Eidgenössische Technische Hochschule Zürich

Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo
Swiss Federal Institute of Technology Zurich

Short-term recurrences for indefinite preconditioning of saddle point problems*

M. Rozložník† and V. Simoncini‡

Research Report No. 2000-08
July 2000

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

*This work was carried out within the framework of the CNR Italy – Academy of Sciences of the Czech Republic bilateral contract CNR/AVCR, 1998–2000.

†Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, CZ-182 07 Prague 8, Czech Republic. E-mail `miro@cs.cas.cz`. Part of this work was supported by the Grant Agency of the Czech Republic under grants No. 101/00/1035 and 201/00/0080.

‡Istituto di Analisi Numerica - CNR, Pavia, Italy. E-mail `val@dragon.ian.pv.cnr.it`.

Short-term recurrences for indefinite preconditioning of saddle point problems*

M. Rozložník† and V. Simoncini‡

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

Research Report No. 2000-08 July 2000

Abstract

In this paper we analyze the null-space projection (constraint) indefinite preconditioner applied to the solution of large-scale saddle point problems. Nonsymmetric Krylov subspace solvers are considered and it is shown that the behavior of short-term recurrence methods can be related to the behavior of preconditioned conjugate gradient method (PCG). Theoretical properties of PCG are studied in detail and simple procedures for correcting possible misconvergence are proposed. The numerical behavior of the scheme on a real application problem is discussed and the maximum attainable accuracy of the approximate solution computed in finite precision arithmetic is estimated.

*This work was carried out within the framework of the CNR Italy – Academy of Sciences of the Czech Republic bilateral contract CNR/AVCR, 1998–2000.

†Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, CZ-182 07 Prague 8, Czech Republic. E-mail miro@cs.cas.cz. Part of this work was supported by the Grant Agency of the Czech Republic under grants No. 101/00/1035 and 201/00/0080.

‡Istituto di Analisi Numerica - CNR, Pavia, Italy. E-mail val@dragon.ian.pv.cnr.it.

1. Introduction. We consider the symmetric indefinite system of linear equations

$$\begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{1.1}$$

where the $n \times n$ matrix block A is symmetric positive definite and the $n \times m$ block B has full column rank. We denote by M the coefficient matrix and for the system (1.1) we also use the notation $Mt = b$ with $t = [x; y]$ and $b = [f; g]$.

Systems of the form (1.1) arise in many application problems such as mixed or mixed–hybrid finite element discretization of partial differential equations and quadratic or nonlinear programming with equality constraints; see [25, 24, 23, 18, 7, 17] and their references. In the partial differential equation context, we are particularly interested in the case in which A (B) corresponds to a zero (one) order operator, such as in mixed formulations of elliptic problems.

Due to the high sparsity of the coefficient matrix, the linear system (1.1) may be efficiently solved using iterative schemes. In order to improve the efficiency of standard iterative solvers some preconditioning technique is commonly employed, such as simple diagonal scaling, incomplete factorization of the system matrix or its inverse, up to problem dependent preconditioning [27, 1, 26, 2, 22]. Block matrices such as that in (1.1) naturally lead to the implementation of ad–hoc algebraic preconditioning strategies that aim to exploit the block structure of the original system; see for instance [23, 10, 3]. Especially attractive is positive definite preconditioning, where symmetric solvers are regularly applicable [25, 29].

In our paper we concentrate on the use of the symmetric but indefinite preconditioner

$$P = \begin{bmatrix} I & B \\ B^T & 0 \end{bmatrix}. \tag{1.2}$$

This choice has be shown to be particularly effective on problems associated with constrained nonlinear programming [23, 17, 18]. More precisely, it can be shown that this preconditioner projects the problem onto the kernel of the constraint operator, and that the constraint equation is exactly satisfied [24, 18].

Due to the indefiniteness of the preconditioning matrix P, the preconditioned system is naturally nonsymmetric so that nonsymmetric solvers must be applied. Although this fact could be considered as a practical drawback, experience on real problems has demonstrated good performance of this approach [23, 17, 4, 18]. The computationally expensive generalized minimum residual (GMRES) method [28] can be applied on the preconditioned system; in practice, however, simplified versions of short–term recurrence methods such as nonsymmetric biconjugate gradient (Bi-CG) or quasi-minimum residual (QMR) [5] methods can also be used.

A thorough analysis of the preconditioner P for a class of magnetostatic problems has been performed in [23], where optimality with respect to the mesh parameter has been shown. In this paper, we instead concentrate on algebraic properties of the preconditioned iteration process and on the connection between short–term recurrence methods and the preconditioned Conjugate Gradient (CG) approach. This analysis is motivated by the theoretical as well as numerical results in [18, 11], where CG and the conjugate residual method were successfully applied to the indefinite system (1.1) preconditioned by the indefinite preconditioner (1.2) for $g = 0$. We show the equivalence between CG and simplified BiCG when right–preconditioning is applied; the convergence analysis of preconditioned CG leads to the development of safeguard strategies

to avoid possible misconvergence of the indefinite CG iteration. We also show that round–off may considerably influence the performance of the applied method, and we provide theoretical results on the behavior of the approximate solution in finite precision arithmetic. As a general result, we derive that the motivation for applying a diagonal pre–scaling of the block matrix A is threefold: (i) together with indefinite preconditioning it leads to independence of the problem size of the iterative solver [23]; (ii) it ensures convergence of the CG method and (iii) it preserves numerical stability of the scheme in finite precision arithmetic.

Finally, we note that general convergence results are given for the residual norm minimizing GMRES method, though a long–term recurrence approach. Nevertheless, GMRES represents a reference method, at least for theoretical purposes, for quasi–optimal cheaper schemes such as BiCG.

The outline of the paper is as follows. In section 2 we study some theoretical properties of a general (nonsymmetric) Krylov subspace method applied to the preconditioned system and the setting for the subsequent sections is described. In section 3 several possible solution methods are discussed and related to previous works. The residual norm minimizing GMRES is studied in detail in section 4 and the related results are compared in subsequent sections with those of short–term recurrence methods. The analysis of the case $g = 0$ starts in section 5. In the subsequent section it is shown that the (theoretical) rate of convergence of the preconditioned GMRES method, up to a small factor, depends only on the spectral distribution of the preconditioned matrix, making this computationally expensive method interesting from a theoretical point of view. The equivalence between simplified BiCG and CG is shown in section 7, so that in the subsequent sections the CG method is analyzed in detail. More precisely, in section 8 we prove that for the preconditioned CG method the M–norm of the error decreases monotonically, whereas the residual norm can show completely different convergence history and it may even diverge unless special measures (correction or suitable scaling) are used to avoid this difficulty. In section 10 it is shown that not only the theoretical rate of convergence (measured by the easily computable residual norm) but also the maximum attainable accuracy level of the approximate solution computed in finite precision arithmetic depends heavily on the scaling of the matrix block A. The use of the CG method applied to the suitably scaled symmetric indefinite system (1.1) together with indefinite preconditioning (1.2) and $g = 0$ is thus theoretically well–justified. Numerical experiments also on a real application problem confirm the described theoretical results. In section 11 we draw our conclusions.

The notation used in this paper is as follows. Matlab notation is always used when possible. Vectors corresponding to the large system will be usually split as $v = [v^{(1)}; v^{(2)}]$ with $v^{(1)} \in \mathbb{R}^n$ and $v^{(2)} \in \mathbb{R}^m$, unless different letters are given to the two block vectors. Given $x \in \mathbb{R}^n$, x^T denotes the transpose vector; the 2–norm and the H–norm of x are defined as $||x||^2 = x^T x = \sum_{i=1}^n x_i^2$ and $||x||_H^2 = x^T H x$, respectively. The norm induced by the vector 2–norm is used for matrices. $\mathbb{P}_k$ indicates the set of polynomials of degree at most k. Finally, $\mathcal{N}(X)$ and $\mathrm{span}\{X\}$ indicate the null and range spaces of the matrix X, respectively.

2. Indefinite preconditioning. Given a starting approximation t_0 and the associated residual $r_0 = b - Mt_0$, the indefinite preconditioner P may be applied either from the right, yielding the system

$$MP^{-1}\hat{t} = r_0 \qquad t = P^{-1}\hat{t}, \tag{2.1}$$

or from the left, so that the system to be solved becomes

$$P^{-1}Mt = P^{-1}r_0, \tag{2.2}$$

(left–right preconditioning will not be considered in this paper, although it does not entail major consequences in the analysis). When standard nonsymmetric systems are preconditioned, the difference between the two approaches in (2.1) and (2.2) is that the former monitors the convergence of the true residual and preconditioned solution, whereas the latter monitors the preconditioned residual and the approximate solution to the original problem. We will see that for our particular problem there may be a close connection between the true residual and the preconditioned residual from the right and left preconditioned method, respectively, and their corresponding approximate solutions may even coincide for certain methods when carefully implemented.

The eigenvalues of $P^{-1}M$ and MP^{-1} are equal, therefore general spectral results can be given in terms of any of the two formulations. We first recall the following result [23, 18, 17]. Here and in the following, $\Pi = B(B^TB)^{-1}B^T$ denotes the orthogonal projector onto span$\{B\}$.

PROPOSITION 2.1. *Let λ be an eigenvalue of MP^{-1}. Then either $\lambda = 1$ or λ is a nonzero eigenvalue of $(I - \Pi)A(I - \Pi)$.*

Due to the positive definiteness of A, the eigenvalues of MP^{-1} are thus all real and positive. Moreover, the eigenvalue $\lambda = 1$ will be isolated if A is scaled so that the nonzero eigenvalues of $(I - \Pi)A(I - \Pi)$ are all smaller or larger than one. Unfortunately, the matrix MP^{-1} is not diagonalizable and the standard analysis on the convergence rate of residual minimizing methods ([13]) cannot be directly applied.

The inverse of the preconditioner P can be written as

$$P^{-1} = \begin{bmatrix} I - \Pi & B(B^TB)^{-1} \\ (B^TB)^{-1}B^T & -(B^TB)^{-1} \end{bmatrix}, \tag{2.3}$$

so that

$$MP^{-1} = \begin{bmatrix} A(I - \Pi) + \Pi & (A - I)B(B^TB)^{-1} \\ 0 & I \end{bmatrix}. \tag{2.4}$$

For brevity, we shall also use the notation

$$MP^{-1} = \begin{bmatrix} G & S \\ 0 & I \end{bmatrix} \tag{2.5}$$

with obvious meaning of G and S. Due to the symmetry of the matrices M and P, the coefficient matrix in the left preconditioned system is partitioned as

$$P^{-1}M = (MP^{-1})^T = \begin{bmatrix} G^T & 0 \\ S^T & I \end{bmatrix}.$$

When solving the right preconditioned system with a Krylov subspace method*, the subspace $K_k(MP^{-1}, r_0)$ is computed, while left preconditioning computes the subspace $K_k(P^{-1}M, P^{-1}r_0)$. Vectors belonging to Krylov subspaces can be written in terms of polynomials; therefore, if $v \in K_{k+1}(M, r_0)$, then $v = \phi(M)r_0$ for some polynomial $\phi \in \mathbb{P}_k$ [27].

