Appeared in: WSEAS Transactions on Mathematics, Volume 2, Issues 1 and 2, 2003, pp. 142-150.

A Zero-Cost Preconditioning for a Class of Indefinite Linear Systems

AVRAM SIDI
Computer Science Department
Technion - Israel Institute of Technology
Haifa 32000
ISRAEL
E-mail: asidi@cs.technion.ac.il http://www.cs.technion.ac.il/~asidi/

Abstract: - We consider the solution by Krylov subspace methods of a certain class of hermitian indefinite linear systems, such as those that arise from discretization of the Stokes equations in incompressible fluid mechanics. We discuss a diagonal preconditioning of these systems that amounts to multiplying some of the equations by -1 while the others are left unchanged. We show that this preconditioning puts all the eigenvalues of the relevant matrix in the open right half plane, enhancing the performance of the Krylov subspace methods in many cases.

Key-Words: - Krylov subspace methods, preconditioning, indefinite linear systems

1 Introduction

In this paper, we consider the numerical solution by Krylov subspace and semi-iterative methods of large sparse linear systems of equations of the form

$$\begin{bmatrix} G & H \\ H^* & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{1}$$

where $G \in \mathbb{C}^{m\times m}$, $H \in \mathbb{C}^{m\times n}$, $m \geq n$, G is hermitian positive definite, and H is of full rank, $u, f \in \mathbb{C}^m$, and $v, g \in \mathbb{C}^n$. Here H^* denotes the hermitian conjugate of H. Obviously, the matrix of the system in (1) is hermitian.

Such linear systems arise in different computational problems. We mention three such problems:
1. Solution of linear least-squares problems of the form

$$\min \|Ax - b\|, \quad A \in \mathbb{C}^{m\times n},\ m \geq n,\ \operatorname{rank}(A) = n. \tag{2}$$

Here $\|y\| = \sqrt{y^*y}$ is the standard Euclidean vector norm. The solution x to this problem is the unique solution of the nonsingular $n \times n$ system $A^*Ax = A^*b$. Defining the residual vector r by $r = b - Ax$, it is easy to see that x and r together satisfy the linear $(m+n)\times(m+n)$ system

$$\begin{bmatrix} I_m & A \\ A^* & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}. \tag{3}$$

Here I_p stands for the $p \times p$ identity matrix.
2. Solution of linearly constrained quadratic programming problems of the form

$$\min (\tfrac{1}{2}x^TAx - r^Tx),\ \text{subject to}\ E^Tx = s, \quad A \in \mathbb{R}^{m\times m},\ E \in \mathbb{R}^{m\times n},\ m \geq n,\ \operatorname{rank}(E) = n. \tag{4}$$

Here A is a positive definite matrix. Introducing a vector $\lambda \in \mathbb{R}^n$ of Lagrange multipliers, it can be shown that the optimal solution x and the vector λ satisfy the linear system

$$\begin{bmatrix} A & E \\ E^T & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} r \\ s \end{bmatrix}. \tag{5}$$

3. Solution of linear systems that arise from suitable finite-difference or finite-element discretizations of the so-called Stokes equations in incompressible fluid mechanics. The Stokes equations, after suitable scaling of the dependent and independent variables and body forces, can be written in the form

$$\Delta\mathbf{v} - \nabla p + \mathbf{f} = 0,\ \ \nabla\cdot\mathbf{v} = 0, \tag{6}$$

where $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ is the Laplacian, $\nabla = \mathbf{i}\partial/\partial x + \mathbf{j}\partial/\partial y + \mathbf{k}\partial/\partial z$ is the gradient operator, $\mathbf{v} = \mathbf{i}v_x + \mathbf{j}v_y + \mathbf{k}v_z$ is the velocity vector of

the fluid, p is the pressure, and $\mathbf{f} = \mathbf{i}f_x + \mathbf{j}f_y + \mathbf{k}f_z$ is some known body force. At least in operator form we can see that these coupled partial differential equations can be expressed in the form of the linear system in (1) as follows:

$$\begin{bmatrix} -\Delta & 0 & 0 & \partial/\partial x \\ 0 & -\Delta & 0 & \partial/\partial y \\ 0 & 0 & -\Delta & \partial/\partial z \\ \partial/\partial x & \partial/\partial y & \partial/\partial z & 0 \end{bmatrix} \begin{bmatrix} v_x \\ v_y \\ v_z \\ p \end{bmatrix} = \begin{bmatrix} f_x \\ f_y \\ f_z \\ 0 \end{bmatrix}. \tag{7}$$

As will be discussed in the next section, the matrix of the linear system in (1) is indefinite, i.e., it has both positive and negative eigenvalues. This fact has been known in connection with the least-squares problem for the matrix of the system in (3).

These linear systems can be solved by suitable fixed-point iterative methods, semi-iterative methods, and Krylov subspace methods. For the development of a class of fixed-point iterative methods, see the paper by Dyn and Ferguson [2]. Three Krylov subspace methods that are relevant for the linear systems above are MINRES, SYMMLQ, and LSQR, due to Paige and Saunders [12], [13]. Of course, we can also apply, e.g., the restarted GMRES of Saad and Schultz [14] or Bi-CGSTAB of van der Vorst [20] or QMR of Freund and Nachtigal [6]. For an exhaustive list of references on this subject, we refer the reader to Björck [1, Chapter 7].

