR iterates has this property. This algorithm can be stated as follows.

Algorithm 4 (Orthodir MR implementation).

```
    0) Choose $x_0 \in C^n$ and set $s_0 = r_0 = b - Ax_0$,
       $q_0 = As_0$, $s_{-1} = q_{-1} = 0$, $\nu_0 = 0$.
    For $k = 0, 1, \dots$
    1) If $q_k = 0$, stop: $x_k$ solves $Ax = b$.
       Otherwise, compute
    2) $\lambda_k = (r_k, q_k)/||q_k||^2$,
       $x_{k+1} = x_k + \lambda_k s_k$, $r_{k+1} = r_k - \lambda_k q_k$,
    3) $\mu_k = (Tq_k, q_k)/||q_k||^2$ and, if $k > 0$, $\nu_k = ||q_k||^2/||q_{k-1}||^2$,
       $s_{k+1} = q_k - (\mu_k + i\sigma)s_k - \nu_k s_{k-1}$,
       $q_{k+1} = Tq_k - \mu_k q_k - \nu_k q_{k-1}$.
```

We remark that $q_k = As_k$ and that the search directions s_k are up to scalar factors identical to the vectors p_k in Algorithm 2.

Next, the results of operation counts for Algorithms 2,3,4 are presented in Table 4.1. Although we solve complex linear systems, most of the scalars (like α_k and β_k in the Lanczos step of Algorithm 2 and 3) occuring in the computations are real. Moreover, on some machines, implementations in real arithmetic are more advantageous. Therefore, we compare work and storage in terms of real quantities. Listed are the number of matrix-vector products $T \cdot v$, $v \in \mathbb{R}^n$, the approximate number m of additional real multiplications per iteration, and the number s of real vectors (of length n) to be stored. The computation of inner products often constitutes a bottle-neck on modern computers. For this reason, we also give the number dp of dot products $x \cdot y$, $x, y \in \mathbb{R}^n$ per iteration. Finally, notice that — based on the simple observation stated in Proposition 3 below — work and storage for the MR and ME/GAL methods can be significantly reduced if the Hermitian part T of the matrix (6) is real. This case occurs frequently in the cited application, and we included the corresponding operation counts in Table 4.1.

Proposition 3. *Let T be real and assume that $r_0 := b - Ax_0 \in \mathbb{R}^n$. Then, all the vectors v_k, $k = 0, 1, \dots$, in Algorithm 2 and 3 are real. In addition, for the MR method, all search directions p_k are real vectors.*

Note that often the right-hand side b is a real vector, and then the standard starting guess $x_0 = 0$ gurantees that r_0 is real. In the general case $b \in \mathbf{C}^n$ and if $\sigma \neq 0$, the condition $r_0 \in \mathbb{R}^n$ can always be fulfilled by choosing the starting vector $x_0 = x_0^{(1)} + ix_0^{(2)}$ appropriately, e.g. $x_0^{(2)} = 0$ and $x_0^{(1)} = \operatorname{Im} b/\sigma$. However, such a strategy might not be desirable, if one already knows a good approximation x_0 for the exact solution of $Ax = b$.

Table 4.1

To explain the numbers given in Table 4.1, a few more comments are necessary. For the ME/GAL algorithm, we have assumed that the Galerkin is, if desired, only computed in the very last step of the iteration. Furthermore, in order to reduce the computational work, note that, in the MR Algorithm 2, one computes the vector $\gamma_k p_k$ instead of p_k. Similarly, in part 4) of Algorithm 3, the vector w_k itself is never needed and, hence, $\eta_k w_k$ is generated directly. Moreover, using fast Givens rotations (e.g. [21, p.158]), we compute the rescaled vector $f_k \tilde{w}_k$ instead of $\tilde{w}_k$ in step 3) of Algorithm 3. Here, $f_k := 1/(c_{k-1} \cos \varphi_k)$ for the case that $s_{k-1} \leq c_{k-1}$ and $|\sin \varphi_k| \leq |\cos \varphi_k|$, and f_k is defined correspondingly for the remaining cases. Note that then only $4n$ real multiplications are needed for updating $f_k \tilde{w}_k$ from $f_{k-1} \tilde{w}_{k-1}$ and v_k.

We conclude this section with a few further remarks. First, Table 4.1 clearly shows that the MR implementation stated in Algorithm 2 is less expensive than the Orthodir Algorithm 4. For real symmetric linear systems, Algorithm 2 and 3 reduce to MINRES and SYMMLQ [30], respectively. Notice that, for the case of complex matrices (6) with T and r_0 real, Algorithm 2 and 3 require only little extra work and storage compared to MINRES and SYMMLQ. Finally, consider real linear systems with matrices (3) $A = I - N$ with $N = -N^T$ (or, equivalently, $A' = iA = T + iI$ with $T = -iN = T^H$ if rewritten in the form (6)). It can be shown, that for this case, the Galerkin part of Algorithm 3 is equivalent to the method of Concus, Golub [6], and Widlund [44]. Also, note that, in [18,39], we have investigated an Orthodir type implementation of the ME approach for the class $A = I - N$.

5. Error bounds

In this section, we derive error bounds for the MR and ME methods. Let $\alpha \leq \lambda_{\min}(T)$ and $\beta \geq \lambda_{\max}(T)$ be given bounds for the extreme eigenvalues of T. Therefore, all eigenvalues of A are contained in the complex line segment $S := [\alpha + i\sigma, \beta + i\sigma]$. For the rest of the paper, we assume that in the Hermitian case $\sigma = 0$, $A = T$ is positive definite and $0 < \alpha < \beta$. This guarantees that $0 \notin S$. We denote by Π_k the set of all complex polynomials of degree at most k.

By the standard technique, using

$$K_k(r_0, A) = \{q(A) r_0 \mid q \in \Pi_{k-1}\} \tag{28}$$

and the expansion (15) of r_0 (note that A and T have the same eigenvectors!), one obtains from (7) the estimate

$$\|r_k^{MR}\| / \|r_0\| \leq \min_{q \in \Pi_{k-1}} \max_{\lambda \in S} |1 - \lambda q(\lambda)| \quad . \tag{29}$$

Similarly, with (10), (28), we deduce from (9) that

$$\|x_* - x_k^{ME}\| / \|x_* - x_0\| \leq \min_{q \in \Pi_{k-1}} \max_{\lambda \in S} |1 - |\lambda|^2 q(\lambda)| \quad . \tag{30}$$

With the linear transformation

$$z = z(\lambda) = \frac{2(i\sigma - \lambda) + \beta + \alpha}{\beta - \alpha} \quad , \tag{31}$$

which maps S onto the unit interval $[-1,1]$, the right-hand side of (29) can be rewritten in the form

$$(E_k(a) :=) \quad \min_{p\in\Pi_k : p(a)=1} \quad \max_{z\in[-1,1]} |p(z)| \tag{32}$$

where

$$a := \frac{2i\sigma + \beta + \alpha}{\beta - \alpha} \quad \notin \ [-1,1] \quad . \tag{33}$$

Furthermore, using the identity

$$4\,|\lambda|^2 = (\beta-\alpha)^2(z(\lambda)-a)(z(\lambda)-\bar{a}) \quad , \quad \lambda \in S \quad ,$$

one easily verifies that the upper bound in (30) is just $E^{(r)}_{k+1}(a)$ where

$$\begin{aligned}(E^{(r)}_k(a) : =) \quad &\min_{p\in\Pi_k(a)} \quad \max_{z\in[-1,1]} |p(z)| \quad , \\ \Pi_k(a) := \{p \in \Pi_k \mid p(a) = p(\bar{a}) = 1 \quad &\text{and, if} \quad a \in \mathbb{R},\ p'(a) = 0\} \quad .\end{aligned} \tag{34}$$