*Given a matrix H and a vector v, a Krylov subspace of at most dimension k is the space spanned by $\{v, Hv, \ldots, H^{k-1}v\}$ and is denoted by $K_k(H, v)$.

We next show that vectors in $K_{k+1}(MP^{-1}, r_0)$ and in $K_{k+1}(P^{-1}M, P^{-1}r_0)$ can in fact be written in terms of polynomials in the matrix G defined in (2.5). These results will be used in the next sections to describe the residual behavior of selected Krylov subspace methods.

LEMMA 2.2. *A vector $v \in K_{k+1}(MP^{-1}, r_0)$ can be written as*

$$v = \phi_k(MP^{-1})r_0 = \begin{bmatrix} \phi_k(G)r_0^{(1)} + \psi_{k-1}(G)Sr_0^{(2)} \\ \phi_k(1)r_0^{(2)} \end{bmatrix}, \quad \phi_k \in \mathbb{P}_k \tag{2.6}$$

where the polynomial ψ_{k-1} is of degree at most $k-1$ and is defined as

$$\psi_{k-1}(\lambda) = \begin{cases} \phi_k'(\lambda) & \lambda = 1 \\ \frac{\phi_k(\lambda) - \phi_k(1)}{\lambda - 1} & \lambda \neq 1. \end{cases} \tag{2.7}$$

Proof. By explicitly writing the polynomial we see that the vector v satisfies

$$v^{(1)} = \phi_k(MP^{-1}r_0)|_{1:n} \quad v^{(2)} = \phi_k(1)r_0^{(2)}.$$

Moreover, since $(MP^{-1})^k r_0|_{1:n} = G^k r_0^{(1)} + G^{k-1}Sr_0^{(2)} + G^{k-2}Sr_0^{(2)} + \cdots + Sr_0^{(2)}$, we obtain for the polynomial $\phi_k(\lambda) = \sum_{i=0}^m \alpha_i \lambda^i$,

$$\begin{aligned} \phi_k(MP^{-1}r_0)|_{1:n} &= \alpha_0 r_0^{(1)} + \alpha_1(Gr_0^{(1)} + Sr_0^{(2)}) + \alpha_2(G^2 r_0^{(1)} + GSr_0^{(2)} + Sr_0^{(2)}) \\ &\quad + \alpha_3(G^3 r_0^{(1)} + G^2 Sr_0^{(2)} + GSr_0^{(2)} + Sr_0^{(2)}) + \cdots \\ &= \phi_k(G)r_0^{(1)} + \psi_{k-1}(G)Sr_0^{(2)}. \end{aligned}$$

The polynomial ψ is defined as

$$\psi_{k-1}(\lambda) = \alpha_1 + (1+\lambda)\alpha_2 + (1+\lambda+\lambda^2)\alpha_3 + \cdots + (1+\lambda+\cdots+\lambda^{k-1})\alpha_k.$$

For $\lambda = 1$, $\psi_{k-1}(1) = \alpha_1 + 2\alpha_2 + \cdots + k\alpha_k = \phi_k'(1)$. For $\lambda \neq 1$ we can write $(1+\lambda+\cdots+\lambda^{k-1}) = (1-\lambda^k)(1-\lambda)^{-1}$ so that

$$\begin{aligned} \psi_{k-1}(\lambda) &= (1-\lambda)^{-1}\left((1-\lambda)\alpha_1 + (1-\lambda^2)\alpha_2 + \cdots + (1-\lambda^k)\alpha_k\right) \\ &= (1-\lambda)^{-1}(\phi(1) - \phi(\lambda)). \end{aligned}$$

□

More comments on the role of ϕ_k and ψ_{k-1} will be given in the next sections.

We next show that a similar relation for the Krylov subspace generated with the left preconditioned matrix can be obtained. We also observe that a polynomial description of an element $w \in K_{k+1}(P^{-1}M, P^{-1}r_0)$ could be also obtained directly from the previous result as $w = P^{-1}\phi_k(MP^{-1})r_0$, yielding however a less insightful relation, at least for general r_0 (cf. section 5 for the case $r_0 = [r_0^{(1)}; 0]$).

LEMMA 2.3. *A vector $w \in K_{k+1}(P^{-1}M, \tilde{r}_0)$ with $\tilde{r}_0 = P^{-1}r_0$ can be written as*

$$w = \phi_k(P^{-1}M)P^{-1}r_0 = \begin{bmatrix} \phi(G^T)\tilde{r}_0^{(1)} \\ S^T\psi_{k-1}(G^T)\tilde{r}_0^{(1)} + \phi_k(1)\tilde{r}_0^{(2)} \end{bmatrix}, \quad \phi_k \in \mathbb{P}_k \tag{2.8}$$

with ψ_{k-1} as in (2.7).

Although left and right preconditioning in general generate different spaces in which an approximate solution is computed, the first block of the approximate solution

to the original problem (1.1) always belongs to the same space, regardless of the side the preconditioner is employed. This is shown in the following proposition.

PROPOSITION 2.4. *Let $t_k = [x_k; y_k]$ be the approximate solution to (1.1) either in $K_k(MP^{-1}, r_0)$ or in $K_k(P^{-1}M, P^{-1}r_0)$. Then $x_k = \phi(G^T)\tilde{r}_0^{(1)}$ for some $\phi \in \mathbb{P}_{k-1}$, where $\tilde{r}_0 = P^{-1}r_0 = [\tilde{r}_0^{(1)}; \tilde{r}_0^{(2)}]$. (The polynomial may not be the same for the two spaces)*

Proof. We first show that t_k belongs to $K_k(P^{-1}M, P^{-1}r_0)$ for both right and left preconditioning. For left preconditioning, the result follows from Lemma 2.3.

Let V_k be a basis of $K_k(MP^{-1}, r_0)$ with

$$MP^{-1}V_k = V_{k+1}H_k \tag{2.9}$$

and $H_k \in \mathbb{R}^{(k+1)\times k}$ upper Hessenberg. It can be shown that $Q_k = P^{-1}V_k$ is a basis of $K_k(P^{-1}M, P^{-1}r_0)$. Let $\hat{t}_k = V_k z_k \in K_k(MP^{-1}, r_0)$ be an approximate solution to the right preconditioned system $MP^{-1}\hat{t} = r_0$. Then the approximate solution t_k to the unpreconditioned system $Mt = r_0$ is computed as $t_k = P^{-1}\hat{t}_k = P^{-1}V_k z_k = Q_k z_k$ so that $t_k \in K_k(P^{-1}M, P^{-1}r_0)$.

Using (2.9), we obtain $P^{-1}MQ_k = Q_{k+1}H_k$, so that the basis $Q_k = [Q_k^{(1)}; Q_k^{(2)}]$ satisfies

$$\begin{bmatrix} G^T & 0 \\ S^T & I \end{bmatrix} \begin{bmatrix} Q_k^{(1)} \\ Q_k^{(2)} \end{bmatrix} = \begin{bmatrix} Q_{k+1}^{(1)} \\ Q_{k+1}^{(2)} \end{bmatrix} H_k$$

and in particular, $G^T Q_k^{(1)} = Q_{k+1}^{(1)} H_k$. Therefore, $\text{span}\{Q_k^{(1)}\} = K_k(G^T, q_1^{(1)})$, where $q_1^{(1)}$ is the first vector of the matrix $Q_k^{(1)}$. Recalling from $t_k = Q_k z_k$ that $x_k = Q_k^{(1)} z_k$, the result follows. □

The proposition above shows that the convergence to the first block of the solution may depend only on the properties of the matrix G.

3. Solution methods. The preconditioned coefficient matrix is nonsymmetric, therefore nonsymmetric solvers seem to be required. Preconditioned GMRES determines an approximate solution in the generated Krylov subspace so as to minimize its residual 2–norm. This optimality condition is obtained by explicitly constructing an orthogonal basis of the computed Krylov subspace [27]. Due to the high computational cost per iteration, GMRES in its original implementation is discarded in practical situations. Quasi-optimal methods are preferred: these give up optimality by omitting the generation of the full orthogonal basis (e.g. restarted GMRES, BiCG, BiCGSTAB).

Classical Lanczos–type approaches such as BiCG employ short–term recurrences to generate the subspace by imposing a bi-orthogonality condition between the basis elements of two distinct subspaces. The computational cost grows only linearly with the number of iterations, while quasi-monotonic behavior of the residual norm may be obtained by employing a smoothing procedure [5, 30]. Given the starting residual r_0 and an auxiliary vector $\tilde{r}_0$, the two Krylov subspaces $K_k(MP^{-1}, r_0)$ and $K_k((MP^{-1})^T, \tilde{r}_0)$ are constructed if right preconditioning is used; the two spaces are usually called right and left Krylov subspaces. Analogously, if left preconditioning is considered, the right and left generated spaces are $K_k(P^{-1}M, P^{-1}r_0)$ and $K_k((P^{-1}M)^T, \tilde{r}_0)$. By comparing the two preconditioning approaches, it is clear that right preconditioning with $\tilde{r}_0 = P^{-1}r_0$ exactly corresponds to reversing the role of right and left spaces in the left preconditioning with $\tilde{r}_0 = r_0$. This consideration,

together with the result of Proposition 2.4, shows that left and right preconditionings of the indefinite problem provide similar information, at least for the first block of the approximate solution vector. We will see that care must be taken in the approximation of the second block when right or left preconditioning is applied. We should remark, however, that often the second block vector refers to terms that do not have physical meaning and therefore are discarded in real applications.

Because of the symmetry of P and M, a lot of redundant information is generated when constructing the right and left spaces. This is clearly seen when choosing $\tilde{r}_0 = P^{-1}r_0$ as auxiliary vector in right preconditioning. Indeed, in this case,

$$((MP^{-1})^T)^k\tilde{r}_0 = (P^{-1}M)^k\tilde{r}_0 = P^{-1}(MP^{-1})^k r_0, \quad \forall k \geq 0,$$

so that vectors in the left space $K_k((MP^{-1})^T, \tilde{r}_0)$ can be simply obtained by premultiplying by P^{-1} vectors in the right space $K_k(MP^{-1}, r_0)$.

This is a special case of the more general $J-$*symmetry* property. A matrix H is called J-symmetric if there exists a nonsingular matrix J such that $H^TJ = JH$, that is H is (real) symmetric with respect to J. It was shown in [15] and later developed in [6] that J–symmetry can be exploited so as to decrease the computational cost of nonsymmetric Lanczos processes. In summary, when the coefficient matrix is J–symmetric, the auxiliary Lanczos recurrence that is used to generate the left space is obtained at low cost from the computed right basis vectors. For right preconditioning, $H = MP^{-1}$ and $J = P^{-1}$, while for left preconditioning, $H = P^{-1}M$ and $J = P$; We refer to [6] for implementation issues concerning J–symmetry. J–symmetry of the preconditioned matrix was used in [23, 24] to enhance the efficiency of iterative solvers on real application problems.