Due to the fact that the matrix of (1) is indefinite, Krylov subspace methods may converge slowly when applied directly to (1). To improve their convergence we may have to apply them with an effective preconditioner. However, the design of such preconditioners is not trivial and their application may entail a non-negligible computational cost. The situation is especially problematic when solving the Stokes equations on unstructured meshes. We thus ask whether we can improve the convergence of the different Krylov subspace methods without having to construct expensive preconditioners. Precisely this is the subject of the present work.

We start by noting that, when applied to a nonsingular linear system $Ax = b$, Krylov subspace methods can be especially effective if the spectrum of A is contained in the interior of a half plane whose boundary is a straight line that passes through the origin. Recall also that the Chebyshev acceleration method of Manteuffel [9] (that is a semi-iterative method) is defined for such matrices. In case A has both positive and negative eigenvalues, there is no such half plane, and Krylov subspace methods may be especially slow.

In this paper, we discuss a simple way of modifying the linear system in (1) that forces the spectrum of its matrix to the open right half of the complex plane. This modification was suggested by Nicolaides [11] already in 1991 in connection with the numerical solution of the Stokes equations by Krylov subspace methods. It amounts to a diagonal preconditioning of the matrix, which, as far as we know, has not been given before. Specifically, Krylov subspace methods are applied directly to the linear system

$$\begin{bmatrix} G & H \\ -H^* & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f \\ -g \end{bmatrix}, \tag{8}$$

The preconditioning matrix is nothing but

$$M = \begin{bmatrix} I_m & 0 \\ 0 & -I_n \end{bmatrix}. \tag{9}$$

Here I_p stands for the $p \times p$ identity matrix. Clearly, the system in (8) has the same solution as that in (1). We also note that, as the matrix of the new system is *not* hermitian, the methods MINRES, SYMMLQ, and LSQR that were mentioned above cannot be applied. Of course, GMRES, Bi-CGSTAB, and QMR are still applicable. It seems that before embarking on the construction of sophisticated preconditioners, one can try the application of Krylov subspace methods to the system in (8) that forms a zero-cost preconditioning of that in (1). This can prove to be very convenient, for example, when solving continuum problems on unstructured meshes.

In the next section, we analyze the spectrum of the matrix of the modified system in (8) and show that all its eigenvalues are in the right half of the complex plane, i.e., they all have positive real parts. This may help enhance the performance of Krylov subspace methods when these are applied to the modified system, and this is the subject of Section 3. In Section 4, we demonstrate the use of the approach discussed here with a numerical example.

For related work, we refer the reader to the recent papers by Wathen and Silvester [21], Silvester and Wathen [19], Golub and Wathen [8], Fischer et al. [5], Fischer and Peherstorfer [4], Murphy, Golub,

and Wathen [10], and Elman, Silvester, and Wathen [3]. In particular, [21] provides a diagonal preconditioner that is different from that presented here. In addition, our Theorem 1 and Corollary 2 are a slight generalization of Lemmas 2.1 and 2.2 of [5]. In the case $G \neq \gamma I_m$, where γ is a scalar, the preconditioner proposed in [5] involves the matrix G^{-1}, so that the preconditioner proposed in the present work is entirely different from that in [5].

The literature that deals with the solution of the linear systems of this work (and their more general forms) and, in particular, of the Stokes equations, is quite rich. For the most recent developments and extensive bibliographies, we refer the reader to the papers above.

2 Theoretical Preliminaries

Our approach is based on the study of the eigenvalue structure of the matrix

$$C(\alpha) = \begin{bmatrix} G & H \\ \alpha H^* & 0 \end{bmatrix}, \quad \alpha \neq 0 \text{ a real scalar}, \quad (10)$$

where, as before, $G \in \mathbb{C}^{m\times m}$, $H \in \mathbb{C}^{m\times n}$, $m \geq n$, G is hermitian positive definite, and H is of full rank.

We begin with the following result, which generalizes those of Lemmas 2.1 and 2.2 in [5]:

Theorem 1 *Let $A \in \mathbb{C}^{m\times n}$ such that $m \geq n$, and $\mathrm{rank}(A) = n$, and denote*

$$F(\alpha) = \begin{bmatrix} I_m & A \\ \alpha A^* & 0 \end{bmatrix}, \quad \alpha \neq 0 \text{ a real scalar}, \quad (11)$$

where I_p stands for the $p \times p$ identity matrix. Then the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{m+n}$ of $F(\alpha)$ are given by

$$\begin{aligned} \lambda_{2k-1} &= \frac{1}{2}\left(1 + \sqrt{1 + 4\alpha\sigma_k^2}\right), \quad k = 1, \ldots, n, \\ \lambda_{2k} &= \frac{1}{2}\left(1 - \sqrt{1 + 4\alpha\sigma_k^2}\right), \quad k = 1, \ldots, n, \\ \lambda_{2n+1} &= \lambda_{2n+2} = \cdots = \lambda_{m+n} = 1, \end{aligned} \quad (12)$$

