Numerische Mathematik manuscript No.
(will be inserted by the editor)

On nonsymmetric saddle point matrices that allow conjugate gradient iterations*

Jörg Liesen[1] **, Beresford N. Parlett[2]

[1] Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany, e-mail: liesen@math.tu-berlin.de

[2] Department of Mathematics, University of California, Berkeley, CA 94720-3840, USA, e-mail: parlett@math.berkeley.edu

Received: date / Revised version: date

Summary Linear systems in saddle point form are usually highly indefinite, which often slows down iterative solvers such as Krylov subspace methods. It has been noted by several authors that negating the second block row of a symmetric indefinite saddle point matrix leads to a nonsymmetric matrix $\mathcal{A}$ whose spectrum is entirely contained in the right half plane. In this paper we study conditions so that $\mathcal{A}$ is diagonalizable with a real and positive spectrum. These conditions are based on necessary and sufficient conditions for positive definiteness of a certain bilinear form, with respect to which $\mathcal{A}$ is symmetric. In case the latter conditions are satisfied, there exists a well defined conjugate gradient (CG) method for solving linear systems with $\mathcal{A}$. We give an efficient implementation of this method, discuss practical issues such as error bounds, and present numerical experiments.

Key words saddle point problem – eigenvalues – conjugate gradient method – Stokes problem

Mathematics Subject Classification (1991): 65F15, 65N22, 65F50

* In memory of Gene Golub (1932-2007), our wonderful friend and colleague, who had a great interest in the conjugate gradient method and the numerical solution of saddle point problems.

** The work of this author was supported by the Emmy Noether-Program and the Heisenberg-Program of the Deutsche Forschungsgemeinschaft.

1 Introduction

Many applications in science and engineering require solving large linear algebraic systems in saddle point form; see [4] for an extensive survey. A typical system matrix is of the form

$$\begin{bmatrix} A & B^T \\ B & -C \end{bmatrix}, \tag{1.1}$$

where $A = A^T \in \mathbb{R}^{n\times n}$ is positive definite ($A > 0$), $B \in \mathbb{R}^{m\times n}$ has rank $r \leq m \leq n$, and $C = C^T \in \mathbb{R}^{m\times m}$ is positive semidefinite ($C \geq 0$). The matrix in (1.1) is congruent to the block diagonal matrix $[A\ 0;\ 0\ S]$, where $S = -(C + BA^{-1}B^T)$, so that $-S = -S^T \geq 0$. Hence this matrix is indefinite, with n positive and $\text{rank}(S)$ negative eigenvalues. Typically, $\text{rank}(S)$ is close to m and therefore, unless m is very small, the matrix in (1.1) is highly indefinite. This feature often slows down iterative solvers such as Krylov subspace methods. For a discussion of the convergence properties of these methods in case of indefinite problems we refer to [4, Section 9.2].

It has been noted by several authors (see [4, p. 23] for references) that, for solving linear algebraic systems, there is no harm in negating the second block row in (1.1), and using as system matrix

$$\mathcal{A} \equiv \begin{bmatrix} A & B^T \\ -B & C \end{bmatrix}. \tag{1.2}$$

Symmetry has been abandoned but what, if anything, has been gained? First, the useful decomposition of a real matrix into its symmetric and skew-symmetric parts is right before our eyes. Standard results, using Rayleigh quotients, show that the eigenvalues of $\mathcal{A}$ lie in a box with real parts in the spectral interval of $A \oplus C$ and imaginary parts in the interval $[0,\ \|B\|]$. Thus, $\mathcal{A}$ is *positive semistable*, i.e. all its eigenvalues have nonnegative real parts. Moreover, it can be shown that there exists a "magic" hidden bilinear form with respect to which $\mathcal{A}$ is symmetric. As shown in this paper, this bilinear form is a proper inner product, if and only if the spectra of A and C are separated, and the norm of B is small enough. When this is satisfied, $\mathcal{A}$ is diagonalizable with nonnegative real eigenvalues. Thus, there exists a conjugate gradient (CG) method for solving linear systems with $\mathcal{A}$.