It already appears from the results given so far that if nonsymmetric short-term recurrence methods are applied, the analysis and the experimental results will substantially differ depending on the choice of the auxiliary vector. In this paper we shall focus on the special choice $\tilde{r}_0 = P^{-1}r_0$ for right preconditioning and $\tilde{r}_0 = r_0$ for left preconditioning, which lead to convenient computational savings as shown above. Moreover, we shall see that these choices of auxiliary vector $\tilde{r}_0$ also entail fundamental theoretical considerations.

4. General convergence results. General convergence results are not easily derived due to the nontrivial Jordan structure of the coefficient matrix MP^{-1}. This however, turns out to be unnecessary, since the block form introduced in (2.4) allows us to write the residual norm in terms of polynomials in G. From these, upper bounds for the residual norm can be readily obtained. More insightful relations can be written when the right–hand side of the system (1.1) is of the form $[f;0]$, that is when $g = 0$. We anticipate that setting $g = 0$ is not restrictive, since the starting approximate solution can be chosen so as to fall in such framework. We shall focus on the general case in this section, while the rest of the paper will be devoted to the analysis for $g = 0$.

For a diagonalizable coefficient matrix $C \in \mathbb{R}^{n\times n}$, a bound on the GMRES residual norm can be given as ([13])

$$||r_k^{\mathrm{GMRES}}|| \leq ||r_0||\kappa_2(Q) \min_{\substack{\phi\in\mathbb{P}_k \\ \phi(0)=1}} \max_{i=1,\ldots,n} |\phi(\lambda_i)|,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of C and $\kappa_2(Q) := ||Q||\,||Q^{-1}||$ is the condition number of its eigenvector basis Q. Although MP^{-1} does not have a full system

of eigenvectors, using the notation and the result of Lemma 2.2, a bound on the convergence of the GMRES residual can be written in terms of polynomials in the matrix $G = A(I - \Pi) + \Pi$. Indeed, the right preconditioned GMRES residual satisfies

$$\begin{aligned} ||r_k^{\mathrm{GMRES}}||^2 &= \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} ||\phi(MP^{-1})r_0||^2 \\ &= \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} \left(||\phi(G)r_0^{(1)} + \psi(G)Sr_0^{(2)}||^2 + |\phi(1)|^2 ||r_0^{(2)}||^2 \right), \end{aligned} \tag{4.1}$$

where the polynomial ψ is of degree at most $k-1$ and is defined through ϕ as in (2.7).

The presence of ψ in (4.1) shows that ϕ is chosen so as to have small derivative at the unit value, which seems to suggest that ϕ will grow only slowly in the neighborhood of one.

Analogously, using (2.8) with $\tilde{r}_0 = P^{-1}r_0$, left preconditioning gives

$$\begin{aligned} ||r_k^{\mathrm{GMRES}}||^2 &= \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} ||\phi(P^{-1}M)\tilde{r}_0||^2 \\ &= \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} \left(||\phi(G^T)\tilde{r}_0^{(1)}||^2 + ||S^T\psi(G^T)\tilde{r}_0^{(1)} + \phi(1)\tilde{r}_0^{(2)}||^2 \right). \end{aligned} \tag{4.2}$$

5. The case $g = 0$. This section serves as introduction to the following sections, where we shall focus on the case in which the original problem satisfies $g = 0$. We note that even though $g \neq 0$, the starting approximate solution t_0 can be chosen so that the starting residual has the form $r_0 = [s_0; 0]$, yielding in practice an equivalent setting as if g were equal to the zero vector. For this reason, we shall assume throughout this and the following sections that $g = 0$ and $t_0 = 0$, so that $r_0 = [f; 0]$.

We start by analyzing right preconditioning, which provides the most unexpected results in practical circumstances. We will show that for $g = 0$ the convergence analysis of GMRES can be carried out by only employing the upper left block matrix G in (2.4). Moreover, we show that simplified BiCG behaves very differently than expected, and that its convergence is strictly related to that of preconditioned CG on the indefinite problem.

In our analysis we will take advantage of some basic properties of matrices P^{-1} and MP^{-1} when applied to a vector $[v; 0]$. Namely, it follows that

$$P^{-1}\begin{pmatrix} v \\ 0 \end{pmatrix} = \begin{pmatrix} (I - \Pi)v \\ (B^TB)^{-1}B^Tv \end{pmatrix}, \quad MP^{-1}\begin{pmatrix} v \\ 0 \end{pmatrix} = \begin{pmatrix} Gv \\ 0 \end{pmatrix}. \tag{5.1}$$

Actually, there is a connection to the solution of the linear least squares problem associated with the matrix B and the right-hand side vector v: while the vector $(I - \Pi)v$ is the least squares residual, the vector $(B^TB)^{-1}B^Tv$ is the least squares solution.

If left preconditioning is used, then the condition $g = 0$ may not lead to significant changes in the generation of the Krylov subspace basis. Indeed, the vector generating the Krylov subspace in such case is $\tilde{r}_0 = [(I-\Pi)r_0^{(1)}; (B^TB)^{-1}B^Tr_0^{(1)}]$ which in general will not have zero blocks. We shall see later on that this fact does not represent a serious difficulty for Lanczos–type methods.

6. The GMRES method. By writing the GMRES residual as $r_k^{\mathrm{GMRES}} = \phi_k(MP^{-1})r_0$, where ϕ_k is the optimal GMRES residual polynomial, the optimality of the residual can be expressed only in terms of the matrix G; see also [23].

COROLLARY 6.1. *With the notation of Proposition 2.2 and for* $r_0 = [s_0; 0]$, *the right preconditioned GMRES residual satisfies*

$$||r_k^{\mathrm{GMRES}}|| = \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} ||\phi(G) s_0||.$$

Assuming $G \equiv A(I - \Pi) + \Pi$ diagonalizable, we obtain

$$||r_k^{\mathrm{GMRES}}|| \leq ||r_0|| \kappa_2(Z) \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} \max_{i=1,\cdots,n} |\phi(\lambda_i)| \tag{6.1}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of G and Z is its eigenvector matrix [23]. Consequently, although the system matrix MP^{-1} is non–diagonalizable, the rate of convergence of preconditioned GMRES depends only on the eigenvalue distribution of the block $A(I - \Pi) + \Pi$ and on the conditioning of its eigenvector basis. In the following proposition, we show that the matrix $A(I - \Pi) + \Pi$ does have a full set of eigenvectors and give a bound for its condition number. This result first appeared in [23] under stricter conditions.

PROPOSITION 6.2. *Let* $Z = [Z_1, Z_2]$ *be a nonsingular eigenvector matrix of* $(A(I - \Pi) + \Pi)$ *with* $(A(I - \Pi) + \Pi)Z_1 = Z_1 L$ *and* $(A(I - \Pi) + \Pi)Z_2 = Z_2$ *with* $L = diag(\ell_{ii})$, $\ell_{ii} \in \mathbb{R}$ *and* $\ell_{ii} \neq 1$. *Then* Z_2 *can be written as* $Z_2 = [Z_{1,2}, Z_{2,2}]$ *with* $span\{Z_{1,2}\} = span\{\Pi\}$, *and* $Z_{1,2}$ *can be chosen so that*

$$\kappa_2(Z) \leq (||[Z_1, Z_{2,2}]|| + 1)^2.$$

Proof. Since all the eigenvalues of $(A(I - \Pi) + \Pi)$ are real, Z can be taken to be real. Left multiplying both sides of $(A(I - \Pi) + \Pi)Z_2 = Z_2$ by $(I - \Pi)$, we obtain $(I - \Pi)A(I - \Pi)Z_2 = (I - \Pi)Z_2$. Let $[u_1, \ldots, u_k] = (I - \Pi)Z_2$. Then for each $j = 1, \ldots, k$ it must be either $u_j = 0$ or u_j is an eigenvector of $(I - \Pi)A$ corresponding to the unit eigenvalue. Then we can write $Z_2 = [Z_{1,2}, Z_{2,2}]$ with $(I - \Pi)Z_{1,2} = 0$ and span$\{Z_{1,2}\}$=span$\{\Pi\}$; moreover, since $Z_{1,2}$ is an eigenbasis corresponding to the eigenvalue one, we can take $Z_{1,2}$ to have orthogonal columns, so that $||Z_{1,2}|| = 1$. Set $\hat{Z} = [Z_1, Z_{2,2}]$ and let $[Y_1, Y_2]$ be such that $[\hat{Z}, Z_{1,2}]^{-1} = [Y_1, Y_2]^T$. It can be shown that Y_1 is an orthogonal basis of eigenvectors of $(I - \Pi)A(I - \Pi)$ corresponding to all its nonzero eigenvalues. Therefore, $||Y_1|| = 1$ and $[Y_1, Z_{1,2}]$ is an orthogonal basis of $\mathbb{R}^n$. Explicitly writing the condition $[Y_1, Y_2]^T[\hat{Z}, Z_{1,2}] = I$, it can also be verified that $Y_2 = -Y_1(\hat{Z}^T Z_{1,2}) + Z_{1,2}$ so that

$$||[Y_1, Y_2]|| = \left\| [Y_1, Z_{1,2}] \begin{bmatrix} I & -\hat{Z}^T Z_{1,2} \\ 0 & I \end{bmatrix} \right\| \leq (1 + ||\hat{Z}||).$$

Using $\kappa_2(Z) = ||[\hat{Z}, Z_{1,2}]|| \cdot ||[Y_1, Y_2]||$, the bound for $\kappa_2(Z)$ follows. ∎

Note that $\kappa_2(Z)$ only depends on the norm of a section of the eigenvector matrix. Since eigenvector norms can be chosen arbitrarily, it follows that $\kappa_2(Z)$ will not be much larger than one. More precisely, the result in Proposition 6.2 shows that $\kappa_2(Z)$ does not depend on the problem dimension. This fact is of great importance when solving systems arising in the discretization of differential equations by means of finite element methods [23].

Using standard results on Chebyshev polynomials to bound the polynomial min–max problem [13], we also obtain

$$\frac{||r_k^{\rm GMRES}||}{||r_0||} \leq 2\gamma \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^k$$

where $\kappa = \lambda_{max}/\lambda_{min}$ stands for the ratio of extremal eigenvalues of the generally nonsymmetric matrix $A(I-\Pi)+\Pi$ (and so it is not its condition number!) and where γ is a constant that bounds $\kappa_2(Z)$.

An analogous result is well known to hold for the M–norm of the relative PCG error, with $\gamma = 1$ and when M and P are positive definite.