We now turn our attention to the two approximation problems (32) and (34). It will be convenient to represent a in the form

$$\begin{aligned} a = a(\psi) = a(R)\cos\psi + ib(R)\sin\psi \quad , \quad R > 1 \quad , \quad 0 \le \psi < 2\pi \quad , \\ a(R) := \frac{1}{2}(R + \frac{1}{R}) \quad , \quad b(R) := \frac{1}{2}(R - \frac{1}{R}) \quad ; \end{aligned} \tag{35}$$

clearly, this is possible for any $a \notin [-1,1]$. For fixed $R > 1$, we set $\mathcal{E}_R = \{a = a(\psi) | 0 \le \psi < 2\pi\}$ and remark that $\mathcal{E}_R$ describes the boundary of an ellipse with foci at ± 1 and semi-axes $a(R)$, $b(R)$.

First, we consider the complex approximation problem (32). Its solution is classical for the case of real a where $T_k(z)/T_k(a)$ is the optimal polynomial. Here, T_k denotes the kth Chebyshev polynomial which, by means of the Joukowsky map, is given by

$$T_k(z) = \frac{1}{2}(v^k + \frac{1}{v^k}) \quad , \quad z = \frac{1}{2}(v + \frac{1}{v}) \quad . \tag{36}$$

For purely imaginary a, the extremal polynomials were found by Freund and Ruscheweyh [19], but for general complex a the solution of (32) is not explicitly known. We now derive a new, very useful upper bound for the optimal value of (32).

Theorem 1. *Let $R > 1$ and $k = 1, 2, \ldots$. Then,*

$$\frac{1}{R^k} < E_k(a) \le \frac{2}{R^k + 1/R^k} \quad , \quad a \in \mathcal{E}_R \quad . \tag{37}$$

Proof. The lower bound follows immediately from an inequality due to S.N. Bernstein (e.g. [29, Theorem 74]). In order to obtain the upper bound, we define the polynomial

$$p(z) = \frac{(R^k - 1/R^k)T_k(z) + 2i\sin(k\psi)}{(R^k - 1/R^k)T_k(a) + 2i\sin(k\psi)} \quad , \tag{38}$$

where ψ is given by the representation (35) of a. Clearly, $p(a) = 1$, and, by using (35), (36), and the fact that $-1 \leq T_k(z) \leq 1$ on $[-1,1]$, one readily verifies that the uniform norm of p on $[-1,1]$ is just the stated upper bound. ∎

Remark 4. For fixed $R > 1$, the upper bound in (37) is optimal, with equality holding for the two real points of $\mathcal{E}_R$. The optimal lower bound is unknown, but it is conjectured to be $2/(R^k + R^{k-2})$ which is just the optimal value of (32) for the two purely imaginary points of $\mathcal{E}_R$ (cf. [19]).

Remark 5. The polynomials (38) were recently introduced by Fischer and Freund [15], and it was shown that they are the optimal solutions for certain constrained approximation problems on ellipses.

Next, we study the approximation problem (34), and we will show that it is closely related to the classical Zolotarev problem

$$\min_{q \in \Pi_{k-2}} \max_{z \in [-1,1]} |z^k + \eta k z^{k-1} - q(z)| \quad , \quad \eta \in \mathbb{R} \quad , \quad k = 2, 3, \ldots \quad . \tag{39}$$

It is well known that there always exists a unique best approximation $q_k(z;\eta)$ for (39) and the corresponding polynomials

$$Z_k(z;\eta) = z^k + \eta k z^{k-1} - q_k(z;\eta) \quad , \quad \eta \in \mathbb{R} \quad , \quad k = 2, 3, \ldots \quad ,$$

are called Zolotarev polynomials. We refer the reader to [4] for a detailed study of these polynomials. Note that

$$Z_k(z;\eta) = 2^{1-k}(1 + |\eta|)^k \, T_k\left(\frac{z + \eta}{1 + |\eta|}\right) \quad \text{for} \quad |\eta| \leq \tan^2 \frac{\pi}{2k} \quad , \tag{40}$$

and for the remaining values of η there are representations of $Z_k(z;\eta)$ in terms of elliptic functions.

Theorem 2. *Let $a = a(\psi) \in \mathcal{E}_R$, $R > 1$, $k = 2, 3, \ldots$. Then, there exists a unique optimal polynomial $p_k(z;a)$ for (34). If $\psi = j\pi/(k-1)$ with an integer $j \not\equiv 0 \bmod k-1$, then*

$$p_k(z;a) = \frac{T_{k-1}(z)}{T_{k-1}(a)} \quad \textit{and} \quad E_k^{(r)}(a) = \frac{2}{R^{k-1} + 1/R^{k-1}} \quad .$$

Otherwise,

$$p_k(z;a) \;=\; \frac{Z_k(z;\eta)}{Z_k(a;\eta)} \tag{41}$$

where $\eta = \eta(a)$ is the unique solution of

$$\operatorname{Im} \; Z_k(a;\eta) = 0 \quad (\textit{resp.} \quad Z_k'(a;\eta) = 0 \; , \; \textit{if} \; a \in \mathbb{R}) \quad , \quad \eta \in \mathbb{R} \quad . \tag{42}$$

In particular, if ψ satisfies for some integer $j \not\equiv 0 \bmod k$

$$\cos\psi = \cos\frac{j\pi}{k} - \frac{\eta \sin^2 \frac{j\pi}{k}}{a(R) + \operatorname{sign}\eta \, \cos\frac{j\pi}{k}} \quad \textit{with} \quad |\eta| \leq \tan^2\frac{\pi}{2k} \quad , \quad \eta \in \mathbb{R} \quad , \tag{43}$$

then

$$p_k(z;a) = \frac{T_k(\dfrac{z+\eta}{1+|\eta|})}{T_k(\dfrac{a+\eta}{1+|\eta|})} \quad \textit{and} \quad E_k^{(r)}(a) = \frac{2}{\rho^k + 1/\rho^k} \tag{44}$$

with ρ defined by

$$\frac{1}{2}(\rho + \frac{1}{\rho}) = a(R) - \frac{|\eta|}{1+|\eta|} \; \frac{b(R)^2}{a(R) + \operatorname{sign}\eta \, \cos\frac{j\pi}{k}} \quad , \quad \rho > 1 \quad . \tag{45}$$