where $\sigma_1, \ldots, \sigma_n$ are the singular values of A, which are all positive by the assumption that A has full rank n. Let us denote by η_k the eigenvector of A^*A corresponding to its eigenvalue σ_k^2, $k = 1, \ldots, n$. When $\alpha \neq -1/(4\sigma_k^2)$, we have $\lambda_{2k-1} \neq \lambda_{2k}$, and the eigenvectors x_{2k-1} and x_{2k} corresponding to λ_{2k-1} and λ_{2k}, respectively, are given by

$$x_j = \begin{bmatrix} (\alpha\sigma_k^2)^{-1}\lambda_j A\eta_k \\ \eta_k \end{bmatrix}, \quad j = 2k-1, 2k. \quad (13)$$

When $\alpha = -1/(4\sigma_k^2)$, hence $\lambda_{2k-1} = \lambda_{2k} = 1/2$, there is one eigenvector x_{2k-1} and one principal vector x_{2k} corresponding to $\lambda_{2k-1} = \lambda_{2k}$ that are given by

$$x_{2k-1} = \begin{bmatrix} -2A\eta_k \\ \eta_k \end{bmatrix}, \quad x_{2k} = \begin{bmatrix} 0 \\ -2\eta_k \end{bmatrix}. \quad (14)$$

The eigenvectors corresponding to the eigenvalue unity are given by

$$x_{2n+k} = \begin{bmatrix} \xi_k \\ 0 \end{bmatrix}, \quad k = 1, \ldots, m-n, \quad (15)$$

where ξ_k are the eigenvectors of AA^* corresponding to its zero eigenvalue that has multiplicity $m-n$ necessarily.

Proof. To prove (12) we start with the fact that

$$(F - I)F = \alpha \begin{bmatrix} AA^* & 0 \\ 0 & A^*A \end{bmatrix}, \quad F \equiv F(\alpha),$$

from which it follows that F has $2n$ eigenvalues λ that satisfy

$$(\lambda - 1)\lambda = \alpha\sigma_k^2, \quad k = 1, \ldots, n,$$

and $m - n$ eigenvalues that satisfy

$$(\lambda - 1)\lambda = 0.$$

The rest of the proof can be carried out by verifying the equalities

$$Fx_j = \lambda_j x_j, \quad j = 2k-1, 2k,$$

when $\alpha \neq -1/(4\sigma_k^2)$, $1 \leq k \leq n$, and

$$Fx_{2k-1} = \tfrac{1}{2}x_{2k-1}, \quad Fx_{2k} = \tfrac{1}{2}x_{2k} + x_{2k-1},$$

when $\alpha = -1/(4\sigma_k^2)$, $1 \leq k \leq n$, and

$$Fx_{2n+k} = x_{2n+k}, \quad k = 1, \ldots, m-n.$$

We leave the details to the reader. ■

What Theorem 1 implies is that the Jordan canonical form of $F(\alpha)$ is diagonal when $\alpha \neq -1/(4\sigma_k^2)$, $k = 1, \ldots, n$, and that it has 2×2 blocks with eigenvalue $1/2$ when $\alpha = -1/(4\sigma_k^2)$ for some $k \in \{1, \ldots, n\}$, the number of these blocks being the same as the number of the σ_k for which $\alpha = -1/(4\sigma_k^2)$. Note also that none of the λ_k vanishes, so that $F(\alpha)$ is nonsingular.

We also observe that the complex eigenvalues of $F(\alpha)$ appear in conjugate pairs whether H [hence $F(\alpha)$] is real or complex.

Corollary 2 *Another implication of Theorem 1 is that, with the exception of the eigenvalue unity, the rest of the spectrum of $F(\alpha)$, namely, $\{\lambda_1, \ldots, \lambda_{2n}\}$, is distributed symmetrically both with respect to the vertical line $\Re\lambda = 1/2$ and with respect to the real axis in the λ-plane. In particular, the following can be deduced from (12):*

(i) *When $\alpha > 0$, all the eigenvalues $\lambda_1, \ldots, \lambda_{2n}$ are real, half of them positive and the other half negative, such that*

$$\lambda_{2k} < 0, \;\; \lambda_{2k-1} > 1, \;\; k = 1, \ldots, n.$$

(ii) *When $-1/(4\sigma_{\max}^2) < \alpha < 0$, all of $\lambda_1, \ldots, \lambda_{2n}$ are real and positive, such that*

$$0 < \lambda_{2k} < 1/2 < \lambda_{2k-1} < 1, \;\; k = 1, \ldots, n.$$

(iii) *When $\alpha \leq -1/(4\sigma_{\max}^2)$, some or all of $\lambda_1, \ldots, \lambda_{2n}$ may be complex with real part $1/2$ and appearing in conjugate pairs, the rest being real positive and in $(0, 1)$.*

Here $\sigma_{\max} = \max\{\sigma_1, \ldots, \sigma_n\}$.