When using left preconditioning, fewer simplifications take place. Using Proposition 2.3 the following relation for the left preconditioned GMRES residual can be simply obtained. The minimization problem (6.3) follows from (5.1) with

$$P^{-1} \left[\begin{array}{c} \phi(G)s_0 \\ 0 \end{array} \right] = \left[\begin{array}{c} (I-\Pi)\phi(G)s_0 \\ (B^TB)^{-1}B^T\phi(G)s_0 \end{array} \right]. \tag{6.2}$$

COROLLARY 6.3. *The left preconditioned GMRES residual norm with $r_0 = [s_0; 0]$ can be written as*

$$\begin{aligned} ||r_k^{\rm GMRES}||^2 &= \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} ||\phi(P^{-1}M)P^{-1}r_0||^2 \\ &= \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} \left(||(I-\Pi)\phi(G)s_0||^2 + ||(B^TB)^{-1}B^T\phi(G)s_0||^2 \right). \end{aligned} \tag{6.3}$$

More directly, from (6.2) we also obtain

$$\begin{aligned} ||r_k^{\rm GMRES}|| &\leq ||P^{-1}|| \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} ||\phi(G)s_0|| \\ &\leq ||P^{-1}||||r_0||\kappa_2(Z) \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} \max_{i=1,\cdots,n} |\phi(\lambda_i)|, \end{aligned}$$

where λ_i's are the eigenvalues of G with corresponding eigenvector matrix Z. The norm of P^{-1} is bounded as (cf. e.g. [25])

$$||P^{-1}|| \leq \max \left\{ \frac{2}{\sqrt{1+4\sigma_{min}(B)^2}-1}, 1 \right\}$$

where $\sigma_{min}(B)$ is the smallest singular value of B.

7. Other nonsymmetric solvers. In this section we briefly discuss nonsymmetric solvers that employ short–term recurrences and which can be used for solving our preconditioned system for $g = 0$.

Motivated by the considerations in section 3, we consider the implementation of simplified Bi-CG as short–term recurrence approach. We next show that for $r_0 = [s_0; 0]$ and provided that $\tilde{r}_0 = P^{-1}r_0$, simplified Bi–CG is equivalent to the CG method applied to the system $Mt = b$ with preconditioner P.

We start by recalling the classical right preconditioned BiCG recurrence: given $r_0, \tilde{r}_0$ and setting $p_0 = r_0$ and $\tilde{p}_0 = \tilde{r}_0$, for $k = 0, 1, \ldots$ we have

P^{-1}-symmetric Bi-CG(MP^{-1})	PCG(M)
$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$	$t_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$
$b - M\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} s_0 \\ 0 \end{pmatrix}$	$r_0 = b - Mt_0 = \begin{pmatrix} s_0 \\ 0 \end{pmatrix}$
$\begin{pmatrix} u_0 \\ v_0 \end{pmatrix} = \begin{pmatrix} s_0 \\ 0 \end{pmatrix}$	$p_0 = P^{-1}r_0$
$k = 0, 1, \ldots$	$k = 0, 1, \ldots$
$\hat{\alpha}_k = \frac{(\begin{pmatrix} s_k \\ 0 \end{pmatrix}, P^{-1}\begin{pmatrix} s_k \\ 0 \end{pmatrix})}{(MP^{-1}\begin{pmatrix} s_k \\ 0 \end{pmatrix}, P^{-1}\begin{pmatrix} s_k \\ 0 \end{pmatrix})}$	$\alpha_k = \frac{(r_k, P^{-1}r_k)}{(p_k, Mp_k)}$
$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} + \hat{\alpha}_k \begin{pmatrix} u_k \\ v_k \end{pmatrix}$	$t_{k+1} = t_k + \alpha_k p_k$
$\begin{pmatrix} s_{k+1} \\ 0 \end{pmatrix} = \begin{pmatrix} s_k \\ 0 \end{pmatrix} - \hat{\alpha}_k MP^{-1}\begin{pmatrix} u_k \\ v_k \end{pmatrix}$	$r_{k+1} = r_k - \alpha_k Mp_k$
$\hat{\beta}_k = \frac{(\begin{pmatrix} s_{k+1} \\ 0 \end{pmatrix}, P^{-1}\begin{pmatrix} s_{k+1} \\ 0 \end{pmatrix})}{(\begin{pmatrix} s_k \\ 0 \end{pmatrix}, P^{-1}\begin{pmatrix} s_k \\ 0 \end{pmatrix})}$	$\beta_k = \frac{(r_{k+1}, P^{-1}r_{k+1})}{(r_k, P^{-1}r_k)}$
$P^{-1}\begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = P^{-1}\begin{pmatrix} s_{k+1} \\ 0 \end{pmatrix} + \hat{\beta}_k P^{-1}\begin{pmatrix} u_k \\ v_k \end{pmatrix}$	$p_{k+1} = P^{-1}r_{k+1} + \beta_k p_k$

FIG. 7.1. *Equivalence of right preconditioned BiCG and PCG for* $r_0 = [s_0; 0]$ *and* $\tilde{r}_0 = P^{-1}r_0$.

$\alpha_k = (\tilde{r}_k, r_k)/(\tilde{p}_k, MP^{-1}p_k)$	
$t_{k+1} = t_k + \alpha_k p_k$	
$r_{k+1} = r_k - \alpha_k MP^{-1}p_k$	$\tilde{r}_{k+1} = \tilde{r}_k - \alpha_k P^{-1}M\tilde{p}_k$
$\beta_k = (\tilde{r}_{k+1}, r_{k+1})/(\tilde{r}_k, r_k)$	
$p_{k+1} = r_{k+1} + \beta_k p_k$	$\tilde{p}_{k+1} = \tilde{r}_{k+1} + \beta_k \tilde{p}_k$

Using J–symmetry (with $J = P^{-1}$) and by setting $\tilde{r}_0 = P^{-1}r_0$ we obtain $\tilde{r}_k = P^{-1}r_k$ for all subsequent $k > 0$, and analogously for $\tilde{p}_k$. Therefore, the iterates $\tilde{r}_k, \tilde{p}_k$ can be computed explicitly from r_k, p_k and the auxiliary "tilde" recurrence can be omitted. The resulting algorithm is nothing but the usual implementation of the CG method preconditioned with the indefinite matrix P [8]. In Figure 7.1 we report the obtained J–symmetric BiCG recurrence versus the Preconditioned CG recurrence for the choice $r_0 = [s_0; 0]$. If we look at the formulae of both algorithms in the figure, it is clear that $\hat{\alpha}_k = \alpha_k$ and $\hat{\beta}_k = \beta_k$ and both algorithms are equivalent for $t_k = [x_k; y_k]$, and if $r_k = [s_k; 0]$, $p_k = P^{-1}[u_k; v_k]$. This condition can be easily proved. Indeed, if $r_0 = [s_0; 0]$ and due to (5.1), the vector $p_0 = P^{-1}r_0 = [p_0^{(1)}; p_0^{(2)}]$ satisfies $B^T p_0^{(1)} = 0$ which gives $Mp_0 = [Ap_0^{(1)} + Bp_0^{(2)}; 0]$. Using induction, one can show for all $j = 0, 1, \ldots$ the properties $B^T p_{j+1}^{(1)} = 0$ and $Mp_{j+1} = [Ap_{j+1}^{(1)} + Bp_{j+1}^{(2)}; 0]$, which imply that r_{j+1} can be written in the form $r_{j+1} = [s_{j+1}; 0]$.

Equivalence can be also shown in the case of left preconditioning. Indeed, the P-symmetric BiCG applied to the preconditioned system with coefficient matrix $P^{-1}M$ and auxiliary vector $\tilde{r}_0 = r_0$, is equivalent to Preconditioned CG: more precisely, the quantities t_k and p_k coincide, while the left preconditioned BiCG residual corresponds to the preconditioned residual iterates $P^{-1}r_k$.

We note that simplified QMR can be also viewed (at least in exact arithmetic) as simplified Bi-CG method with the QMR residual smoothing procedure applied on its top; cf. for instance [14, 30].

8. Preconditioned CG. In light of the considerations of the previous section, we see that simplified BiCG for $g = 0$ reduces to standard preconditioned CG applied on (1.1) with preconditioner P. Clearly, the indefiniteness of both M and P does not make the algorithm robust, and breakdown may occur, as observed in [18, 19]; however, in [18] safeguard strategies were suggested to overcome possible breakdown, which the authors encountered at convergence stage. In this section we give a closer look at the behavior of CG on the indefinite system (1.1) and give explicit formula describing the possible (mis)convergence of the method.

Given the linear system $Mt = b$, initial guess t_0 with $r_0 = b - Mt_0$ and the preconditioner P, the preconditioned CG algorithm (PCG) generates iterates t_k with residuals $r_k = b - Mt_k$ and preconditioned residuals $z_k = P^{-1}r_k$, $k = 0, 1, \ldots$ such that the error $e_k = t - t_k$ satisfies

$$e_k \in e_0 + \{z_0, \cdots, z_k\} \quad e_k^T M z_j = e_k^T M P^{-1} M e_j = 0 \quad j = 0, \ldots, k.$$

If P and M were positive definite then the M–norm of e_k would be minimized over $e_0 + \{z_0, \cdots, z_k\}$. The error e_k can then be written in the form $e_k = \phi_k(P^{-1}M)e_0$, where ϕ_k is the CG polynomial of degree k such that $\phi_k(0) = 1$. The residual vector r_k satisfies $r_k = M\phi_k(P^{-1}M)e_0$ and

$$r_k \perp \{z_0, \cdots, z_k\}.$$

We have already shown that since $r_0 = [s_0; 0]$, then all subsequent r_j have second block component equal to zero, that is $r_j = [s_j; 0]$, $j = 0, 1, \ldots, k$. In particular, this implies that the approximate solution $[x_k; y_k]$ satisfies $B^T x_k = 0$ or, equivalently, $B^T e_k^{(1)} = 0$. The preconditioned residuals z_j, $j = 0, 1, \ldots, k$ then satisfy the relation $Mz_j = [(I - \Pi)s_j; 0]$, so that the M–orthogonality of the error $e_k = [e_k^{(1)}; e_k^{(2)}]$ gives

$$0 = e_k^T M z_j = (e_k^{(1)})^T (I - \Pi) s_j \quad j = 0, \ldots, k.$$

Therefore, the condition on the error is only imposed on the first block component. The presence of $(I - \Pi)$ shows that only the component of the error in the kernel of B is forced to be orthogonal to the previous residuals, and the error is minimized (in the indefinite M–norm) only in a subspace of the kernel space of B. Moreover, since $B^T e_k^{(1)} = 0$ for $k = 0, \ldots$, then

$$e_k^T M e_k = (e_k^{(1)})^T A e_k^{(1)} > 0 \quad \forall e_k^{(1)} \neq 0 \tag{8.1}$$

so that $||e_k||_M$ is always non–negative. Next result follows from the properties of the preconditioned residual in the PCG method.