Some of the theoretical results in this paper, particularly those concerning the definiteness of the bilinear form, generalize previous work in [8] and [5]. In these papers the focus is on cases with $C = 0$. Moreover, unlike in [8,5], we provide an implementation of a CG method for solving linear systems with $\mathcal{A}$. In this context we discuss some subtleties concerning the choice of the inner product for constructing a well defined CG method for $\mathcal{A}$. Recent related work on CG methods in non-standard inner products and with particular emphasis on saddle point problems is reported in [18,17]. In [18], some known examples for such non-standard inner products including those in [8,5] are surveyed, and a general strategy for generating new examples from known ones is discussed (so-called combination preconditioning). The paper [17] focusses on a particular example, namely the Bramble-Pasciak preconditioner [6], and presents some new variants of this approach.

Note that transforming a saddle point system with a matrix (1.1) into one with a matrix of the form (1.2) may be considered a form of *preconditioning*. This is the viewpoint taken, e.g., in [18]. It should not be expected, however, that just by negating the second block row the speed of convergence of a Krylov subspace method can be improved significantly. For instance, the matrices (1.1) and (1.2) have the same singular values, and hence a possible ill-conditioning of (1.1) is not overcome by the transformation into (1.2). To really improve the speed of convergence, one must combine the negation of the second block row with a suitable preconditioning technique, for example (inexact) block diagonal preconditioning; see [4, Section 10.1.1] for a detailed discussion of this technique. We believe that the general idea of negation is of interest, since the behavior of Krylov subspace methods for problems with positive eigenvalues is better understood than for problems with eigenvalues on both sides of the origin; see [4, Section 9.2] for more details. Problems that are better understood can potentially be preconditioned more effectively. The construction of such preconditioning techniques, which typically should be tuned to the application at hand, is beyond the scope of this paper.

The paper is organized as follows. In Sections 2 and 3 we analyze properties of the matrix $\mathcal{A}$ and the bilinear form with respect to which $\mathcal{A}$ is symmetric. In Section 4 we construct a CG method for solving linear systems with $\mathcal{A}$, discuss when this method is well defined, and provide an efficient implementation. In Section 5 we discuss some practical issues in the context of our CG method, including the conditioning of the CG inner product matrix and error bounds. In Section 6 we present some numerical experiments that show the effectiveness of the method.

Notation. The bilinear form defined by a (real) symmetric matrix G is denoted by $(u,v)_G \equiv v^T G u$. In case G is positive definite, this bilinear form is an inner product, and the associated norm is given by $\|u\|_G \equiv (u,u)_G^{1/2}$. In case $G = I$, i.e. the Euclidean inner product and norm, we skip the index and simply write (u,v) and

$\|u\|$. A matrix M is called symmetric with respect to a (real) symmetric matrix G, or shortly *G-symmetric*, if GM is symmetric, or, equivalently, $(Mu,v)_G = (u,Mv)_G$ for all u,v. The matrix M is called *G-definite*, if GM is definite, i.e. $(GMu,u) = (Mu,u)_G \neq 0$ for all $u \neq 0$. Of course, if $G = I$, we simply write symmetric and definite as usual.

2 Analysis of the matrix $\mathcal{A}$

We consider a matrix $\mathcal{A} \in \mathbb{R}^{(n+m)\times(n+m)}$ as in (1.2), with symmetric positive definite $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}$ of rank $r \leq m \leq n$, and symmetric positive semidefinite $C \in \mathbb{R}^{m\times m}$. As shown in [4, Theorem 3.6], the matrix $\mathcal{A}$ is *positive semistable*, meaning that all its eigenvalues have nonnegative real parts, and in case B has full rank m, $\mathcal{A}$ is *positive stable*, i.e. all its eigenvalues have positive real parts.

The following lemma leads, in a natural manner, to the special inner product mentioned above.