Proof. Writing $p \in \Pi_k(a)$ in the form $p(z) = 1-(z-a)(z-\bar{a})q(z)$, $q \in \Pi_{k-2}$, one recognizes (34) as a linear Chebyshev approximation problem, for which, since $a \notin [-1,1]$, Haar's condition is satisfied. Standard results from approximation theory (see e.g. [29]) guarantee that there always exists a unique optimal polynomial $p_k(z;a)$ for (34). Moreover, because of the symmetry of the problem with respect to the real axis, p_k is a real polynomial, and p_k is charcterized by assuming its maximum absolute value at at least k points in $[-1,1]$ with alternating signs. This alternation property implies that p_k has degree $k-1$ or k. First, consider the case $k-1$. Since the scalar multiples of T_{k-1} are the only polynomials of degree $k-1$ with an alternating set of length at least k, we conclude that $p_k(z;a) = T_{k-1}(z)/T_{k-1}(a)$, and, in view of $p_k \in \Pi_k(a)$, this case occurs iff $T_{k-1}(a) \in \mathbb{R}$ and $a \notin \mathbb{R}$. With (35), (36), one readily verifies that these are just the points $a = a(\psi)$ with $\psi = j\pi(k-1)$, $j \neq 0 \bmod k-1$. Now we turn to the case that p_k is of degree k. Since the optimal polynomials for the Zolotarev problem (39) are characterized by the same alternating property as p_k, it follows that p_k is of the form (41) with a suitable $\eta \in \mathbb{R}$. In order to guarantee $p_k \in \Pi_k(a)$, η must be the solution of (42).
Now, let $\eta \in \mathbb{R}$, $|\eta| \leq \tan^2 \frac{\pi}{2k}$. With (40) and (36), we conclude that a satisfies (42) iff

$$(\tilde{a} :=) \; \frac{a+\eta}{1+|\eta|} = \frac{1}{2}(\rho + \frac{1}{\rho})\cos\frac{j\pi}{k} + \frac{i}{2}(\rho - \frac{1}{\rho})\sin\frac{j\pi}{k} \tag{46}$$

for some $\rho > 1$ and some integer $j \neq 0 \bmod k$. By using the representation (35) of a and by equating the real (resp. imaginary) parts of (46), one arrives at two real nonlinear equations for the unknowns $\cos\psi$ and ρ, and a straightforward, but lenghty calculation shows that the solutions are given by (43) and (45). Finally, note that the first identity in (44) is a consequence of (41) and (40); the second one follows from $E_k^{(r)}(a) = 1/|T_k(\tilde{a})|$ and (46). ∎

For general a, (41) and (42) lead to rather complicated and not very useful formulae for $E_k^{(r)}(a)$ in terms of elliptic integrals. Next, we derive simple bounds for this quantity.

Theorem 3. *Let $R > 1$ and $k = 2, 3, \ldots$. Then, for $a = a(\psi) \in \mathcal{E}_R$*

$$\frac{2}{R^k + 1/R^k} \;\leq\; E_k^{(r)}(a) \;\leq\; 2\frac{b_{k-1}(R)|f_{k-1}(\psi)| + b_k(R)|f_k(\psi)|}{b_{2k-1}(R) + b_1(R) f_{2k-1}(\psi)} \quad \Big(=: B_k^{(r)}(a)\Big) \tag{47}$$

where

$$b_j(R) = \frac{1}{2}(R^j - \frac{1}{R^j}) \quad , \quad f_j(\psi) = \begin{cases} \sin(j\psi)/\sin\psi & \textit{if } \sin\psi \neq 0 \\ (-1)^{(j-1)l} j & \textit{if } \psi = l\pi \end{cases} \quad .$$

Both bounds in (47) are attained if $\psi = j\pi/k$, $j \neq 0 \bmod k$. In addition, the upper estimate is sharp for $\psi = j\pi/(k-1)$, $j \neq 0 \bmod k-1$.

Proof. Duffin and Schaeffer [9] showed that for any real polynomial of degree at most k, $|p(z)| \leq M$ on $[-1,1]$ implies $|p(a)| \leq M(R^k + 1/R^k)/2$ for all $a \in \mathcal{E}_R$. Application of this result to $p_k(z;a)$ yields the lower bound in (47). In order to obtain the upper bound, we consider polynomials $p(z) = \gamma T_k(z) + \delta T_{k-1}(z) \in \Pi_k(a)$ with $\gamma, \delta \in \mathbb{R}$. With (35) and (36), one readily verifies that $p \in \Pi_k(a)$ iff γ and δ satisfy

$$\begin{pmatrix} (R^k + 1/R^k)\cos(k\psi) & (R^{k-1} + 1/R^{k-1})\cos(k-1)\psi \\ (R^k - 1/R^k) f_k(\psi) & (R^{k-1} - 1/R^{k-1}) f_{k-1}(\psi) \end{pmatrix} \begin{pmatrix} \gamma \\ \delta \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix} \quad .$$

A routine calculation shows that this linear system has a unique solution and that

$$\max_{z \in [-1,1]} |p(z)| = |\gamma| + |\delta| = B_k^{(r)}(a) \quad .$$

Finally, the statements on the sharpness of (47) follow from Theorem 2. ∎

Note that the bounds in (47) are asymptotically optimal, and we have the following

Corollary 1. *Let $R > 1$ and $a =\in \mathcal{E}_R$. Then,*

$$\lim_{k \to \infty} (E_k^{(r)}(a))^{1/k} = \lim_{k \to \infty} (B_k^{(r)}(a))^{1/k} = \frac{1}{R} \quad .$$

The typical behavior of the optimal values of (32) and (34) and the bounds stated in Theorems 1 and 3 is illustrated in Fig. 5.1. For fixed $R = 1.103\ldots$ and $k = 30$, the four curves

$$E_k(a) \leq \frac{2}{R^k + 1/R^k} \leq E_k^{(r)}(a) \leq B_k^{(r)}(a) \quad , \quad a = a(\psi) \in \mathcal{E}_R \quad , \quad 0 \leq \psi \leq \pi/2$$

are plotted. Note that $E_k(a) = E_k(\bar{a}) = E_k(-a)$ (and analogous for $E_k^{(r)}(a)$), and hence it suffices to consider only the points a in the first quadrant.

Fig. 5.1

The follwing theorem summarizes our results on error bounds for the MR and ME methods. For the special case of matrices $A = T + i\sigma I$ with positive definite Hermitian part T, we also derive an error bound for the GAL method.

Theorem 4. *Let $\alpha \leq \lambda_{\min}(T)$ and $\beta \geq \lambda_{\max}(T)$ be given, and assume that $0 < \alpha < \beta$ if $\sigma = 0$. Let a be given by (33), and let R be the unique solution of*

$$\frac{1}{2}(R + \frac{1}{R}) = \frac{\sqrt{\beta^2 + \sigma^2} + \sqrt{\alpha^2 + \sigma^2}}{\beta - \alpha} \quad , \quad R > 1 \quad . \tag{48}$$

Then, for $k = 1, \ldots$:
a)

$$\frac{\|b - Ax_k^{MR}\|}{\|b - Ax_0\|} \;\leq\; E_k(a) \;\leq\; \frac{2}{R^k + 1/R^k} \; . \tag{49}$$

b)

$$\frac{\|x_\star - x_k^{ME}\|}{\|x_\star - x_0\|} \;\leq\; E_{k+1}^{(r)}(a) \;\leq\; B_{k+1}^{(r)}(a) \; . \tag{50}$$

c) If T is positive definite, then

$$\frac{\|x_\star - x_k^{GAL}\|_T}{\|x_\star - x_0\|_T} \;\leq\; \sqrt{1 + \frac{\sigma^2(\sqrt{\kappa} - \frac{1}{\sqrt{\kappa}})^2}{4\sigma^2 + \alpha^2(\sqrt{\kappa} + \frac{1}{\sqrt{\kappa}})^2}} \;\; \frac{2}{R^k + 1/R^k} \tag{51}$$

where $\kappa = \beta/\alpha$.