PROPOSITION 8.1. *Let $e_0 = [e_0^{(1)}; e_0^{(2)}]$ be the starting error of PCG. Then*

$$||\phi_k(P^{-1}M)e_0||_M = \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} ||\phi(P^{-1}M)e_0||_M = \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} ||\phi((I - \Pi)A(I - \Pi))e_0^{(1)}||_A.$$

Proof. We have to prove for every polynomial ϕ that

$$||\phi(P^{-1}M)e_0||_M = ||\phi((I - \Pi)A(I - \Pi))e_0^{(1)}||_A.$$

Since $B^T e_0^{(1)} = B^T x = 0$, we have that $Me_0 = [Ae_0^{(1)} + Be_0^{(2)}; 0]$ and therefore $P^{-1}Me_0 = [(I-\Pi)Ae_0^{(1)}; \star]$. It also follows that $\phi(P^{-1}M)e_0 = [\phi((I-\Pi)A)e_0^{(1)}; \star]$. Since $e_0^{(1)} = (I-\Pi)e_0^{(1)}$, and using a similar approach as in (8.1) we obtain

$$||\phi(P^{-1}M)e_0||_M^2 = ||\phi((I-\Pi)A(I-\Pi))e_0^{(1)}||_A^2.$$

□

Since $e_0^{(1)} = (I-\Pi)e_0^{(1)}$, the M–norm of the error $e_k = \phi_k(P^{-1}M)e_0$ is minimized only over the set of nonzero eigenvalues of $(I-\Pi)A(I-\Pi)$. We thus have the following bound

$$\min_{\substack{\phi\in\mathbb{P}_k \\ \phi(0)=1}} ||\phi(P^{-1}M)e_0||_M \leq ||e_0^{(1)}||_A \min_{\substack{\phi\in\mathbb{P}_k \\ \phi(0)=1}} \max_{\lambda\in[\alpha,\beta]} |\phi(\lambda)| \tag{8.2}$$

where $[\alpha, \beta]$ is the smallest interval containing the nonzero eigenvalues of $(I-\Pi)A(I-\Pi)$. Using once more standard Chebyshev polynomial results, we see that the M–norm of the error decreases asymptotically at least as the optimal Chebyshev polynomial on $[\alpha, \beta]$. On the other hand, the residual norm of PCG (both the preconditioned residual and the true residual) does not obey the corresponding asymptotic rule and the convergence curve may differ dramatically. This is due to the fact that the quantity $||e_k||_M$ may be zero for nonzero e_k, with $e_k^{(1)} = 0$ and $e_k^{(2)} \neq 0$ (cf. (8.1)), showing that $||\cdot||_M$ is not a definite norm. We next show that this is the reason why the energy norm (the M–norm) fails to describe the convergence of the PCG residual on this problem. The residual $r_k = b - Mt_k$ satisfies

$$r_k = M\phi_k(P^{-1}M)e_0 = \begin{bmatrix} \phi_k(A(I-\Pi)+\Pi)s_0 \\ 0 \end{bmatrix}.$$

Let $A(I-\Pi)+\Pi = Z\Lambda Z^{-1}$ be the eigenvalue factorization of $A(I-\Pi)+\Pi$. Then

$$\begin{aligned} ||r_k|| &= ||\phi_k(A(I-\Pi)+\Pi)s_0|| \\ &\leq \kappa_2(Z)\,||s_0||\,||\phi_k(\Lambda)|| \leq \kappa_2(Z)\,||s_0||\,\max\{\phi_{max}, |\phi_k(1)|\} \end{aligned} \tag{8.3}$$

where $\phi_{max} = \max_{\lambda\in[\alpha,\beta]} |\phi_k(\lambda)|$ and ϕ_k is the optimal PCG polynomial in $[\alpha, \beta]$. While ϕ_{max} decreases as expected, $|\phi_k(1)|$ might not decrease (if it does at all) at the same rate if $1 \notin [\alpha, \beta]$. Therefore, the rate at which the bound of $||r_k||$ decreases depends on the value of the PCG polynomial ϕ_k at $\lambda = 1$. A similar dependence was already observed for GMRES. However, it is more crucial for PCG, since the optimal polynomial ϕ_k is minimized in $[\alpha, \beta]$, which might not contain the value 1, whereas in GMRES it does contain it. Assuming, however, that $1 \in [\alpha, \beta]$ and using the standard result on Chebyshev polynomials in (8.3) (see e.g. [13]), the following estimate holds for the relative residual norm of PCG

$$\frac{||r_k||}{||r_0||} \leq 2\gamma \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k, \tag{8.4}$$

where γ is a constant close to one that bounds $\kappa_2(Z)$ (see Proposition 6.2), and κ stands for the ratio of extremal nonzero eigenvalues of the symmetric positive semi-definite matrix $(I-\Pi)A(I-\Pi)$, which can be bounded further by $\kappa(A)$. At first sight, this result may sound unexpected. Nevertheless, the convergence of PCG becomes natural when recalling the equivalence between indefinite preconditioning and the null-space method (cf. [24]).

We shall see in the next section that the problem can be scaled so that the condition $1 \in [\alpha, \beta]$ is satisfied and the bound (8.4) holds.

The typical situation when $1 \notin [\alpha, \beta]$ is the occurrence of breakdown before the residual has dropped below the required (sufficiently small) tolerance. Nevertheless, there is a remedy how to avoid the unsuccessful termination of the PCG method. Since the first part of the error $e_k^{(1)}$ converges to zero, in exact arithmetic the computation terminates with the breakdown $(r_k, P^{-1}r_k) = (e_k, Me_k) = 0$ which results in $e_k^{(1)} = 0$. Then using $s_k = Ae_k^{(1)} + Be_k^{(2)} = Be_k^{(2)}$ we can correct the approximate solution (see [19]) as

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} + \begin{pmatrix} e_k^{(1)} \\ e_k^{(2)} \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} + \begin{pmatrix} 0 \\ (B^TB)^{-1}B^Ts_k \end{pmatrix}. \tag{8.5}$$

In particular, this shows that checking the residual norm may be misleading, and may lead to pessimistic expectation on the obtained approximation. In the lack of better knowledge of estimates on the error norm (cf. for instance [9]), it is clearly desirable that this correction step be avoided and that the method terminate successfully on both components of the error. This is discussed in the next section.

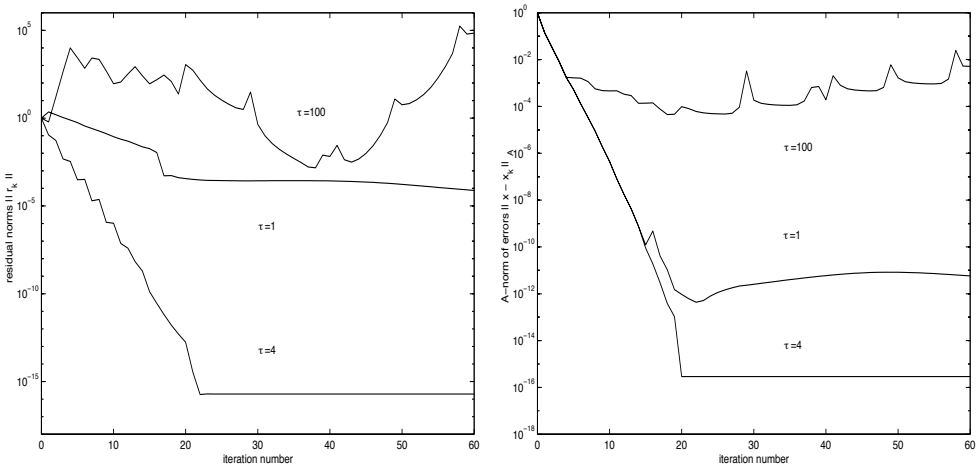

FIG. 9.1. *Residual norm (left) and error M–norm (right) history of PCG for various values of τ.*

9. Conjugate gradients and diagonal scaling. In the previous section we have shown that while the M–norm of the PCG error must necessarily decrease, the 2–norm of the residual may not decrease at the same rate as the iteration proceeds, or may not converge at all. The rate of convergence, when measured by the norm of residual, strongly depends on the value of the PCG polynomial at the eigenvalue 1, which may be outside the interval that contains the nonzero eigenvalues of $(I - \Pi)A(I-\Pi)$. This problem, however, can be easily overcome by pre–scaling the original coefficient matrix as described below.

If A is symmetric positive definite and D =diag(A), then the eigenvalues of the matrix $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ are either all ones or are contained in a nontrivial interval $[\alpha, \beta]$ strictly including the unit value[†]. However, this fact does not necessarily imply that

[†]This can be shown in a number of ways. Martin Gutknecht proposed the following: *The $n \times n$*

also the spectral interval of the projected matrix $(I-\Pi)D^{-\frac{1}{2}}AD^{-\frac{1}{2}}(I-\Pi)$ includes the unit value, although this is usually the case. Standard theory only ensures that the nonzero eigenvalues of $(I-\Pi)D^{-\frac{1}{2}}AD^{-\frac{1}{2}}(I-\Pi)$ are contained in a subset of $[\alpha,\beta]$, which may or may not include the unit value. Nevertheless, this problem can be solved by means of a simple scalar scaling of A as follows. Let $v\in\mathbb{R}^n$ be any nonzero vector with unit norm such that $v=(I-\Pi)v$ and let $\chi=v^TAv>0$. Then the smallest interval containing the nonzero eigenvalues of the matrix $F_\chi=(I-\Pi)(\chi^{-1}A)(I-\Pi)$ includes the unit value. Indeed, let $\lambda_{min},\lambda_{max}$ be the nonzero smallest and largest eigenvalues of F_χ, respectively. Then

$$\lambda_{max}=\max_{0\neq x}\frac{x^TF_\chi x}{x^Tx}\geq v^TF_\chi v=1$$

and, using standard variational arguments (see e.g. [8]),

$$\lambda_{min}=\min_{0\neq x\perp\text{span}\{B\}}\frac{x^TF_\chi x}{x^Tx}\leq v^TF_\chi v=1.$$

In terms of the quantities in the original problem, the theory above is recovered by simply rescaling the saddle point problem as

$$D_\chi^{-\frac{1}{2}}MD_\chi^{-\frac{1}{2}}\hat{t}=D_\chi^{-\frac{1}{2}}b\qquad\hat{t}=D_\chi^{-\frac{1}{2}}t\qquad D_\chi=\text{diag}(\chi I,\chi^{-1}I),$$

and then using the corresponding indefinite preconditioner. It should be also mentioned that scaling with D_χ does not affect the contraint matrix B.

As a general implementation rule, we suggest to first scale A by its diagonal, which in several applications makes the preconditioned problem independent of the mesh parameter, and then employ the additional scaling matrix D_χ to ensure that the preconditioned CG method converges at the expected convergence rate.

In the following examples we show the behavior of PCG with respect to the location of the interval $[\alpha,\beta]$. We emphasize that analogous results could be obtained by using simplified BiCG with right preconditioning.