We now turn to the matrix $C(\alpha)$ in (10). Below, $\rho(G)$ and $\lambda_{\min}(G)$ stand, respectively, for the spectral radius and the smallest eigenvalue of G, and $\sigma_{\max}(H)$ stands for the largest singular value of H, as usual. Also, $\|\cdot\|$ stands both for the vector l_2-norm and the matrix norm induced by it. Thus, $\|G\| = \rho(G)$ and $\|H\| = \sigma_{\max}(G)$.

Theorem 3 *The matrix $C(\alpha)$ is nonsingular for all $\alpha \neq 0$. In addition, the following hold:*

(i) *For every $\alpha \neq 0$, $C(\alpha)$ has at most $m - n$ eigenvalues λ_k that are also in the spectrum of G (hence are positive) with corresponding eigenvectors $x_k = [\xi_k^{\mathrm{T}} \; 0^{\mathrm{T}}]^{\mathrm{T}}$, $\xi_k \in \mathbb{C}^m$ such that $G\xi_k = \lambda_k \xi_k$ and $H^*\xi_k = 0$. As for the remaining eigenvalues (at least $2n$ in number), we have the following: (1) When $\alpha > 0$, they all are real; some of them are positive and the rest are negative. (2) When $\alpha < 0$, they all are in the open right half of the complex plane; the real ones satisfy $0 < \lambda_k < \rho(G)$, while for the complex ones, we have $0 < \frac{1}{2}\lambda_{\min}(G) < \Re\lambda_k < \frac{1}{2}\rho(G)$. [These results are valid in parts (ii) and (iii) that follow.]*

(ii) *The matrix $C(1)$ is hermitian indefinite and has m positive and n negative eigenvalues. With $\Theta \equiv \max\{\rho(G), \sigma_{\max}(H)\}$, the positive eigenvalues satisfy*

$$0 < \lambda_k \leq \frac{1 + \sqrt{5}}{2}\,\Theta, \tag{16}$$

while the negative ones satisfy

$$\frac{1}{2}\big[\lambda_{\min}(G) - \sqrt{5}\,\Theta\big] \leq \lambda_k < 0. \tag{17}$$

(iii) *The matrix $C(-1)$ has all its eigenvalues in the open right half plane. When G and H are real, the spectrum of $C(-1)$ is distributed symmetrically with respect to the real axis. Furthermore, each eigenvalue λ_k of $C(-1)$ satisfies*

$$0 < \Re\lambda_k \leq \rho(G), \;\; |\Im\lambda_k| \leq \sigma_{\max}(H). \tag{18}$$

Proof. Letting $A \equiv G^{-1/2}H$, we start by observing that

$$C(\alpha) = KF(\alpha)K, \tag{19}$$

where

$$K = \begin{bmatrix} G^{1/2} & 0 \\ 0 & I_n \end{bmatrix}, \;\; F(\alpha) = \begin{bmatrix} I_m & A \\ \alpha A^* & 0 \end{bmatrix}. \tag{20}$$

Obviously, K is a hermitian positive definite matrix, and hence (19) is a congruence, namely, $C(\alpha) = KF(\alpha)K^*$. Furthermore, A has full rank, so that Theorem 1 applies to the matrix $F(\alpha)$ with $\alpha \neq 0$. Thus, since $F(\alpha)$ is nonsingular, so is $C(\alpha)$.

To prove part (i), we proceed as follows: Let (λ, x) be an eigenpair of $C(\alpha)$, i.e., $C(\alpha)x = \lambda x$.

Writing $x = [\xi^T \ \eta^T]^T$, where $\xi \in \mathbb{C}^m$, $\eta \in \mathbb{C}^n$, this gives

$$G\xi + H\eta = \lambda\xi, \ \ \alpha H^*\xi = \lambda\eta. \tag{21}$$

Now $\xi \neq 0$ must hold; otherwise, we would have from (21) that $H\eta = 0$, hence $\eta = 0$ since H has full rank, which would imply that $x = 0$.

Similarly, from (21), there may be eigenvectors with $\eta = 0$, provided $G\xi = \lambda\xi$ and $\xi \in \mathcal{N}(H^*)$, where $\mathcal{N}(B)$ denotes the null space of B. Actually, by the fact that $\lambda \neq 0$, and from the second of the equations in (21), $\eta = 0$ if and only if $H^*\xi = 0$. Also, because $H^* \in \mathbb{C}^{n\times m}$ and has full rank, the dimension of $\mathcal{N}(H^*)$ is exactly $m - n$. Thus, $C(\alpha)$ has at most $m - n$ eigenvectors of the form $x = [\xi^T \ 0^T]^T$ with corresponding eigenvalues λ such that $G\xi = \lambda\xi$, for which $H^*\xi = 0$ necessarily. For the remaining eigenvectors $x = [\xi^T \ \eta^T]^T$ (at least $2n$ in number), we have that $\eta \neq 0$ and hence $H^*\xi \neq 0$. We turn to these eigenvectors next.