Proof of part c). We set $e_k = x_\star - x_k$ and $\mu_j = (T^j e_k, e_k)$, $j = -1, 0, 1$. With (6) and since $r_k = Ae_k$, one obtains

$$(r_k, e_k) = \mu_1 + i\sigma\mu_0 \quad \text{and} \quad \|r_k\|_{T^{-1}}^2 = \mu_1 + \sigma^2\mu_{-1} \quad . \tag{52}$$

Now let $u \in x_0 + K_k$ be arbitrary. By (8), $(r_k, u - x_k) = 0$, and therefore

$$(r_k, e_k) = (r_k, x_\star - u) = (T^{-1/2} r_k, T^{1/2}(x_\star - u)) \quad . \tag{53}$$

By application of the Cauchy-Schwartz inequality to (53) and with (52), we arrive at

$$\mu_1^2 + \sigma^2\mu_0^2 \;\leq\; \|x_\star - u\|_T^2 \, (\mu_1 + \sigma^2\mu_{-1}) \quad . \tag{54}$$

Next, recall that, by the Kantorovich inequality (e.g. [26, p. 83]),

$$s^2\mu_1\mu_{-1} \;\leq\; \mu_0^2 \quad \text{where} \quad s := \left(\frac{1}{2}(\sqrt{\kappa} + \frac{1}{\sqrt{\kappa}})\right)^{-1} \quad . \tag{55}$$

Using (55) and the estimate $\mu_1/\mu_{-1} \geq \lambda_{\min}(T^2) = \alpha^2$, we obtain from (54)

$$\mu_1 \leq \|x_\star - u\|_T^2 \, \frac{1 + \sigma^2\mu_{-1}/\mu_1}{1 + \sigma^2 s^2 \mu_{-1}/\mu_1} \leq \|x_\star - u\|_T^2 \, \frac{\alpha^2 + \sigma^2}{\alpha^2 + \sigma^2 s^2} \quad . \tag{56}$$

Since $u \in x_0 + K_k$ is arbitrary, $\|x_\star - u\|_T$ in (56) can be replaced by

$$\min_{u \in x_0 + K_k} \|x_\star - u\|_T \;=\; \min_{p \in \Pi_k : p(0) = 1} \|p(A)e_0\|_T \quad . \tag{57}$$

By expanding e_0 into orthonormal eigenvectors of T (cf. (15)) and with (31), (32), (33), and (37), we obtain

$$\min_{p \in \Pi_k : p(0) = 1} \|p(A)e_0\|_T \;\leq\; \|e_0\|_T \, E_k(a) \;\leq\; \|e_0\|_T \, \frac{2}{R^k + 1/R^k} \quad . \tag{58}$$

Finally, combining (56) – (58) yields the desired bound (51). ■

Remark 6. For the special case of $\sigma = 0$, (51) and (49) reduce to the usual error bounds (see e.g. [39]) for the classical conjugate gradient and conjugate residual algorithms.

Remark 7. We excluded Hermitian indefinite matrices $A = T$. Error bounds for this case can be found in Chandra [5] for the MR method and in [40,18,41] for the ME method.

Remark 8. For the GAL method there are no satisfactory error bounds for the general class of matrices (6).

6. Polynomial preconditioning

Polynomial preconditioning aims at speeding up the convergence of conjugate gradient type methods for the solution of $Ax = b$ by applying them to one of the two equivalent linear systems

$$s(A)Ax = s(A)b \tag{59}$$

(left preconditioning), or

$$s(A)Ay = b \quad , \quad x = s(A)y \tag{60}$$

(right preconditioning). Here s is a suitably chosen polynomial of small degree. For the case of Hermitian positive definite A, the idea goes back to Rutishauser [34] who proposed polynomial preconditioning in the fifties as a remedy for roundoff in the classical CG algorithm. The recent revival [25] of Rutishauser's method and the general interest in polynomial preconditioning is mainly motivated by the attractive features of this technique for vector and parallel computers (see [35] for a survey).

In this section, we study polynomial preconditioning for the class of matrices (6) $A = T + i\sigma I$. Let $l \geq 2$ be any fixed integer. We seek a polynomial $s \in \Pi_{l-1}$ with the following two properties:

(i) the coefficient matrix $s(A)A$ of (59) and (60) is again a shifted Hermitian matrix of the form (6) and

(ii) the convergence of conjugate gradient type methods, applied to the preconditioned systems (59) or (60), is speeded up optimally.

As in the previous section, let $\alpha, \beta \in \mathbb{R}$ be given such that

$$\alpha \leq \mu \leq \beta \quad \text{for all eigenvalues } \mu \text{ of } T \quad , \tag{61}$$

and assume that $0 < \alpha < \beta$ if $\sigma = 0$. Our criteria for optimal convergence in (ii) will be based on (61) as the only available information on the spectrum A and on the error bounds stated in Theorem 4.

First, consider the requirement (i). For any $s \in \Pi_{l-1}$, we can represent $s(A)A$ in the form

$$s(A)A = (T + i\sigma I)s(T + i\sigma I) = q(T) + i\tau I \tag{62}$$

with $q \in \Pi_l$ and $\tau \in \mathbb{R}$. Note that s, q, and τ is equivalent to

$$(\mu + i\sigma)s(\mu + i\sigma) \equiv q(\mu) + i\tau \quad \text{and} \quad \tau := iq(-i\sigma) \quad . \tag{63}$$

Since $q(T)$ is Hermitian iff q is a real polynomial, it follows from (62) that (i) is fulfilled iff $q \in \Pi_l^{(r)} := \{ \Pi_l \mid q \text{ has real coefficients } \}$ and $\tau \in \mathbb{R}$. Therefore, from now on, it is assumed that $s \in \Pi_{l-1}$ satisfies (63) with $q \in \Pi_l^{(r)}$ and $\tau \in \mathbb{R}$.

Next, we turn to the question of optimal choice of q and τ. A first, very tempting strategy is to require $\tau = 0$ and to choose q such that $s(A)A = q(T)$ is positive definite. The preconditioned system (59) can then be solved by the standard CG method. Clearly, $q(T) \approx I$ should approximate the identity matrix as best as possible. Using (61) and (63), we conclude that such an optimal q is given as the best approximation in

$$\min_{q\in\Pi_l^{(r)}:q(-i\sigma)=0} \quad \max_{\mu\in[\alpha,\beta]} |1-q(\mu)| \quad . \tag{64}$$

For positive definite matrices $A = T$, this approach just leads to Rutishauser's method [34]. For the non-Hermitian case $\sigma \neq 0$, (64) turns out to be equivalent to the approximation problem (34), and we have the following

Theorem 5. *Let $\sigma \neq 0$ and $l \geq 2$. Then, there exists a unique best approximation in (64) given by*

$$q^\star(\mu) = 1 - p_l(\frac{\beta+\alpha-2\mu}{\beta-\alpha}; a) \quad , \quad a = \frac{\beta+\alpha+2i\sigma}{\beta-\alpha} \quad , \tag{65}$$

where $p_l(z;a)$ is the extremal polynomial of (34) (for $k = l$) with optimal value $E_l^{(r)}(a)$ (cf. Theorem 2). Moreover, the matrix $s(A)A = q(T)$ is positive definite with eigenvalues in $[1 - E_l^{(r)}(a), 1 + E_l^{(r)}(a)]$, and for the iterates x_k of the CG method, applied to (59), the estimates

$$\frac{||x_\star - x_k||_{q(T)}}{||x_\star - x_0||_{q(T)}} \leq \frac{2}{\tilde{R}^k + 1/\tilde{R}^k} \quad , \quad k = 1,2,\ldots \quad , \quad \tilde{R} := \frac{1+\sqrt{1-(E_l^{(r)}(a))^2}}{E_l^{(r)}(a)} \quad , \tag{66}$$

hold.