Lemma 2.1 *Let the matrix*

$$\mathcal{J} \;\equiv\; \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} \tag{2.1}$$

be conformally partitioned with $\mathcal{A}$. Then

(1) $\mathcal{A}$ is $\mathcal{J}$-symmetric, i.e. $\mathcal{J}\mathcal{A} = \mathcal{A}^T\mathcal{J} = (\mathcal{J}\mathcal{A})^T$,

and, for any polynomial p,

(2a) $p(\mathcal{A})$ is $\mathcal{J}$-symmetric, i.e. $\mathcal{J}p(\mathcal{A}) = p(\mathcal{A}^T)\mathcal{J} = (\mathcal{J}p(\mathcal{A}))^T$, and
(2b) $\mathcal{A}$ is $\mathcal{J}p(\mathcal{A})$-symmetric, i.e.
$(\mathcal{J}p(\mathcal{A}))\mathcal{A} = \mathcal{A}^T(p(\mathcal{A}^T)\mathcal{J}) = (\mathcal{J}p(\mathcal{A})\mathcal{A})^T$.

Proof Item (1) follows by straightforward computation. Using (1), we see that

$$\mathcal{J}\mathcal{A}^2 = (\mathcal{J}\mathcal{A})\mathcal{A} = (\mathcal{A}^T\mathcal{J})\mathcal{A} = \mathcal{A}^T(\mathcal{J}\mathcal{A}) = (\mathcal{A}^T)^2\mathcal{J},$$

which implies $(2a)$ by induction. Using items (1) and $(2a)$,

$$(\mathcal{J}p(\mathcal{A}))\mathcal{A} = (\mathcal{J}\mathcal{A})p(\mathcal{A}) = \mathcal{A}^T(\mathcal{J}p(\mathcal{A})) = \mathcal{A}^T(p(\mathcal{A}^T)\mathcal{J}),$$

which proves $(2b)$. □

Item (1) in Lemma 2.1 shows that $\mathcal{A}$ is symmetric with respect to the symmetric *indefinite* matrix $\mathcal{J}$. This fact has been exploited before, e.g., in [5] (also see the discussion in [4, p. 25]). Our goal here is to determine whether there exists a symmetric *positive definite* matrix with respect to which $\mathcal{A}$ is symmetric. Our starting point is the relation (2*b*) in Lemma 2.1. It shows that $\mathcal{A}$ is symmetric with respect to any matrix of the form $\mathcal{J}p(\mathcal{A})$, where p is any polynomial. As shown by item (2*a*), any matrix of the form $\mathcal{J}p(\mathcal{A})$ is itself symmetric. Therefore it suffices to show conditions under which $\mathcal{J}p(\mathcal{A})$ is positive definite. Obviously, when p is of degree zero, this cannot be satisfied. Choice of a p of degree exceeding one seems too expensive. Our ansatz will therefore be a polynomial p of degree one. Without loss of generality we may take the leading coefficient of p to be equal to one, and write our polynomial in the form $p(\zeta) = \zeta - \gamma$, for some yet to be determined parameter $\gamma \in \mathbb{R}$. Hence we ask: When is

$$\mathcal{M}(\gamma) \;\equiv\; \mathcal{J}p(\mathcal{A}) \;=\; \mathcal{J}(\mathcal{A}-\gamma I) \;=\; \begin{bmatrix} A-\gamma I & B^T \\ B & \gamma I - C \end{bmatrix} \tag{2.2}$$

a positive definite matrix? The complete answer to this question is given in the following theorem.

Theorem 2.2 *The symmetric matrix $\mathcal{M}(\gamma)$ is positive definite if and only if*

$$\lambda_{\min}(A) > \gamma > \lambda_{\max}(C)\,, \tag{2.3}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalue, respectively, and

$$\left\| (\gamma I - C)^{-1/2} B (A-\gamma I)^{-1/2} \right\| < 1\,. \tag{2.4}$$