Solving the second of the equations in (21) for η, and substituting in the first equation there, we obtain

$$G\xi + \frac{\alpha}{\lambda} H H^*\xi = \lambda\xi, \tag{22}$$

where we have used the fact that $\lambda \neq 0$ always. Multiplying both sides of this equality by ξ^*, and normalizing ξ such that $\xi^*\xi = \|\xi\|^2 = 1$ (which is possible by the fact that $\xi \neq 0$), we obtain the quadratic equation

$$\begin{gathered}\lambda^2 - p\lambda - \alpha q = 0, \\ p = \xi^*G\xi > 0, \ \ q = \xi^*HH^*\xi = \|H^*\xi\|^2 > 0.\end{gathered} \tag{23}$$

The solutions of this equation for λ are given by

$$\lambda_\pm = \frac{1}{2}\left(p \pm \sqrt{p^2 + 4\alpha q}\right). \tag{24}$$

There are two cases to consider:

1. $\alpha > 0$: In this case, $\lambda_\pm$ are all real and $\lambda_+ \geq p > 0$, while $\lambda_- < 0$. In short, when $\alpha > 0$, $C(\alpha)$ has only real positive and real negative eigenvalues.

2. $\alpha < 0$: In this case, there are two different situations to consider: (a) When $0 \leq p^2 + 4\alpha q < p^2$, we have that $\lambda_\pm$ are both real and positive with $0 < \lambda_- < \lambda_+ < p$. Thus, all such eigenvalues are real positive and satisfy $0 < \lambda_- < \rho(G)$. (b) When $p^2 + 4\alpha q < 0$, both of the eigenvalues $\lambda_\pm$ are complex, $\lambda_- = \overline{\lambda_+}$, and $\Re\lambda_\pm = \frac{1}{2}p > 0$. Thus, for all such eigenvalues $\frac{1}{2}\lambda_{\min}(G) < \Re\lambda_\pm < \frac{1}{2}\rho(G)$.

For the proof of part (ii), we proceed as follows: By the fact that $F(1)$ and $C(1)$ are both hermitian and by $C(1) = KF(1)K^*$, the Sylvester law of inertia (see, e.g., Golub and Van Loan [7]) applies, and $C(1)$ and $F(1)$ have the same number of positive and negative eigenvalues. Invoking now part (i) of Corollary 2, we conclude that $C(1)$ has m positive and n negative eigenvalues. To prove (16), we bound the expression for λ_+ given in (24) from above, and to prove (17), we bound the expression for λ_- there from below.

To prove part (iii), we consider $x^*C(-1)x = \lambda x^*x$, where (λ, x) is an eigenpair of $C(-1)$ as before, which gives

$$\xi^*G\xi + \xi^*H\eta - \eta^*H^*\xi = \lambda(\xi^*\xi + \eta^*\eta), \tag{25}$$

from which

$$\Re\lambda = \frac{\xi^*G\xi}{\xi^*\xi + \eta^*\eta}, \ \ \Im\lambda = \frac{2\Im(\xi^*H\eta)}{\xi^*\xi + \eta^*\eta}. \tag{26}$$

By the fact that $\xi \neq 0$, the first of the equalities in (26) implies $\Re\lambda > 0$. Again, by the fact that $\xi \neq 0$, it also follows from (26) that

$$\Re\lambda \leq \frac{\xi^*G\xi}{\xi^*\xi} \leq \|G\| \tag{27}$$

and

$$|\Im\lambda| \leq \frac{2|\xi^*H\eta|}{\|\xi\|^2 + \|\eta\|^2} \leq \frac{2\|H\|\,\|\xi\|\,\|\eta\|}{\|\xi\|^2 + \|\eta\|^2} \leq \|H\|. \tag{28}$$

Here we have made use of the fact that $2\|\xi\|\|\eta\| \leq \|\xi\|^2 + \|\eta\|^2$. The results in (18) now follow. ■

3 General Use of the Preconditioner

As the matrix of the system in (1), namely, $C(1)$, has both positive and negative eigenvalues by Theorem 3, the solution of (1) by Krylov subspace methods without preconditioning may not always be effective. The solution by Krylov subspace methods of the equivalent system

$$\begin{bmatrix} G & H \\ -H^* & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f \\ -g \end{bmatrix} \tag{29}$$

instead may be more effective as the matrix $C(-1)$ of this system has all its eigenvalues in the open right half plane. Recall that $C(-1)$ is obtained from $C(1)$ by a simple zero-cost diagonal preconditioning:

$$C(-1) = \begin{bmatrix} I_m & 0 \\ 0 & -I_n \end{bmatrix} C(1). \tag{30}$$

Because the matrix $C(-1)$, unlike $C(1)$, is not hermitian, there is no possibility of applying the methods MINRES or SYMMLQ or LSQR. Methods such as the restarted GMRES or Bi-CGSTAB or QMR may be used freely.