Proof. The linear transformation $z(\mu) = (\beta + \alpha - 2\mu)/(\beta - \alpha)$ maps $[\alpha, \beta]$ onto $[-1, 1]$. Moreover, $p(z(\mu)) = 1 - q(\mu)$ defines a one-to-one correspondence between all $q \in \Pi_l^{(r)}$ with $q(-i\sigma) = 0$ and all real polynomials $p \in \Pi_l(a)$. This shows that (64) and (34) are equivalent (recall that the optimal polynomial for (34) is real), and, hence, $q^\star$ is indeed the unique best approximation in (64). The error bounds (66) follow from (51) and (48) (with $\sigma = 0$, $\alpha = 1 - E_l^{(r)}(a)$, and $\beta = 1 + E_l^{(r)}(a)$). ∎

Recall (see Fig. 5.1) that for fixed l of moderate size and fixed R, $E_l^{(r)}(a)$ strongly depends on the position of a on the ellipse $\mathcal{E}_R$. In particular, if a is close to the real points of the ellipse, $E_l^{(r)}(a)$ is significantly larger than for the other points of $\mathcal{E}_R$. Therefore, (66) suggests that the polynomial (65) will yield a poor preconditioner for matrices A which are nearly Hermitian positive definite. This will be confirmed by numerical results presented in the next section. Therefore, in order to obtain a polynomial preconditioner which is satisfactory for all $a \in \mathcal{E}_r$, it is crucial to treat τ in (62) as a free parameter, and, next, we determine optimal choices of p and τ for speeding up the MR and ME algorithms.

First, consider the MR method. Here right preconditioning (60) is the more natural choice between (59) and (60), since residual vectors for (60) are also residual vectors of the original linear system. Let y_k denote the kth iterate of the MR algorithm applied to $s(A)Ay = b$, and set $x_k^{PP} = s(A)y_k$. Moreover, let x_k be the kth approximation generated by the MR method applied to the original system $Ax = b$. Then, assuming that $x_0 = x_0^{PP}$, it follows with (60) that $K_k(s(A)r_0, s(A)A) \subset K_{kl}(r_0, A)$ and $x_k^{PP}, x_{kl} \in x_0 + K_{kl}(r_0, A)$. Hence, the minimization property (7) implies that $||b - Ax_{kl}|| \leq ||b - Ax_k^{PP}||$. Therefore, in view of (49), we conclude that, based on (61) as the only information on the spectrum of A, the best possible choice of $s \in \Pi_{l-1}$ is one which guarantees the estimates

$$\frac{||b - Ax_k^{PP}||}{||b - Ax_0||} \leq \frac{2}{R^{kl} + 1/R^{kl}} \quad , \quad k = 1, 2, \ldots \quad , \tag{67}$$

with R defined in (48). We call $s \in \Pi_{l-1}$ an *optimal* polynomial preconditioner for the MR algorithm if it leads to the error bounds (67).

Similarly, for the ME method with left polynomial preconditioning (59), the error bounds (50) and Corollary 1 suggest that the best possible choice of $s \in \Pi_{l-1}$ is one for which the iterates x_k^{PP} satisfy

$$\frac{||x_* - x_k^{PP}||}{||x_* - x_0||} \leq E_{k+1}^{(\tau)}(\tilde{a}) \quad , \quad k = 1, 2, \ldots \quad , \tag{68}$$

for some $\tilde{a} \in \mathcal{E}_{R^l}$. A polynomial $s \in \Pi_{l-1}$ is called an *optimal* preconditioner for the ME approach if it guarantees (68).

With this notion of optimality, we can now state the main result of this section as follows.

Theorem 6. *Let $l \geq 2$. Then,*

$$s_{l-1}(\lambda) = \frac{q_l(\lambda - i\sigma) + i\tau}{\lambda} \tag{69}$$

where

$$q_l(\mu) = T_l(\frac{2\mu - \beta - \alpha}{\beta - \alpha}) - \operatorname{Re} T_l(-a) \quad \textit{and} \quad \tau = -\operatorname{Im} T_l(-a) \quad , \tag{70}$$

is an optimal polynomial preconditioner for the MR and ME methods. Here, T_l denotes the lth Chebyshev polynomial (cf. (36)) and a is given in (65).

Proof. First, note that, by (70), $q_l(-i\sigma) = -i\tau$, and thus (69) defines indeed a polynomial $s \in \Pi_{l-1}$. Next, consider the preconditioned matrix $\tilde{A} = s(A)A$. With (61) and since T_l maps the interval $[-1, 1]$ onto itself, it follows that the eigenvalues of the Hermitian part $q_l(T)$ of $\tilde{A}$ are contained in $[\tilde{\alpha}, \tilde{\beta}]$ where $\tilde{\alpha} := -1 - \operatorname{Re} T_l(-a)$ and $\tilde{\beta} := 1 - \operatorname{Re} T_l(-a)$. Now we apply Theorem 4 (with $\alpha = \tilde{\alpha}$, $\beta = \tilde{\beta}$, and $\sigma = \tau$) and note that, by (35) and (36),

$$\tilde{a} = \frac{\tilde{\beta} + \tilde{\alpha} + 2i\tau}{\tilde{\beta} - \tilde{\alpha}} = -T_l(-a) \in \mathcal{E}_{R^l} \quad .$$

The error bounds (67) and (68) are then an immediate consequence of parts a) and b), respectively, of Theorem 4. Hence s_{l-1} is an optimal polynomial preconditioner, and the proof is complete. ■

Remark 9. In [10], Eiermann, Li, and Varga developed a general theory for polynomial preconditioning for asymptotically optimal semi-iterative methods. In particular, by means of Theorem 6 from [10], one can show that the polynomial preconditioner (69) is also best possible for semi-iterative procedures for the class of matrices (6).

Remark 10. For the GAL approach, there are in general no error bounds on which we could base the choice of a best possible polynomial s. However, in analogy to the case of real symmetric matrices (see [42,40,41]), preconditioning for the GAL method can be motivated by its close connection (cf. (27)) to the ME algorithm. Therefore, we regard (69) also as an optimal polynomial preconditioner for the GAL method.

Finally, note that polynomial preconditioning is very easily incorporated into the MR and ME/GAL Algorithms 2 and 3. Right preconditioning leads to slightly more economical implementations, and only this choice is considered in the sequel. The idea is to apply the CG type methods to the linear system $s_{l-1}(A)Ay = b - Ax_0$ with starting guess $y_0 = 0$. The resulting iterates y_k of the MR and ME/GAL approaches are generated by Algorithm 2 and 3, respectively, modified in the following way: substitute y_k for x_k, replace in (24) σ by τ (defined in (70)), and, finally, perform in 2) the following Lanczos step

$$v = z^{(k)} - \tilde{\alpha}_k v_k - \beta_k v_{k-1} \quad \text{where} \quad z^{(k)} := T_l\Big(\frac{2}{\beta-\alpha}T - \frac{\beta+\alpha}{\beta-\alpha}I\Big)v_k \,, \ \tilde{\alpha}_k := (z^{(k)}, v_k) \,, \qquad (71)$$

and set $\alpha_k = \tilde{\alpha}_k - \mathrm{Re}\, T_l(-a)$. We remark that for this computation only T_l, but never the complex polynomial (69), is used. The actual preconditioner s_{l-1} appears only in the translation of the y_k into the corresponding iterates

$$x_k = x_0 + s_{l-1}(A)y_k \qquad (72)$$

for the original system $Ax = b$. However, we do not need to generate x_k in each step. Note that the norm $||r_k||$ of the residual $r_k = b - Ax_k$ is available (cf. Section 3) from the procedure generating y_k, and the iteration is stopped as soon as $||r_k||$ is sufficiently reduced. Hence, x_k is computed only once, namely in the very last step of the algorithm.