Proof It is easy to see that $\mathcal{M}(\gamma) > 0$ holds only if $A - \gamma I > 0$, or, equivalently, $\lambda_{\min}(A) > \gamma$. If this holds, $\mathcal{M}(\gamma)$ is congruent to $(A-\gamma I) \oplus S$, where

$$S = (\gamma I - C) - B(A-\gamma I)^{-1}B^T.$$

Therefore, $\mathcal{M}(\gamma)$ is positive definite, if, and only if, $\lambda_{\min}(A) > \gamma$ and $S > 0$. The second inequality is equivalent to

$$\gamma I - C > B(A-\gamma I)^{-1}B^T.$$

The matrix on the right hand side is positive semidefinite, which implies $\gamma I - C > 0$, or, equivalently, $\gamma > \lambda_{\max}(C)$. Finally, $\gamma I - C > B(A-\gamma I)^{-1}B^T$ is equivalent to

$$I > \left((\gamma I - C)^{-1/2} B (A-\gamma I)^{-1/2}\right)\left((\gamma I - C)^{-1/2} B (A-\gamma I)^{-1/2}\right)^T,$$

which in turn is equivalent to (2.4). □

3 Sufficient conditions

From Theorem 2.2 we can derive some useful *sufficient* conditions that make $\mathcal{M}(\gamma)$ positive definite.

Corollary 3.1 *The matrix $\mathcal{M}(\gamma)$ is symmetric positive definite when (2.3) holds, and, in addition,*

$$\|B\|^2 \;<\; (\lambda_{\min}(A) - \gamma)\,(\gamma - \lambda_{\max}(C))\,. \tag{3.1}$$

For $\gamma = \widehat{\gamma} \equiv \frac{1}{2}(\lambda_{\min}(A) + \lambda_{\max}(C))$, the right hand side of (3.1) is maximal, and (3.1) reduces to

$$2\|B\| \;<\; \lambda_{\min}(A) - \lambda_{\max}(C)\,. \tag{3.2}$$

Proof A simple computation shows that

$$\begin{aligned}&\left\|(\gamma I - C)^{-1/2} B (A-\gamma I)^{-1/2}\right\| \;\le\\ &\|(\gamma I - C)^{-1/2}\|\,\|B\|\,\|(A-\gamma I)^{-1/2}\| \;=\\ &(\gamma - \lambda_{\max}(C))^{-1/2}\,\|B\|\,(\lambda_{\min}(A)-\gamma)^{-1/2}\,.\end{aligned}$$

Hence $\mathcal{M}(\gamma) > 0$, if (2.3) holds and the right hand side is less than one, which is equivalent to (3.1). The second part follows from another simple computation. □

The conditions for positive definiteness of $\mathcal{M}(\gamma)$ yield sufficient conditions so that $\mathcal{A}$ is diagonalizable with a positive real spectrum.

Corollary 3.2 *If there exists a $\gamma \in \mathbb{R}$ so that $\mathcal{M}(\gamma)$ is positive definite, then $\mathcal{A}$ has a nonnegative real spectrum and a complete set of eigenvectors that are orthonormal with respect to the inner product defined by $\mathcal{M}(\gamma)$. In case B has full rank, the spectrum of $\mathcal{A}$ is real and positive.*

Proof We know that the matrix $\mathcal{A}$ is $\mathcal{M}(\gamma)$-symmetric. If $\mathcal{M}(\gamma)$ is positive definite, then $\mathcal{M}(\gamma)$ defines an inner product, and hence $\mathcal{A}$ has real eigenvalues and a complete set of eigenvectors that are orthonormal with respect to this inner product (see, e.g., [9, Chapter IX]). Moreover, $\mathcal{A}$ is known to be positive semistable (positive stable if B has full rank), so that its eigenvalues indeed must be real and nonnegative (real and positive). □

We now show that any saddle point matrix as specified in (1.1) with full rank B and $\lambda_{\min}(A) > \lambda_{\max}(C)$ can be *scaled* to give a matrix of the form (1.2) which is diagonalizable with real and positive eigenvalues. Let $\alpha > 0$ be a real parameter and consider

$$\begin{bmatrix} \alpha I & 0 \\ 0 & -I \end{bmatrix} \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} \alpha I & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} \alpha^2 A & \alpha B^T \\ -\alpha B & C \end{bmatrix} \equiv \mathcal{A}(\alpha) \,. \tag{3.3}$$