In case the matrices G and H have different magnitudes, we can "balance" $C(-1)$ by "preconditioning" H with a suitable nonsingular matrix such that the "preconditioned" H will be comparable to G in size. This can be achieved by replacing (29) by

$$\begin{bmatrix} G & \tilde{H} \\ -\tilde{H}^* & 0 \end{bmatrix} \begin{bmatrix} u \\ \tilde{v} \end{bmatrix} = \begin{bmatrix} f \\ -\tilde{g} \end{bmatrix}, \tag{31}$$

where

$$\tilde{H} = HS, \;\; \tilde{v} = S^{-1}v, \;\; \tilde{g} = S^*g. \tag{32}$$

In particular, we can choose S to be a diagonal scaling matrix. It is obvious that the spectrum of the "balanced" $C(-1)$ will also lie in the open right half plane. Then the application of Krylov subspace methods to the system in (32) can be more appropriate.

If GMRES(k), the restarted GMRES for some fixed integer k, is the Krylov subspace method being used, then its performance can be further enhanced as we discuss next. As $C(-1)$ has all its eigenvalues λ_k in the open right half plane, hence $|\arg \lambda_k| < \pi/2$ for all k, for every real scalar ω that satisfies

$$0 < \omega < \frac{2\cos\beta}{\rho(C(-1))}, \;\; \beta = \max_k |\arg \lambda_k|, \tag{33}$$

the Richardson iterative method defined by

$$\begin{bmatrix} u_{s+1} \\ v_{s+1} \end{bmatrix} = \begin{bmatrix} u_s \\ v_s \end{bmatrix} + \omega \left(\begin{bmatrix} f \\ -g \end{bmatrix} - \begin{bmatrix} G & H \\ -H^* & 0 \end{bmatrix} \begin{bmatrix} u_s \\ v_s \end{bmatrix} \right), \tag{34}$$

converges. (For a proof of this fact, see Sidi [16].) With ω chosen in this way, we now apply GMRES in the following manner: We choose a positive integer j in addition to k. Starting with some arbitrary initial vector $x_0 = [u_0^T \; v_0^T]^T$, we generate the vectors $x_1, x_2, \ldots, x_j$ by the Richardson iterative method as in (34). We next apply k steps of GMRES to the linear system in (29) with x_j serving as the initial vector for this purpose. Setting x_0 equal to the outcome of this application of GMRES, we repeat the same steps as many times as is necessary. Thus, each such cycle uses $j + k$ matrix-vector products with the matrix $C(-1)$. When applied in this form, GMRES has been denoted GMRES(j, k). [Note that GMRES$(0, k)$ is simply GMRES(k).]

This approach was originally incorporated into the computer program that implements vector extrapolation methods in the paper Sidi [17]. The analysis of GMRES(j, k) that is provided in Sidi and Shapira [18] shows, at least in some cases of interest, that the j Richardson iterations that precede the application of GMRES have a very beneficial effect. Specifically, if one cycle of GMRES(k), for some k, reduces the initial error by a given factor, then one cycle of GMRES(j, k), with $j > 0$, can reduce the initial error by a larger factor. It has been observed numerically that, starting with the same initial vector, GMRES(j, k), with some $j > 0$, requires a smaller number of matrix-vector products than GMRES(k) to reduce the initial error by a given factor, at least in some problems of interest. In addition, GMRES(j, k) has been observed to converge even in situations where GMRES(k) stagnates. Note that the storage requirements of both GMRES(k) and GMRES(j, k) are the same. In both cases, we need to store about k vectors in $\mathbb{C}^{m+n}$ (or in $\mathbb{R}^{m+n}$) in the present case.

It is interesting that, even when ω is such that the Richardson iterative method of (34) diverges (but slowly), GMRES(j, k) with moderate $j > 0$ (to prevent fast growth of the iteration vectors $x_1, x_2, \ldots, x_j$), can still produce very good convergence.

For details on these matters and for examples involving the numerical solution of partial differential equations, we refer the reader to [18]. Further applications (to indefinite systems) have been given in Shapira et al [15].

4 A Numerical Example

In this section, we demonstrate the validity of the arguments we presented in the previous section on a simple but instructive example. We choose in (10)

$$G = \mathrm{diag}\,(\mu_1, \mu_2, \ldots, \mu_m) \in \mathbb{C}^{m\times m},$$
$$H = \mathrm{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_n) \in \mathbb{C}^{m\times n}, \qquad (35)$$

where μ_k and σ_k are all positive. Thus, μ_k are the eigenvalues of G and σ_k are the singular values of H. With the technique of the proof of Theorem 1, it can be shown that the eigenvalues λ_j of $C(\alpha)$ are given as in

$$\lambda_{2k-1} = \frac{1}{2}\left(\mu_k + \sqrt{\mu_k^2 + 4\alpha\sigma_k^2}\right), \quad k = 1, \ldots, n,$$
$$\lambda_{2k} = \frac{1}{2}\left(\mu_k - \sqrt{\mu_k^2 + 4\alpha\sigma_k^2}\right), \quad k = 1, \ldots, n,$$
$$\lambda_{2n+k} = \mu_{n+k}, \quad k = 1, \ldots, m-n. \qquad (36)$$