Finally, notice that $z^{(k)}$ in (71) can be obtained by performing l steps of the classical Chebyshev semi-iterative method (see Golub and Varga [22]). More precisely, setting

$$z_j^{(k)} := T_j\Big(\frac{2}{\beta-\alpha}T - \frac{\beta+\alpha}{\beta-\alpha}I\Big)v_k \quad , \quad T' := T - \frac{\beta+\alpha}{2}I \quad , \quad \omega = \frac{2}{\beta-\alpha} \quad , \qquad (73)$$

the three-term recurrence formula of the Chebyshev polynomials leads to the following

Algorithm 5 (Computation of $z^{(k)}$ in (71)).

```
0) Set $z_0^{(k)} = v_k$ and $z_1^{(k)} = \omega\ T'v_k$.
1) For $j = 2, \dots, l$, compute
   $z_j^{(k)} = 2\omega T' z_{j-1}^{(k)} - z_{j-2}^{(k)}$.
2) Set $z^{(k)} = z_l^{(k)}$.
```

Remark 11. The computation of $z^{(k)}$ via Algorithm 5 requires $2l$ matrix-vector products $T \cdot v$, $v \in \mathbb{R}^n$, and $2l$ additional real multiplications. If T and r_0 are real (cf. Section 4), all $z_j^{(k)}$ are real too, and the work is halved.

Similarly, using (69), (70), and again the three-term recurrence formula of the Chebyshev polynomials, a routine calculation shows that the following algorithm just yields the iterate (72).

Algorithm 6 (Computation of x_k in (72)).

```
0) Set $h_o^{(k)} = y_k$ and $h_1^{(k)} = 2\omega(T'y_k - (\frac{\beta+\alpha}{2} + i\sigma)y_k)$.
1) For $j = 2, \dots, l-1$, compute
   $h_j^{(k)} = 2\omega T' h_{j-1}^{(k)} - h_{j-2}^{(k)} + 2T_j(-a)y_k$.
2) Set $x_k = x_0 + \omega h_{l-1}^{(k)}$.
```

7. Numerical experiments

We have performed numerical experiments with all algorithms considered in this paper in numerous cases. Mostly, linear systems arising from the complex Helmholtz equation (5) were used as test problems. In this section, we present a few typical results of these experiments.

Consider (5) on the unit square $(0,1) \times (0,1)$ with Dirichlet boundary conditions, and assume that $\sigma_1, \sigma_2 \in \mathbb{R}$ are constant coefficients. Then, approximating (5) by finite differences on a uniform $m \times m$ grid with mesh size $h = 1/(m+1)$ leads to a linear system with coefficient matrix

$$A = T + i\sigma I \quad , \quad T := A_0 - \sigma_1 h^2 I \quad , \quad \sigma := \sigma_2 h^2 \ . . \tag{74}$$

Here A_0 is the symmetric positive definite matrix arising from the usual five-point discretization of $-\Delta$. Note that the eigenvalues of A_0 are known, and for our experiments with polynomial preconditioning we have used the true values

$$\alpha = \lambda_{\min}(A_0) - \sigma_1 h^2 \quad , \quad \beta = \lambda_{\max}(A_0) - \sigma_1 h^2 \tag{75}$$

of the extreme eigenvalues of T (cf. (61)). For the constants in (74), values of the form $\sigma_1 = \sigma_1(\psi)$, $\sigma_2 = \sigma_2(\psi)$ were chosen. Here $0 \le \psi \le \pi/2$ is a parameter such that the points $a(\psi) = (\beta + \alpha + 2i\sigma)/(\beta - \alpha)$ all lie on the same ellipse $\mathcal{E}_R$, $R > 1$ fixed, with ψ

describing the position of $a(\psi)$ on $\mathcal{E}_R$ (see (33) and (35)). The case $\psi = 0$ corresponds to a symmetric positive definite matrix (74), and for our experiments, we have chosen $R > 1$ such that $A = A_0$ for $\psi = 0$. Moreover, notice that with increasing ψ, the symmetric part T of (74) becomes more and more indefinite and $\alpha = -\beta$ for $\psi = \pi/2$. Also, the shift σ increases with ψ. Finally, we remark that the error bounds of Theorem 4 suggest that the MR and ME methods should display similar convergence rates for all ψ.

For our numerical examples, the mesh size $h = 1/64$ was used resulting in matrices (74) of dimension $n = 3969$. All the computations were done on a Cray-2. The exact solution $x_\star$ was generated with random components in $[-1,1] + i[-1,1]$, and then the right-hand side was set to $b := Ax_\star$. As starting vector $x_0 = 0$ was chosen. As stopping criterion, we used

$$||b - Ax_k|| \leq 10^{-6} ||b - Ax_0|| \quad . \tag{76}$$

In the following tables, for several values of ψ (stated in degree!) and the various CG type methods, we list the number of iterations which were necessary to reach (76). A "$\star$" indicates that the process still had not converged after 200 steps. In Table 7.1 the results for the MR, ME, and GAL Algorithms 2 resp. 3 (without preconditioning) are given. The Tables 7.2, 7.3, and 7.4 display the behavior of the three methods combined with the polynomial preconditioner (69) with $l = 6, 11,$ and 16, respectively. Also listed are the results for the ZPCG method consisting of the classical CG algorithm with Zolotarev polynomial preconditioner (65) (see Theorem 5).

Table 7.1

Table 7.2

Table 7.3

Table 7.4

From these results, we draw the following conclusions. If used without preconditioning, the MR method appears to be superior to the ME and GAL approaches. However, note that the stopping criterion (76) is based on the norm of the residual, and this is more favorable for the MR method. A comparison based on the Euclidian norm of the error vector $x_\star - x_k$ displays a similar convergence behavior for the ME and MR approaches. In combination with polynomial preconditioning, the performance of all three methods PPMR, PPME, and PPGAL is nearly identical. Also, note that the polynomial (69) yields a very efficient preconditioner which reduces the number of iterations significantly in all examples. Finally, as already suspected in the previous section, the strategy leading to the ZPCG method is a very dangerous one, and the algorithm even fails to converge if A is close to a positive definite matrix.

We conclude this section with two further remarks. First, note that all the results for the PPMR, PPME, and PPGAL metods were obtained with right polynomial preconditioning (RPP) (cf. (60)). Experiments with left polynomial preconditioning (LPP) (see (59)) gave nearly identical results. However, since implementations of RPP are slightly more economical, we therefore recommend RPP over LPP. Finally, recall that for our tests, the true extreme eigenvalues (75) of T were used. Of course, in general, such information is not available. However, it is possible to obtain good estimates of these quantities after relatively few steps of the Lanczos Algorithm 1.

Acknowledgement. Part of this work was done while the author was visiting the Computer Science Department of Stanford University. I would like to thank Gene Golub for his warm hospitality. I am also indebted to him for bringing the complex Helmholtz equation to my attention.