Corollaries 3.1 and 3.2 show that this matrix $\mathcal{A}(\alpha)$ is diagonalizable with real and positive eigenvalues whenever

$$2\,\alpha\,\|B\| \;<\; \alpha^2\,\lambda_{\min}(A) - \lambda_{\max}(C)\,. \tag{3.4}$$

This condition can always be satisfied by choosing α large enough. Of course, if $\|B\| \gg \lambda_{\min}(A)$, then α must be chosen very large, and this might cause numerical problems, or may be incompatible with the application at hand. For example, in case of an equality-constrained optimization problem, where B represents the constraints (see [4, Section 1.1]), scaling with a very large α transforms the original saddle point matrix (in the limit $\alpha \to \infty$) into one that represents an unconstrained problem.

Our results above generalize several results that appeared in the literature: Fischer et al. [8] consider $\mathcal{A}$ with $A = \alpha I > 0$ and $C = 0$. They show that $\mathcal{A}$ has a nonnegative real spectrum when $2\|B\| < \alpha$, which in this case is equivalent to (3.2). Benzi and Simoncini [5, Section 3] consider (in our notation) a matrix $\mathcal{M}(\gamma)$ with $A = A^T > 0$ and $C = 0$. In [5, Proposition 3.1] they show that $\mathcal{M}(\widehat{\gamma})$ is positive definite when

$$4\lambda_{\max}(BA^{-1}B^T) < \lambda_{\min}(A). \tag{3.5}$$

Note that

$$\begin{aligned} 4\lambda_{\max}(BA^{-1}B^T) &= 4\|BA^{-1/2}\|^2 \\ &\le 4\,\|B\|^2\,\|A^{-1/2}\|^2 \;=\; \frac{4\|B\|^2}{\lambda_{\min}(A)}\,. \end{aligned} \tag{3.6}$$

Condition (3.2) with $C = 0$ is equivalent to $4\|B\|^2/\lambda_{\min}(A) < \lambda_{\min}(A)$. Thus (3.2) implies (3.5).

One of the conditions that make $\mathcal{M}(\gamma)$ symmetric positive definite is (2.3), namely $\lambda_{\min}(A) > \gamma > \lambda_{\max}(C)$. Hence, Corollary 3.2 only yields a condition for a (nonnegative) real spectrum of $\mathcal{A}$ when the spectra of A and C are separated. This separation appears to be essential, as seen by considering a 2×2 matrix of the form

$$\begin{bmatrix} \alpha & \beta \\ -\beta & \eta \end{bmatrix}, \quad \alpha > 0\,,\ \eta > 0\,.$$

This matrix has real eigenvalues (necessarily nonnegative) only when its discriminant $(\alpha - \eta)^2 - 4\beta^2$ is nonnegative. Thus β must vanish if $\alpha = \eta$, and, by extension to $\mathcal{A}$, if the spectra of A and C overlap, i.e. $\lambda_{\max}(C) \geq \lambda_{\min}(A)$, and $B \neq 0$, it is most unlikely that the spectrum of $\mathcal{A}$ is real.

The discriminant condition we just have derived turns out to be a special case of [5, Proposition 2.5]. In that result the matrix $\mathcal{A}$ is assumed to be of the form (1.2) with $C = \eta I > 0$. If $\mathcal{A} \in \mathbb{R}^{2\times 2}$, then the condition that both eigenvalues of $\mathcal{A}$ are real is $(\alpha + \eta)^2 \geq 4(\beta^2 + \alpha\eta)$, or, equivalently, $(\alpha - \eta)^2 - 4\beta^2 \geq 0$.