We consider the following setting: $n=25, s=5$,

$$A=\text{tridiag}(1,\underline{4},1)\in\mathbb{R}^{n\times n}\quad B=\text{rand}(n,s)\quad f=\text{rand}(n,1),\quad g=0.$$

The nonzero eigenvalues of $(I-\Pi)A(I-\Pi)$ are in the interval $[\alpha,\beta]=[2.1268,5.8275]$. We consider two diagonal scalings of A that provide matrices $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ whose spectral interval is shifted. Since the diagonal of A is constant, this simply amounts to considering matrices of the form $D=\tau I$. We shall denote by $[\alpha_\tau,\beta_\tau]$ the corresponding eigenvalue interval. Clearly, $\tau=1$ gives the original matrix, while $\tau>1$ shifts $[\alpha_\tau,\beta_\tau]$ towards zero. The value $\tau=4$ is optimal in the sense that it corresponds to the choice $D=$diag(A). No scaling with χ, as described in section 9, is carried out.

In Figure 9.1 (Left) the exact residual norm history of PCG for $\tau=1,4,100$ is reported, while Figure 9.1 (Right) shows the corresponding M–norm of the error. Both residual and error (M)–norms fall to machine precision level with the prescribed asymptotic convergence behavior for $\tau=4$. For $\tau=1$, $[\alpha_1,\beta_1]=[2.1268,5.8275]$ and

matrix $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ has trace n. Since the trace is the sum of its (positive) eigenvalues, then either all eigenvalues are equal to 1, or there exist at least one eigenvalue less than 1 and one eigenvalue which is greater than 1, that is $1\in]\alpha,\beta[$.

the residual norm does not decrease at the same rate as the M–norm of the error, since the residual polynomial might not be small at the unit value. This is clearly observed in the figures. It should be mentioned, however, that we do not expect the residual to grow unboundlessly because of the constraint $\phi(0) = 1$ (cf. e.g. Proposition 8.1). Mitigating effects on the residual norm (cf. Figure 9.1 (Left)) no longer take place for $\tau = 100$, since $\alpha_\tau < \beta_\tau < 1$ and $\phi(1)$ may be substantially larger than one. Surprisingly, complete failure of the method is reported for $\tau = 100$, where at least the A–norm of the error should converge to zero, in exact arithmetic. In fact, finite precision arithmetic computation is responsible for this failure. The behavior of PCG on the indefinite problem in finite precision arithmetic is discussed in section 10.

We next show the same kind of behavior on a real application problem. We consider the potential fluid flow problem in a rectangular domain with homogeneous Neumann conditions and Dirichlet conditions imposed on a part of the boundary [20, 22]. General prismatic discretization of the domain is used and a mixed–hybrid finite element formulation is considered [16, 20]. The lowest order Raviart-Thomas finite element approximation to the problem leads to the symmetric indefinite system of the form (1.1) of total dimension 868. The positive definite block A represents a discrete form of the tensor in the Darcy law describing the physical properties (hydraulic permeability) of the porous medium in the domain. The off–diagonal block B describes the geometry of the domain and the fulfillment of Neumann boundary conditions. The dependence of the spectrum of M on the discretization parameter (mesh size) was analyzed in [21] and the rate of convergence of unpreconditioned MINRES method applied to the indefinite system (1.1) was estimated. The eigenvalues of the matrix $(I - \Pi)A(I - \Pi)$ are contained in $[4 \cdot 10^{-3}, 8 \cdot 10^{-2}]$. In Figure 9.2 we report the convergence history of preconditioned CG and GMRES on the unscaled (left plot) and scaled (right plot) problems. Scaling with χ was not necessary on this problem. The reported residual is the true residual given by the current approximate solution. In Figure 9.2 (Left), the GMRES residual norm converges towards its maximum accuracy with the expected asymptotic slope. The spectral distribution explains the divergence of the CG residual, while the M–norm of the CG error converges to its final accuracy after few iterations. The connection between the behavior of the error and the residual of PCG in finite precision arithmetic is discussed in detail in the next section.

Figure 9.2 (Right) confirms that scaling optimally cures the problem, and maximum accuracy is obtained with both methods.

10. Behavior in finite precision arithmetic. We have experimentally observed in the previous section that round–off may take large part in the finite precision behavior of PCG on the indefinite problem. In this section we discuss the maximum attainable accuracy of the preconditioned CG scheme, measured in terms of the A–norm of the error $x - \bar{x}_k$, where $\bar{x}_k$ is the first part of the approximate solution $\bar{t}_k$ computed in finite precision arithmetic. Computed quantities will be identified by upper bar.

It is well known that there is a limitation in the accuracy of the approximate solution computed via the CG recurrence. Namely, the residual norm obtained directly from the computed iterates $\bar{t}_k$ as $||b - M\bar{t}_k||$ cannot decrease below a certain level, which is called the maximum attainable accuracy of the scheme. Using the theory of Greenbaum and after slight modification of Theorem 1 given in [12] we can formulate the following proposition.

PROPOSITION 10.1. *Assuming that the initial residual r_0 is computed exactly,*

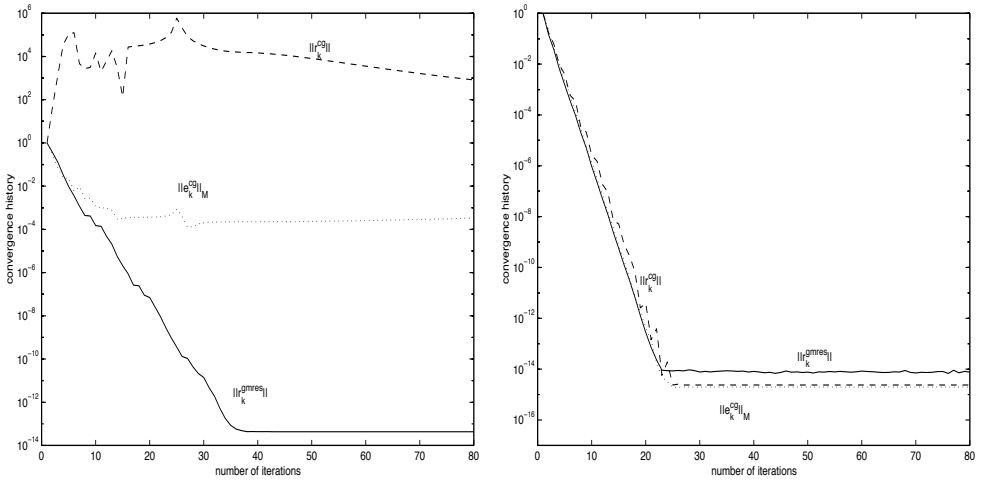

FIG. 9.2. *Convergence history of PCG and PGMRES on real application problem. Left: original problem; Right: scaled problem.*

the gap between the true residual $b - M\bar{t}_k$ *and the recursively computed residual* $\bar{r}_k$ *can be bounded as*

$$(10.1)\ ||(b - M\bar{t}_k) - \bar{r}_k|| \leq \varepsilon k||M|| \left(||t|| + (6 + 2\mu(n+m)^{1/2}) \max_{j=0,\ldots,k} ||t - \bar{t}_j|| \right),$$

where μ *stands for the maximum number of nonzeros per row in the matrix* M*, and* ε *denotes the machine precision.*

If we assume that the method converges, we can expect that even the norm of the recursively computed residual $\bar{r}_k$ will decrease far below the machine precision level. Consequently, from the bound for the gap we receive the bound for the maximum attainable accuracy level (measured by the true residual norm) which depends on the largest error norm during the whole process of convergence. It was shown by Greenbaum ([12]) that the growth in the norm does not occur for the error or residual norm minimizing methods (with respect to any positive–definite norm). Unfortunately, since in our case the "M–norm" of the error is minimized, and M is indefinite and does not induce a norm, these results cannot be applied directly to our scheme. The right–hand side of (10.1) can be further bounded in terms of the residual norm using $\varepsilon||t - \bar{t}_j|| \leq \varepsilon||M^{-1}||\,||\bar{r}_j|| + O(\varepsilon^2)$, therefore the bound on $||(b - M\bar{t}_k) - \bar{r}_k||$ depends in general on the maximum residual norm during the iteration steps $j = 0, \ldots, k$. We assume, however, that our problem is well–scaled and that the norm of the computed residual $||\bar{r}_k||$ converges far below machine precision. Under these assumptions, convergence is usually monotonic or nearly monotonic. Thus the maximum attainable accuracy, measured by the true residual norm, can be assumed to be at the level $p(k, \mu, n+m)\varepsilon\kappa(M)||\bar{r}_0||$, which is the level one gets for the standard CG algorithm (see [12]). Here, the term $p(k, \mu, n+m)$ stands for a low degree polynomial in k, μ and $n+m$ and it does not play an important role in our considerations. The fact that the numerical behavior of this scheme depends heavily on the size of computed residuals is already known and it was analyzed in [11], where iterative refinement techniques and other residual update strategies were proposed in order to reduce the errors caused by large residuals. For the A–norm of the error $x - \bar{x}_k$ the following bound holds in our case.

PROPOSITION 10.2. *The A–norm of the error $x - \bar{x}_k$ can be bounded as*

$$||x - \bar{x}_k||_A \leq \gamma_1\gamma_2||\Pi(x - \bar{x}_k)|| + \gamma_3||(I - \Pi)A(I - \Pi)(x - \bar{x}_k)||$$

where $\gamma_1 = ||A||^{1/2}, \gamma_2 = (1 + (\kappa(A))^{1/2})$ *and* $\gamma_3 = ||A^{-1}||^{1/2}$.

Proof. Since $\Pi x = 0$, the A–norm of the error $\bar{e}_k = x - \bar{x}_k$ can be written as

$$\begin{aligned} ||\bar{e}_k||_A^2 &= (\Pi A\bar{e}_k, \Pi\bar{e}_k) + ((I - \Pi)A\bar{e}_k, (I - \Pi)\bar{e}_k) \\ &= (A\bar{e}_k, \Pi\bar{e}_k) + ((I - \Pi)A(I - \Pi)\bar{e}_k, \bar{e}_k) + ((I - \Pi)A\Pi\bar{e}_k, \bar{e}_k). \end{aligned} \tag{10.2}$$

Using some manipulation we get

$$\begin{aligned} ||\bar{e}_k||_A^2 &\leq ||A\bar{e}_k||||\Pi\bar{e}_k|| + ||(I - \Pi)A(I - \Pi)\bar{e}_k||||\bar{e}_k|| + ||(I - \Pi)A\Pi\bar{e}_k||||\bar{e}_k|| \\ &\leq ||A||^{1/2}||\bar{e}_k||_A||\Pi\bar{e}_k|| + ||(I - \Pi)A(I - \Pi)\bar{e}_k||||A^{-1}||^{1/2}||\bar{e}_k||_A \\ &\quad + ||I - \Pi||\,||A||\,||\Pi\bar{e}_k||\,||A^{-1}||^{1/2}\,||\bar{e}_k||_A \end{aligned}$$

and the result follows. □

The first term on the right–hand side should be zero in exact arithmetic and it describes the departure of the computed iterate $\bar{x}_k$ from the null-space of B^T. The second term will converge to zero in exact arithmetic (see Proposition 8.1). By using a small modification of the proof in Proposition 10.2 we can get from (10.2) the following statement.