Let us denote by x_j the eigenvector corresponding to the the eigenvalue λ_j, $j = 1, \ldots, m+n$. Then, $x_1, x_2, \ldots, x_{2n}$ are given as in

$$x_j = \begin{bmatrix} e_k^{(m)} \\ \alpha\sigma_k\lambda_j^{-1}e_k^{(n)} \end{bmatrix}, \quad j = 2k-1, 2k, \qquad (37)$$

if $\alpha \neq -\mu_k^2/(4\sigma_k^2)$, hence $\lambda_{2k-1} \neq \lambda_{2k}$. In case $\alpha = -\mu_k^2/(4\sigma_k^2)$, hence $\lambda_{2k-1} = \lambda_{2k} = \mu_k/2$, there is one eigenvector x_{2k-1} and one principal vector x_{2k} corresponding to $\lambda_{2k-1} = \lambda_{2k}$ that are given by

$$x_{2k-1} = \begin{bmatrix} e_k^{(m)} \\ -\mu_k/(2\sigma_k)e_k^{(n)} \end{bmatrix}, \quad x_{2k} = \begin{bmatrix} 0 \\ \sigma_k^{-1}e_k^{(n)} \end{bmatrix}. \qquad (38)$$

The eigenvectors x_{2n+k} corresponding to the eigenvalues λ_{2n+k} are given by

$$x_{2n+k} = \begin{bmatrix} e_{n+k}^{(m)} \\ 0 \end{bmatrix}, \quad k = 1, \ldots, m-n. \qquad (39)$$

Here $e_j^{(s)}$ is the jth standard basis vector in $\mathbb{C}^s$.

The numerical results we present here have been obtained by choosing

$$\mu_k = 1/k^2, \; k = 1, \ldots, m,$$
$$\sigma_k = 2^{p-k}, \; k = 1, \ldots, n; \;\; p \text{ a positive integer.} \qquad (40)$$

In Tables 1–3, we present the results obtained for the case $m = 35$ and $n = 15$, with $p = 2$, $p = 4$, and $p = 6$, respectively. In each case, the exact solution to the linear system in (1) has been chosen as 1 for each of the unknowns.

i	$\|r_i^+\|$	$\|e_i^+\|$	$\|r_i^-\|$	$\|e_i^-\|$
5	$4.60D-05$	$1.45D+00$	$3.31D-06$	$6.41D-01$
10	$5.08D-06$	$8.28D-01$	$6.08D-07$	$1.76D-01$
15	$1.05D-06$	$2.48D-01$	$2.05D-09$	$2.61D-04$
20	$1.00D-08$	$2.60D-03$	$5.62D-14$	$1.58D-08$
25	$2.19D-09$	$6.02D-04$	$5.13D-18$	$2.86D-15$

Table 1: Numerical results from GMRES(25) on the example of Section 4 with $p = 2$ in (40). Here $r_i^{\pm}$ and $e_i^{\pm}$ are respectively the residual vector and the error vector at the end of the ith cycle of GMRES(25) and $\|\cdot\|$ stands for the vector l_2-norm. r_i^+ and e_i^+ are related to the system (1), and r_i^- and e_i^- are related to the system (8). The initial vector is zero.

Acknowledgement

The author would like to thank Professor Roy Nicolaides for discussions on the application of iterative methods during a visit to NASA Glenn Research Center in the summer of 1991. This paper grew out of those discussions and is a slightly extended version of a report, with the same title, that was completed in April 2002.

i	$\|r_i^+\|$	$\|e_i^+\|$	$\|r_i^-\|$	$\|e_i^-\|$
1	$2.62D-03$	$2.40D+00$	$1.36D-03$	$2.00D+00$
2	$7.85D-04$	$1.77D+00$	$1.72D-04$	$1.20D+00$
3	$5.01D-04$	$1.32D+00$	$1.03D-04$	$9.24D-01$
4	$3.49D-04$	$1.36D+00$	$6.12D-05$	$7.07D-01$
5	$2.93D-04$	$1.25D+00$	$2.41D-05$	$4.01D-01$
6	$2.54D-04$	$1.26D+00$	$8.92D-06$	$1.06D-01$
7	$2.27D-04$	$1.19D+00$	$3.25D-06$	$5.58D-02$
8	$2.05D-04$	$1.20D+00$	$6.00D-07$	$4.71D-03$
9	$1.87D-04$	$1.14D+00$	$8.97D-08$	$1.19D-03$
10	$1.73D-04$	$1.15D+00$	$1.99D-09$	$2.81D-06$
11	$1.61D-04$	$1.09D+00$	$4.20D-11$	$6.83D-07$
12	$1.00D-04$	$9.05D-01$	$1.27D-13$	$1.08D-10$
13	$7.12D-05$	$7.77D-01$	$1.05D-17$	$2.52D-15$
14	$3.65D-05$	$5.32D-01$	$8.75D-18$	$1.06D-15$
15	$2.65D-05$	$4.58D-01$	$7.30D-18$	$9.09D-16$