References

[1] Ashby, S.F., Manteuffel, T.A., Saylor, P.E.: A taxonomy for conjugate gradient methods. Preprint UCRL-98508, Lawrence Livermore National Laboratory, March 1988

[2] Barbour, I.M., Behilil, N.-E., Gibbs, P.E., Rafiq, M., Moriarty, K.J.M., Schierholz, G.: Updating fermions with the Lanczos method. *J. Comput. Phys.* **68**, 227-236 (1987)

[3] Bayliss, A., Goldstein, C.I., Turkel, E.: The numerical solution of the Helmholtz equation for wave propagation problems in underwater acoustics. *Comp. and Maths. with Appls.* **11**, 655-665 (1985)

[4] Carlson, B.C., Todd, J.: Zolotarev's first problem – the best approximation by polynomials of degree $\leq n-2$ to $x^n - n\sigma x^{n-1}$ in $[-1,1]$. *Aequationes Math.* **26**, 1-33 (1983)

[5] Chandra, R.: Conjugate gradient methods for partial differential equations. Ph. D. Thesis, Computer Science Department, Research Report 129 Yale University, January 1978

[6] Concus, P., Golub, G.H.: A generalized conjugate gradient method for nonsymmetric systems of linear equations. In: Computing methods in applied sciences and engineering (R. Glowinski and J.L. Lions, eds.), pp. 56-65. Lecture Notes in Economics and Mathematical Systems 134. Berlin, Heidelberg, New York: Springer 1976

[7] Concus, P., Golub, G.H., O'Leary, D.P.: A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations. In: Sparse matrix computations (J.R. Bunch and D.J. Rose, eds.), pp. 309-332. New York: Academic Press 1976

[8] Delfour, M., Fortin, M., Payre, G.: Finite-difference solutions of a non-linear Schrödinger equation. *J. Comput. Phys.* **44**, 277-288 (1981)

[9] Duffin, R., Schaeffer, A.C.: Some properties of functions of exponential type. *Bull. Amer. Math. Soc.* **44**, 236-240 (1938)

[10] Eiermann, M., Li, X., Varga, R.S.: On hybrid semiiterative methods. *SIAM J. Numer. Anal.* **26**, 152-168 (1989)

[11] Eisenstat, S.C., Elman, H.C., Schultz, M.H.: Variational iterative methods for nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.* **20**, 345-357 (1983)

[12] Elmore, W.C., Heald, M.A.: Physics of waves. New York: McGraw-Hill 1969

[13] Faber, V., Manteuffel, T.: Necessary and sufficient conditions for the existence of a conjugate gradient method. *SIAM J. Numer. Anal.* **21**, 352-362 (1984)

[14] Faber, V., Manteuffel, T.: Orthogonal error methods. *SIAM J. Numer. Anal.* **24**, 170-187 (1987)

[15] Fischer, B., Freund, R.: On the constrained Chebyshev approximation problem on ellipses. *J. Approx. Theory*, to appear

[16] Fletcher, R.: Conjugate gradient methods for indefinite systems. In: Numerical Analysis Dundee 1975 (G.A. Watson, ed.), pp. 73-89. Lecture Notes in Mathematics 506. Berlin, Heidelberg, New York: Springer 1976

[17] Fridman, V.M.: The method of minimum iterations with minimum errors for a system of linear algebraic equations with a symmetrical matrix. *USSR Comput. Math. and Math. Phys.* **2**, 362-363 (1963)

[18] Freund, R.: Über einige cg-ähnliche Verfahren zur Lösung linearer Gleichungssysteme. Doctoral Thesis, Universität Würzburg, F.R. of Germany, May 1983

[19] Freund, R., Ruscheweyh, St.: On a class of Chebyshev approximation problems which arise in connection with a conjugate gradient type method. *Numer. Math.* **48**, 525-542 (1986)

[20] Goldstein, C.I.: Multigrid preconditioners applied to three-dimensional parabolic equation type models. In: Computational acoustics: wave propagation (D. Lee, R.L. Sternberg, M.H. Schultz, eds.), pp. 57-74. Amsterdam: North-Holland 1988

[21] Golub, G.H., Van Loan, C.F.: Matrix computations. Baltimore: The Johns Hopkins University Press 1983

[22] Golub, G.H., Varga, R.S.: Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods. *Numer. Math.* **3**, 147-168 (1961)

[23] Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards* **49**, 409-436 (1952)

[24] Householder, A.S.: The theory of matrices in numerical analysis. New York: Blaisdell 1964

[25] Johnson, O.G., Micchelli, C.A., Paul, G.: Polynomial preconditioners for conjugate gradient calculations. *SIAM J. Numer. Anal.* **20**, 362-376 (1983)

[26] Joubert, W.D., Young, D.M.: Necessary and sufficient conditions for the simplification of generalized conjugate-gradient algorithms. *Lin. Alg. Appl.* **88/89**, 449-485 (1987)

[27] Lanczos, C.: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Standards* **45**, 255-282 (1950)

[28] Marfurt, K.J.: Finite element modeling of elastodynamic and electromagnetic wave propagation for geophysical exploration. In: Computing methods in applied sciences and engineering VII (R. Glowinski and J.L. Lions, eds.), pp. 517-547. Amsterdam: North Holland 1986

[29] Meinardus, G.: Approximation of functions: Theory and numerical methods. Berlin, Heidelberg, New York: Springer 1967

[30] Paige, C.C., Saunders, M.A.: Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.* **12**, 617-629 (1975)

[31] Peterson, A.F., Mittra, R.: Method of conjugate gradients for the numerical solution of large-body electromagnetic scattering problems. *J. Opt. Soc. Am. A* **2**, 971-977 (1985)

[32] Pierce, A.D.: Acoustics: An introduction to its physical principles and applications. New York: McGraw-Hill 1981

[33] Rapoport, D.: A nonlinear Lanczos algorithm and the stationary Navier-Stokes equation. Ph. D. Thesis, Department of Mathematics, New York University, October 1978

[34] Rutishauser, H.: Theory of gradient methods. In: Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary value problems, pp. 24-49. Mitteilungen aus dem Institut für Angewandte Mathematik an der ETH Zürich **8**. Basel: Birkhäuser 1959

[35] Saad, Y.: Krylov subspace methods on supercomputers. *SIAM J. Sci. Stat. Comput.*, to appear

[36] Saad, Y., Schultz, M.H.: Conjugate gradient-like algorithms for solving nonsymmetric linear systems. *Math. Comp.* **44**, 417-424 (1985)

[37] Schultz, M.H., Lee, D., Jackson, K.R.: Application of the Yale sparse technique to solve the 3-dimensional parabolic wave equation. In: Recent progress in the development and application of the parabolic equation (P.D. Scully-Power and D. Lee, eds.). Naval Underwater Systems Center, Technical Document 7145, May 1984

[38] Sidi, A.: Extrapolation vs. projection methods for linear systems of equations. *J. Comput. Appl. Math.* **22**, 71-88 (1988)

[39] Stoer, J.: Solution of large linear systems of equations by conjugate gradient type methods. In: Mathematical programming – The state of the art (A. Bachem, M. Grötschel, and B. Korte, eds.), pp. 540-565. Berlin, Heidelberg, New York, Tokyo: Springer 1983