COROLLARY 10.3. *The A–norm of the error $x - \bar{x}_k$ can be bounded as*

$$||x - \bar{x}_k||_A \leq \gamma_1||\Pi(x - \bar{x}_k)|| + \gamma_3||(I - \Pi)(f - A\bar{x}_k - B\bar{y}_k)||. \tag{10.3}$$

The bound on $||x - \bar{x}_k||_A$ consists of two parts the first of which is related to the departure of $\bar{x}_k$ from the null-space of B^T; the second part is related to the projection of the residual $f - A\bar{x}_k - B\bar{y}_k$ onto $\mathcal{N}(B^T)$. We next give some computable bounds for the A–norm of the error in terms of the bound in Proposition 10.1. In exact arithmetic the second part of the residual $r_k = [s_k; 0]$ should be zero. For the computed vector $\bar{r}_k = [\bar{s}_k^{(1)}; \bar{s}_k^{(2)}]$ this is no longer the case and we have

$$(b - M\bar{t}_k) - \bar{r}_k = \begin{bmatrix} (f - A\bar{x}_k - B\bar{y}_k) - \bar{s}_k^{(1)} \\ B^T(x - \bar{x}_k) - \bar{s}_k^{(2)} \end{bmatrix}. \tag{10.4}$$

From Proposition 10.1 it also follows that the residual $\bar{s}_k^{(1)}$ is a good approximation to the true one $f - A\bar{x}_k - B\bar{y}_k$, provided we are above the limiting accuracy level given by the bound (10.1). This implies that the second term in the right–hand side of (10.3) is close to the computable quantity $||(I - \Pi)\bar{s}_k^{(1)}||$. For the first term in (10.3) we can write

$$||\Pi(x - \bar{x}_k)|| \leq \delta_1||B^T(x - \bar{x}_k)||, \tag{10.5}$$

where $\delta_1 = (\sigma_{min}(B))^{-1}$. It immediately follows from (10.4) that

$$||B^T(x - \bar{x}_k) - \bar{s}_k^{(2)}|| \leq ||(b - M\bar{t}_k) - \bar{r}_k|| \tag{10.6}$$

and, again, provided that the residuals are above the level of maximum attainable accuracy, the second part of the updated residual $\bar{s}_k^{(2)}$ is a good approximation to the

quantity $B^T(x-\bar{x}_k)$. So we can use (10.5) to obtain the bound in terms of $||\bar{s}_k^{(2)}||$ which is also easily computable. The A–norm of the error $x-\bar{x}_k$ is thus well–approximated (from above) by the maximum between the quantities $\gamma_1\delta_1||\bar{s}_k^{(2)}||$ and $\gamma_3||(I-\Pi)\bar{s}_k^{(1)}||$. In the case when the recursively computed residual $\bar{r}_k$ converges ultimately below the machine precision level, then $||(I-\Pi)\bar{s}_k^{(1)}||$ and $||\bar{s}_k^{(2)}||$ also converge below the machine precision level and the quantities $B^T(x-\bar{x}_k)$ in (10.6) and $f-A\bar{x}_k-B\bar{y}_k$ in Corollary 10.3 can be bounded using Proposition 10.1. As a consequence, we obtain a bound on the level of maximum attainable accuracy of the method, measured by $||x-\bar{x}_k||_A$. On the other hand, if the system is badly scaled so that its unit eigenvalue is at the exterior of the spectral interval of $(I-\Pi)A(I-\Pi)$, then the quantities $\gamma_3||(I-\Pi)\bar{s}_k^{(1)}||$ and $\gamma_1\delta_1||\bar{s}_k^{(2)}||$ may remain at a much higher level. This leads to low accuracy of the computed $\bar{x}_k$ which is reflected in large $||x-\bar{x}_k||_A$. We can summarize the considerations above by saying that a proper scaling not only ensures the convergence of the residual norm in exact arithmetic, but also allows us to obtain a satisfactory level of maximum attainable accuracy of the computed approximate solution $\bar{x}_k$.

We have already noticed at the end of section 8 that, in the general case, y_k may not converge to the solution y at all, so one can hardly expect some accuracy in the computed approximate solution $\bar{y}_k$, unless the correction step (8.5) is used. Nevertheless, assuming that the problem is well scaled, y_k does converge and further considerations based on Proposition 10.1 can be made and the accuracy of the computed second block $\bar{y}_k$ can be estimated. Indeed, we have

$$||B(y-\bar{y}_k)|| \leq ||f-A\bar{x}_k-B\bar{y}_k|| + ||A(x-\bar{x}_k)||. \tag{10.7}$$

Considering (10.7) and using the inequality $||A(x-\bar{x}_k)|| \leq ||A||^{1/2}||x-\bar{x}_k||_A$ we get the bound on $||y-\bar{y}_k||$

$$||y-\bar{y}_k|| \leq \delta_1 \left(||f-A\bar{x}_k-B\bar{y}_k|| + \gamma_1||x-\bar{x}_k||_A\right). \tag{10.8}$$

Considering the inequality from (10.4)

$$||(f-A\bar{x}_k-B\bar{y}_k)-\bar{s}_k^{(1)}|| \leq ||(b-M\bar{t}_k)-\bar{r}_k|| \tag{10.9}$$

and assuming further that $||\bar{r}_k||$ is beyond the level of machine precision, the first term in (10.8) can be bounded using Proposition 10.1. Together with the bounds on $||x-\bar{x}_k||_A$, this gives us the level of maximum attainable accuracy of the scheme, measured by $||y-\bar{y}_k||$. In the case the residual $\bar{r}_k$ is above its level of maximum attainable accuracy, the norm $||y-\bar{y}_k||$ is well approximated by the maximum between the quantities $\delta_1||\bar{s}_k^{(1)}||$, $(\gamma_2-1)\delta_1||(I-\Pi)\bar{s}_k^{(1)}||$ and $\gamma_1\delta_1^2||\bar{s}_k^{(2)}||$.

In the following we report numerical experiments on the finite arithmetic behavior of the computed quantities generated during the CG recurrence. We consider the same 30×30 example as before and solve the system scaled by τ, for $\tau = 100, 4, 1$. In Figure 10.1 the true residual norm of PCG for $\tau = 1$ is reported (solid line). Since the method does not converge to the high accuracy level on the original problem, the solid line coincides fully with the norm of the updated residual vector $||\bar{r}_k||$. The norm of the departure from $\mathcal{N}(B^T)$, measured by $||\Pi\bar{x}_k||$ (dotted line), remains close to the level of machine precision and is well–approximated by the term $\gamma_1\delta_1||\bar{s}_k^{(2)}||$ (not reported in the plot).

It is immediately clear from Figure 10.1 that the error $||x-\bar{x}_k||_A$ (dashed line) is determined by the second term of the bound (10.3) in Corollary 10.3. Due to the

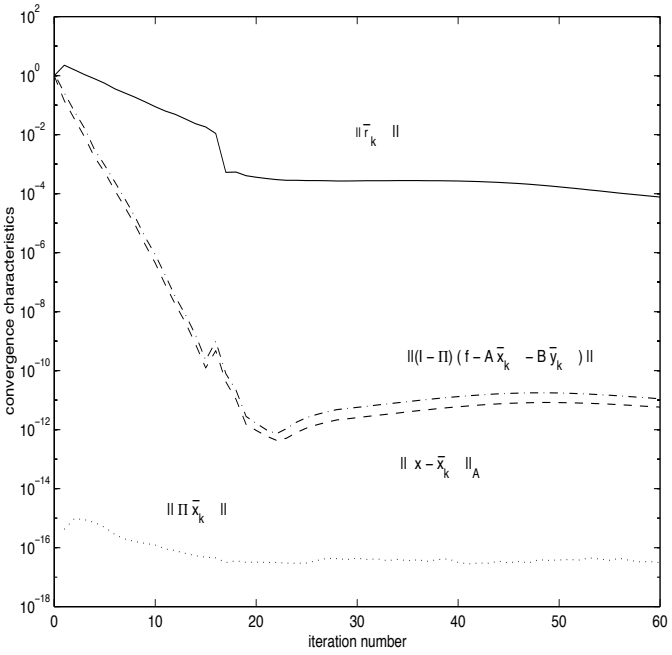

FIG. 10.1. *Behavior in finite precision arithmetic. Original problem.*

poor convergence of the residual norm, the quantity $||(I-\Pi)(f-A\bar{x}_k-B\bar{y}_k)||$ (dash–dotted line) coincides with $||(I-\Pi)\bar{s}_k^{(1)}||$. It is clear that in the case $\tau = 1$ this term determines the level of accuracy of the computed approximate solution $\bar{x}_k$.

Figure 10.2 shows the same quantities as Figure 10.1 for $\tau = 100$. For $\tau = 100$, the problem becomes even more badly scaled and the residual norm (either of the true or updated residual - their difference is almost invisible) does not converge at all. Moreover, the departure from $\mathcal{N}(B^T)$ is no longer close to the level of machine precision and actually reaches the level of $||x-\bar{x}_k||_A$. This indicates that for very irregular residual behavior (or, in other words, very badly scaled problems) the first term in (10.3) may play an important role.

Figure 10.3 illustrates the behavior of PCG on the problem with optimal scaling $\tau = 4$. Both norms of the true and updated residual converge almost monotonically; while the true residual norm remains stagnating at machine precision level, the quantity $||\bar{r}_k||$ (solid line) converges even far beyond this level. Consequently the terms $||\Pi\bar{x}_k||$ and $||(I-\Pi)(f-A\bar{x}_k-B\bar{y}_k)||$ remain close to machine precision leading to a very accurate (whole) approximate solution $\bar{t}_k$.

11. Conclusions. Indefinite preconditioning has recently shown to be particularly attractive for solving saddle point problems arising from constrained nonlinear programming. Short–term recurrence nonsymmetric methods are applicable, at a cost comparable to that of symmetric solvers. However, numerical experience indicated that convergence was not always guaranteed (cf. [18, 19] for the indefinite CG method).