Table 2: Numerical results from GMRES(25) on the example of Section 4 with $p = 4$ in (40). Here $r_i^{\pm}$ and $e_i^{\pm}$ are respectively the residual vector and the error vector at the end of the ith cycle of GMRES(25) and $\|\cdot\|$ stands for the vector l_2-norm. r_i^+ and e_i^+ are related to the system (1), and r_i^- and e_i^- are related to the system (8). The initial vector is zero.

i	$\|r_i^+\|$	$\|e_i^+\|$	$\|r_i^-\|$	$\|e_i^-\|$
1	$8.10D-03$	$3.72D+00$	$1.13D-02$	$4.05D+00$
2	$4.74D-03$	$2.68D+00$	$4.54D-03$	$2.76D+00$
3	$3.46D-03$	$2.31D+00$	$1.99D-03$	$1.43D+00$
4	$2.81D-03$	$2.03D+00$	$5.66D-05$	$5.10D-02$
5	$2.45D-03$	$1.88D+00$	$1.56D-06$	$1.07D-03$
6	$2.25D-03$	$1.81D+00$	$6.98D-09$	$5.52D-06$
7	$1.00D-03$	$1.06D+00$	$3.43D-11$	$2.56D-08$
8	$6.68D-04$	$7.20D-01$	$3.52D-15$	$9.09D-13$
9	$3.07D-04$	$2.79D-01$	$9.93D-16$	$1.09D-15$
10	$3.31D-05$	$2.08D-02$	$4.10D-18$	$7.93D-16$

Table 3: Numerical results from GMRES(25) on the example of Section 4 with $p = 6$ in (40). Here $r_i^{\pm}$ and $e_i^{\pm}$ are respectively the residual vector and the error vector at the end of the ith cycle of GMRES(25) and $\|\cdot\|$ stands for the vector l_2-norm. r_i^+ and e_i^+ are related to the system (1), and r_i^- and e_i^- are related to the system (8). The initial vector is zero.

References

[1] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, 1996.

[2] N. Dyn and W.E. Ferguson, Jr. The numerical solution of equality-constrained quadratic programming problems. *Math. Comp.*, 41:165–170, 1983.

[3] H.C. Elman, D.J. Silvester, and A.J. Wathen. Performance and analysis of saddle point preconditioners for the discrete steady-state Navier-Stokes equations. *Numer. Math.*, 90:665–688, 2002.

[4] B. Fischer and F. Peherstorfer. Chebyshev approximation via polynomial mappings and the convergence behaviour of Krylov subspace methods. *Electr. Trans. Numer. Anal.*, 12:205–215, 2001.

[5] B. Fischer, A. Ramage, D.J. Silvester, and A.J. Wathen. Minimal residual methods for augmented systems. *BIT*, 38:527–543, 1998.

[6] R. Freund and N. Nachtigal. QMR: A quasi-minimal residual method for non-Hermitian linear systems. *Numer. Math.*, 60:315–339, 1991.

[7] G.H. Golub and C.F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, London, third edition, 1996.

[8] G.H. Golub and A.J. Wathen. An iteration for indefinite systems and its application to the Navier-Stokes equations. *SIAM J. Sci. Comput.*, 19:530–539, 1998.

[9] T.A. Manteuffel. The Tchebychev iteration for nonsymmetric linear systems. *Numer. Math.*, 28:307–327, 1977.

[10] M.F. Murphy, G.H. Golub, and A.J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM J. Sci. Comput.*, 21:1969–1972, 2000.

[11] R. Nicolaides. Private communication.

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

[13] C.C. Paige and M.A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Software*, 8:43–71, 1982.

[14] Y. Saad and M.H. Schultz. GMRES: A generalized minimal residual method for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7:856–869, 1986.

[15] Y. Shapira, M. Israeli, A. Sidi, and U. Zrahia. Preconditioning spectral element schemes for definite and indefinite problems. *Numer. Methods Partial Differential Eqs.*, 15:535–543, 1999.

[16] A. Sidi. Application of vector extrapolation methods to consistent singular linear systems. *Appl. Numer. Math.*, 6:487–500, 1990.

[17] A. Sidi. Efficient implementation of minimal polynomial and reduced rank extrapolation methods. *J. Comp. Appl. Math.*, 36:305–337, 1991. Originally appeared as NASA TM-103240 ICOMP-90-20.

[18] A. Sidi and Y. Shapira. Upper bounds for convergence rates of acceleration methods with initial iterations. *Numer. Algorithms*, 18:113–132, 1998.

[19] D. Silvester and A. Wathen. Fast iterative solution of stabilised Stokes systems part II: Using general block preconditioners. *SIAM J. Numer. Anal.*, 31:1352–1367, 1994.

[20] H. van der Vorst. Bi-CGSTAB: A fast and smoothly convergent variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 13:631–644, 1992.

[21] A. Wathen and D. Silvester. Fast iterative solution of stabilised Stokes systems. part I: Using simple diagonal preconditioners. *SIAM J. Numer. Anal.*, 30:630–649, 1993.