Next let us consider a more representative matrix

$$\mathcal{A} = \left[\begin{array}{ccc|cc} 1 & 0 & 0 & \beta & 0 \\ 0 & 2 & 0 & 0 & \beta \\ 0 & 0 & 3 & 0 & 0 \\ \hline -\beta & 0 & 0 & 2\eta & -\eta \\ 0 & -\beta & 0 & -\eta & 2\eta \end{array}\right], \quad \beta \neq 0\,, \quad \eta \geq 0\,.$$

Here $\lambda_{\min}(A) = 1$, B has full rank, $\|B\| = |\beta|$, and $\lambda_{\max}(C) = 3\eta$. By (3.2), $\mathcal{A}$ is diagonalizable with a positive and real spectrum when

$$2|\beta| < 1 - 3\eta\,.$$

For $\eta = 1/12$ the above condition is $|\beta| < 3/8 = 0.375$, while MATLAB [13] shows that $\mathcal{A}$ has five distinct real and positive eigenvalues for $|\beta|$ as large as 0.405. For larger $|\beta|$, $\mathcal{A}$ has nonreal eigenvalues. On the other hand, if we choose $\beta = 1/2$, then the above condition is not satisfied for any $\eta \geq 0$, and indeed a MATLAB computation reveals that the matrix $\mathcal{A}$ is not diagonalizable for $\eta = 0$, and has nonreal eigenvalues for $\eta > 0$. Therefore, the sufficient condition (3.2) cannot be relaxed, in general.

4 Construction of a conjugate gradient (CG) method

In this section we construct a CG method for solving linear systems with the matrix $\mathcal{A}$.

4.1 The (generic) CG method

As a starting point we consider the following (generic) statement of the CG method based on a given inner product $(u, v)_G = v^T G u$ for solving a linear system of the form $Mx = b$, cf. [2, (4.1a)–(4.1g)]. An alternative approach is discussed in Remark 4.5.

Algorithm 4.1 (The Conjugate Gradient (CG) method)
Input: System matrix M, right hand side b, inner product matrix G, initial guess x_0.
Initialize: $r_0 = b - Mx_0$, $p_0 = r_0$.
For $i = 0, 1, \ldots$ until convergence:

$$\alpha_i = \frac{(x - x_i, p_i)_G}{(p_i, p_i)_G} \tag{4.1}$$

$$x_{i+1} = x_i + \alpha_i p_i \tag{4.2}$$

$$r_{i+1} = r_i - \alpha_i M p_i \tag{4.3}$$

$$\beta_{i+1} = -\frac{(r_{i+1}, p_i)_G}{(p_i, p_i)_G} \tag{4.4}$$

$$p_{i+1} = r_{i+1} + \beta_{i+1} p_i \tag{4.5}$$

In case $M = M^T > 0$ and $G = M$, this algorithm corresponds to the classical Hestenes and Stiefel implementation of CG [11]. Its requirements and most important properties are summarized in the following result (see [2] for proofs and further details).

Theorem 4.1 *Suppose that the matrix G is symmetric positive definite and hence defines an inner product. If the matrix M is G-symmetric and G-definite, then (until convergence) the CG method stated in Algorithm 4.1 has the following properties:*

(1) The iterates and residuals satisfy

$$x_i \in x_0 + K_i(M, r_0) \quad \textit{and} \quad r_i = b - Mx_i \in r_0 + MK_i(M, r_0),$$

where $K_i(M, r_0) \equiv \mathrm{span}\{r_0, Mr_0, \ldots, M^{i-1}r_0\}$ is the ith Krylov subspace generated by M and r_0.

(2) The direction vectors $p_i \in r_0 + MK_i(M, r_0)$ *satisfy*

$$(p_i, p_j)_G = 0 \quad \textit{for} \quad i \neq j.$$

(3) The error vector $x - x_{i+1} \in (x - x_0) + K_{i+1}(M, r_0)$ *satisfies*

$$(x - x_{i+1}, u)_G = 0 \quad \textit{for all} \quad u \in \operatorname{span}\{p_0, \dots, p_i\}.$$

(4) The method is optimal in the G*-norm, i.e.*

$$\begin{aligned}\|x - x_{i+1}\|_G &= \min_{u \in x_0 + K_{i+1}(M, r_0)} \|x - u\|_G \\ &= \min_{p \in \pi_{i+1}} \|p(M)(x - x_0)\|_G,\end{aligned}$$

where π_{i+1} *denotes the set of polynomials of degree at most* $i+1$ *and with value one at the origin.*

When the assumptions of Theorem 4.1 on M and G are satisfied, and hence the assertions (1)–(4) hold, we say that Algorithm 4.1 is *well defined* for M and G.