In this paper we have shown that there is a tight connection between short–term recurrence methods such as BiCG and the indefinite CG method used in [18]. More

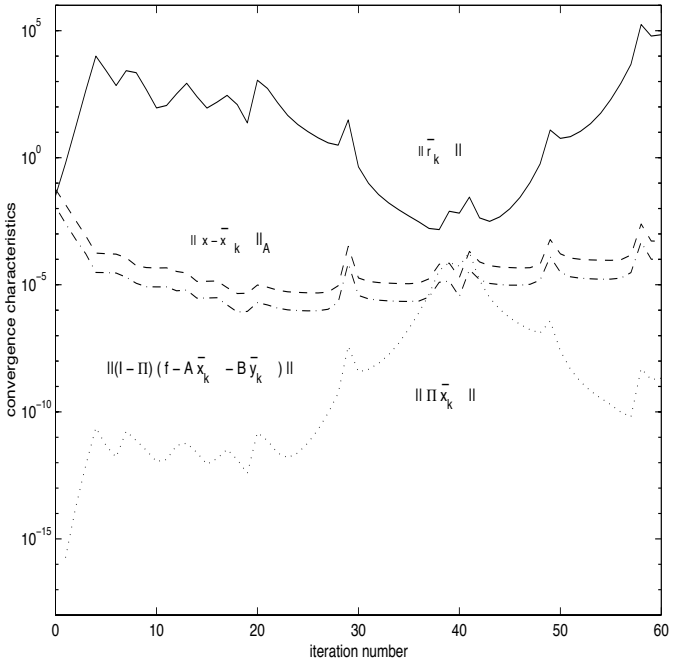

FIG. 10.2. *Behavior in finite precision arithmetic. Diagonal scaling* $D = \tau I$ *with* $\tau = 100$.

precisely, they are equivalent for a special choice of auxiliary vector, with which BiCG simplifies. Moreover, we have proved that the convergence of preconditioned CG strongly depends on the location of the unit eigenvalue with respect to the rest of the spectrum, so that if 1 is not an extreme eigenvalue, then convergence of preconditioned CG on the indefinite problem is guaranteed. We have shown that this condition is not restrictive, as it can be easily satisfied by scaling the original matrix. Scaling turns out to be fundamental also for the stability of the method.

In spite of its indefiniteness, we have thus shown that the scaled problem can be efficiently solved using CG with indefinite preconditioning at the same asymptotic convergence rate as that given by preconditioned CG on a positive definite problem (cf. (8.4)).

Finally, it is interesting to note that numerical experiments related to the work in [23] showed that similar considerations with respect to the behavior of PCG seem to also hold for the problem

$$\begin{bmatrix} A & B \\ B^T & -C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}$$

with C positive semidefinite, $\mathcal{N}(B) \neq \emptyset$ and $C + B^T B$ positive definite, which includes a wider class of problems than that treated in this paper.

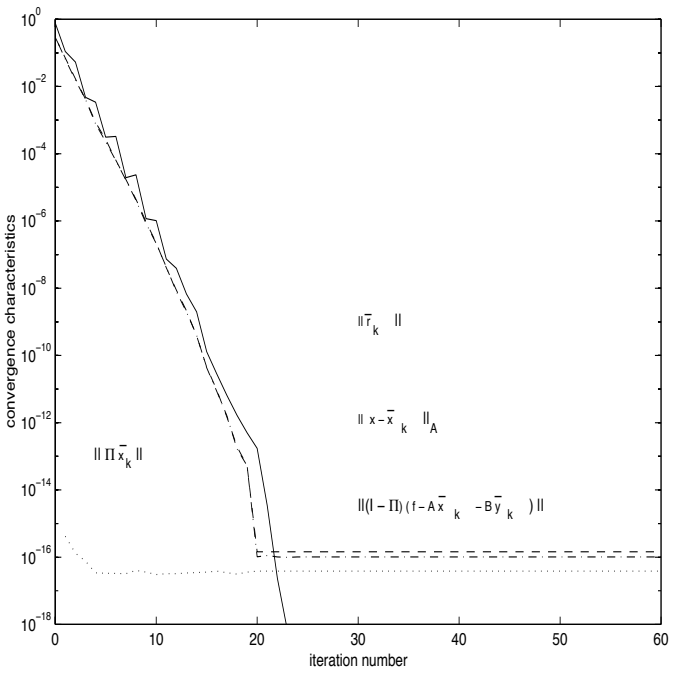

FIG. 10.3. *Behavior in finite precision arithmetic. Diagonal scaling $D = \tau I$ with $\tau = 4$.*

Acknowledgments. The authors would like to thank L. Lukšan for enlightening discussions on [18] and [19]. We thank M. Gutknecht for fruitful conversations.

REFERENCES

[1] D. N. ARNOLD, R. S. FALK, AND R. WINTHER, *Preconditioning in $h(div)$ and applications*, Math. Comp., 66 (1997), pp. 957–984.

[2] J. H. BRAMBLE AND J. E. PASCIAK, *A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems*, Math. Comp., 50 (1988), pp. 1–17.

[3] H. C. ELMAN, D. J. SILVESTER, AND A. J. WATHEN, *Iterative methods for problems in computational fluid dynamics*, in Iterative Methods in Scientific Computing, C. T. Chan and G. Golub, eds., Springer-Verlag, 1997.

[4] R. E. EWING, R. D. LAZAROV, P. LU, AND P. S. VASSILEVSKI, *Preconditioning indefinite systems arising from mixed finite element discretization of second-order elliptic problems*, in Notes in Mathematics, Springer, 1990, pp. 28–43.

[5] R. W. FREUND AND N. M. NACHTIGAL, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.

[6] ———, *Software for simplified Lanczos and QMR algorithms*, Applied Numerical Mathematics, 19 (1995), pp. 319–341.

[7] P. E. GILL, W. MURRAY, D. B. PONCELEON, AND M. A. SAUNDERS, *Preconditioners for indefinite systems arising in optimization*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 292–311.

[8] G. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, 3rd ed., 1996.

[9] G. H. GOLUB AND G. MEURANT, *Matrices, moments and quadrature II; how to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705.

[10] 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 (1998), pp. 530–539.

[11] N. I. M. GOULD, M. E. HRIBAR, AND J. NOCEDAL, *On the solution of equality constrained quadratic programming problems arising in optimization*, Tech. Rep. RAL-TR-1998-069, Rutherford Appleton Lab., Oxfordshire, 1998.

[12] A. Greenbaum, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–551.
[13] ———, *Iterative methods for solving linear systems*, SIAM, PA, 1997.
[14] M. Gutknecht and M. Rozložník, *Residual smoothing techniques: Do they improve the limiting accuracy of iterative solvers?*, Tech. Rep. 99-22, SAM ETH Zurich, 1999. BIT, To appear.
[15] K. C. Jea and D. M. Young, *On the simplification of generalized conjugate-gradient methods for nonsymmetrizable linear systems*, Lin. Alg. Appl., 52–53 (1983), pp. 399–417.
[16] E. F. Kaasschieter and A. J. M. Huijben, *Mixed-hybrid finite elements and streamline computation for the potential flow problem*, Numerical Methods for Partial Diff. Equations, 8 (1992), pp. 221 –266.
[17] C. Keller, N. I. M. Gould, and A. J. Wathen, *Constraint preconditioning for indefinite linear systems*, Tech. Rep. NA 99/01, Oxford University Computing Laboratory, 1999. SIMAX, To appear.
[18] L. Lukšan and J. Vlček, *Indefinitely preconditioned inexact Newton method for large sparse equality constrained non-linear programming problems*, Num. Linear Algebra and Appl., 5 (1998), pp. 219–247.
[19] ———, *Conjugate gradient methods for saddle point systems*, in Proc. of the 13-th Summer School on Software and Algorithms of Numerical Mathematics, I. Marek, ed., Nečtiny, Czech Republic, Sept. 6-10, 1999, pp. 223–230.
[20] J. Maryška, M. Rozložník, and M. Tůma, *Mixed-hybrid finite element approximation of the potential fluid flow problem*, J. Comput. Appl. Math., 63 (1995), pp. 383 –392.
[21] ———, *The potential fluid flow problem and the convergence rate of the minimal residual method*, Numer. Linear Algebra with Appl., 3 (1996), pp. 525 –542.
[22] ———, *Schur complement systems in the mixed-hybrid finite element approximation of the potential fluid flow problem*, tech. rep., 2000. SISC, To appear.
[23] I. Perugia and V. Simoncini, *Block–diagonal and indefinite symmetric preconditioners for mixed finite element formulations*, Num. Linear Algebra with Appl., (2000), pp. 1–32. To appear.
[24] I. Perugia, V. Simoncini, and M. Arioli, *Linear algebra methods in a mixed approximation of magnetostatic problems*, SIAM J. Sci. Comput., 21 (1999), pp. 1085–1101.
[25] T. Rusten and R. Winther, *A preconditioned iterative method for saddle point problems*, SIAM J. Matrix Analysis and Appl., 13 (1992), pp. 887–904.
[26] ———, *Substructure preconditioners for elliptic saddle point problems*, Math. Comp., 60 (1993), pp. 23–48.
[27] Y. Saad, *Iterative methods for sparse linear systems*, The PWS Publishing Company, 1996.
[28] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (July 1986), pp. 856–869.
[29] D. Silvester and A. Wathen, *Fast iterative solution of stabilized Stokes systems part II: using general block preconditioners*, SIAM J. Numer. Anal., 31 (1994), pp. 1352–1367.
[30] L. Zhou and H. F. Walker, *Residual smoothing techniques for iterative methods*, SIAM J. Sci. Comput., 15 (1994), pp. 297–312.

Research Reports

No.	Authors	Title
00-08	M. Rozložník, V. Simoncini	Short-term recurrences for indefinite preconditioning of saddle point problems
00-07	P. Houston, C. Schwab, E. Süli	Discontinuous *hp*-Finite Element Methods for Advection-Diffusion Problems
00-06	W.P. Petersen	Estimation of Weak Lensing Parameters by Stochastic Integration
00-05	M.H. Gutknecht	A Matrix Interpretation of the Extended Euclidean Algorithm
00-04	M.J. Grote	Nonreflecting Boundary Conditions for Time Dependent Wave Propagation
00-03	M.H. Gutknecht	On Lanczos-type methods for Wilson fermions
00-02	R. Sperb, R. Strebel	An alternative to Ewald sums. Part 3: Implementation and results
00-01	T. Werder, K. Gerdes, D. Schötzau, C. Schwab	*hp* Discontinuous Galerkin Time Stepping for Parabolic Problems
99-26	J. Waldvogel	Jost Bürgi and the Discovery of the Logarithms
99-25	H. Brunner, Q. Hu, Q. Lin	Geometric meshes in collocation methods for Volterra integral equations with proportional time delays
99-24	D. Schötzau, Schwab	An *hp* a-priori error analysis of the DG time-stepping method for initial value problems
99-23	R. Sperb	Optimal sub- or supersolutions in reaction-diffusion problems
99-22	M.H. Gutknecht, M. Rozložník	Residual smoothing techniques: do they improve the limiting accuracy of iterative solvers?
99-21	M.H. Gutknecht, Z. Strakoš	Accuracy of Two Three-term and Three Two-term Recurrences for Krylov Space Solvers
99-20	M.H. Gutknecht, K.J. Ressel	Look-Ahead Procedures for Lanczos-Type Product Methods Based on Three-Term Lanczos Recurrences
99-19	M. Grote	Nonreflecting Boundary Conditions For Elastodynamic Scattering
99-18	J. Pitkäranta, A.-M. Matache, C. Schwab	Fourier mode analysis of layers in shallow shell deformations
99-17	K. Gerdes, J.M. Melenk, D. Schötzau, C. Schwab	The *hp*-Version of the Streamline Diffusion Finite Element Method in Two Space Dimensions
99-16	R. Klees, M. van Gelderen, C. Lage, C. Schwab	Fast numerical solution of the linearized Molodensky problem