[40] Stoer, J., Freund, R.: On the solution of large indefinite systems of linear equations by conjugate gradient algorithms. In: Computing methods in applied sciences and engineering V (R. Glowinski and J.L. Lions, eds.), pp. 35-53. Amsterdam: North Holland 1982

[41] Szyld, D.B.: A two-level iterative method for large sparse generalized eigenvalue calculations. Ph. D. Thesis, Department of Mathematics, New York University, October 1983

[42] Szyld, D.B., Widlund, O.: Applications of conjugate gradient type methods to eigenvalue calculations. In: Advances in computer methods for partial differential equations III (R. Vichnevetsky and R.S. Stepleman, eds.), pp. 167-173. New Brunswick: IMACS 1979

[43] Trummer, M.: An efficient implementation of a conformal mapping method based on the Szegö kernel. *SIAM J. Numer. Anal.* **23**, 853-872 (1986)

[44] Widlund, O.: A Lanczos method for a class of nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.* **15**, 801-812 (1978)

	T· v	m	dp	s
MR Algorithm 2	2	18n	4	12
ME/GAL Algorithm 3	2	18n	4	10
Orthodir Algorithm 4	2	26n	8	14
If T and r_0 are real:				
MR Algorithm 2	1	9n	2	7
ME/GAL Algorithm 3	1	13n	2	7
If $A = T$ and r_0 are real:				
MINRES [30]	1	8n	2	6
SYMMLQ [30]	1	8n	2	5

Table 4.1

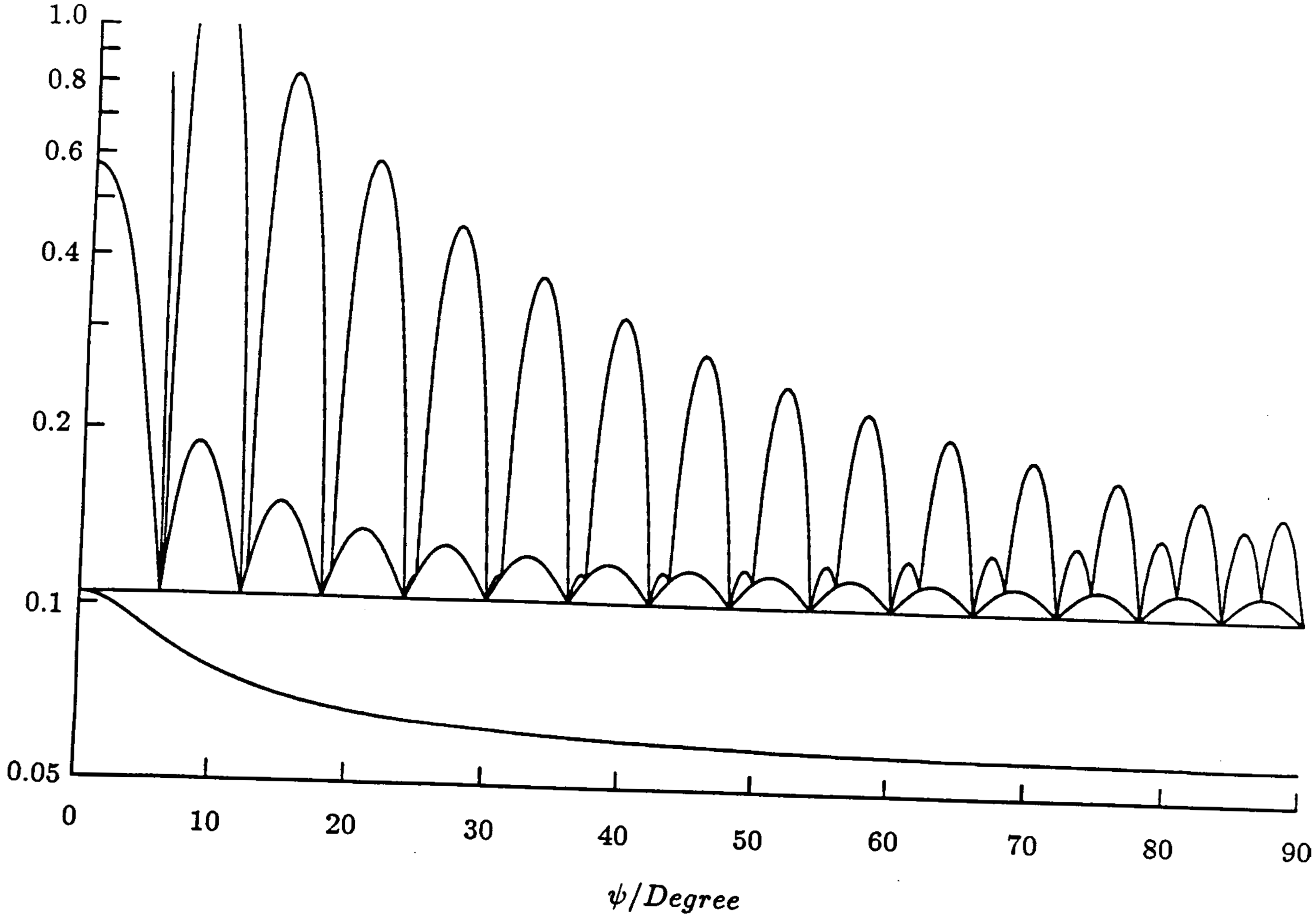

Fig. 5.1

$\psi/Degree$	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90
MR	120	126	148	165	175	183	190	197	203	208	212	217	221	224	228	232	234	237	239
ME	183	177	166	186	191	210	210	215	224	231	236	237	244	245	250	252	259	260	263
GAL	129	144	165	182	198	208	213	222	225	231	236	240	244	248	253	255	259	261	264

Table 7.1

$\psi/Degree$	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90
PPMR	47	47	47	47	47	47	48	47	47	47	47	47	47	47	47	47	47	47	47
PPME	63	47	47	47	47	47	64	47	47	47	47	47	63	47	47	47	47	47	63
PPGAL	49	49	49	49	50	50	50	50	50	49	49	49	49	49	49	49	50	50	50
ZPCG	$\star$	$\star$	148	99	74	59	49	56	62	63	59	53	48	53	57	58	56	52	49

Table 7.2 , l=6

$\psi/Degree$	0	5	10	15	20	25	30	35	40	45
PPMR	26	26	26	26	26	26	26	26	26	26
PPME	33	26	27	29	26	26	28	27	26	27
PPGAL	28	28	28	28	28	28	28	28	28	28
ZPCG	⋆	87	44	29	32	34	29	29	31	29

$\psi/Degree$	50	55	60	65	70	75	80	85	90
PPMR	26	26	26	26	26	26	26	26	26
PPME	30	27	26	30	27	26	26	28	27
PPGAL	28	28	28	28	28	28	28	28	28
ZPCG	27	30	29	27	29	29	27	28	29

Table 7.3 , l=11

$\psi/Degree$	0	5	10	15	20	25	30	35	40	45
PPMR	18	18	18	18	18	18	18	18	18	18
PPME	23	19	18	18	17	17	18	18	17	23
PPGAL	20	20	19	19	20	20	19	19	19	20
ZPCG	146	41	21	23	21	20	21	19	20	19

$\psi/Degree$	50	55	60	65	70	75	80	85	90
PPMR	18	18	18	18	18	18	18	18	18
PPME	17	18	18	17	17	17	17	17	23
PPGAL	19	19	19	19	19	19	19	19	20
ZPCG	20	19	20	19	19	20	19	20	19

Table 7.4 , l=16