Remark 4.2 In case M is G-symmetric and G-*semi*definite (i.e. M is singular), Algorithm 4.1 may still be well defined. However, this will not hold globally for any right hand side b, but just for $b \in \operatorname{Range}(M)$, so that a solution of the linear system exists. Details of the CG method for singular matrices are well discussed in [3, Section 11.2.8].

Note that the numerator of α_i in (4.1) contains the unknown solution vector x. Therefore an essential practical requirement for the CG method, that is not mentioned in Theorem 4.1, is that the inner product matrix G must be chosen so that the scalar α_i is *computable.* In the classical CG method of Hestenes and Stiefel, the system matrix M is assumed symmetric positive definite, $G = M$, and then

$$(x - x_i, p_i)_G = (x - x_i, p_i)_M = (M(x - x_i), p_i) = (r_i, p_i),$$

which is a computable quantity.

4.2 A CG method for $\mathcal{A}$

From now on we assume that we are given a linear system of the form $\mathcal{A}x = b$ where $\mathcal{A}$ is as in (1.2) with $A = A^T > 0$, full rank B, and $C = C^T \geq 0$. Generalizations of the following algorithms and results may be easily obtained for rank deficient B, cf. Remark 4.2, but here we have opted for a clean rather than the most general presentation.

A CG method for solving this system requires a symmetric positive definite matrix defining an inner product. To make this CG method well defined, the matrix $\mathcal{A}$ and the inner product matrix should satisfy the assumptions of Theorem 4.1. Moreover, the inner product matrix must be chosen so that α_i is computable. The latter requirement is *not* satisfied by the matrix $\mathcal{M}(\gamma)$ in (2.2), since for this matrix we cannot evaluate the quantity $(x - x_i, p_i)_{\mathcal{M}(\gamma)}$ unless we know the solution vector x. Therefore, even if $\mathcal{M}(\gamma)$ is positive definite, this matrix cannot be used in Algorithm 4.1 to give a computable CG method for solving $\mathcal{A}x = b$.

To make α_i computable, our choice for the inner product will be the matrix $\mathcal{M}(\gamma)\mathcal{A}$. From item (2*b*) in Lemma 2.1 we know that this matrix is symmetric. The following result gives a sufficient condition when this matrix is positive definite.

Lemma 4.3 *If the (symmetric) matrix $\mathcal{M}(\gamma)$ is positive definite, then for all nonnegative integers k the matrix $\mathcal{M}(\gamma)\mathcal{A}^k$ is also symmetric positive definite.*

Proof Item (2*b*) in Lemma 2.1 shows that $\mathcal{M}(\gamma)\mathcal{A} = \mathcal{A}^T\mathcal{M}(\gamma)$ and therefore, by induction,

$$\mathcal{M}(\gamma)\mathcal{A}^k = (\mathcal{A}^T)^k\mathcal{M}(\gamma) = (\mathcal{M}(\gamma)\mathcal{A}^k)^T$$

for all $k \geq 0$, i.e. $\mathcal{M}(\gamma)\mathcal{A}^k$ is symmetric. Moreover,

$$\mathcal{A}^k = \mathcal{M}(\gamma)^{-1}(\mathcal{A}^T)^k\mathcal{M}(\gamma),$$

which means that $\mathcal{A}^k$ is normal with respect to the symmetric positive definite matrix $\mathcal{M}(\gamma)$. A result of Givens [10, Theorem 2] implies that the $\mathcal{M}(\gamma)$-field of values of $\mathcal{A}^k$, i.e. the set of all

$$\frac{(\mathcal{A}^k u, u)_{\mathcal{M}(\gamma)}}{(u, u)_{\mathcal{M}(\gamma)}}, \quad u \neq 0,$$

is equal to the convex hull of the eigenvalues of $\mathcal{A}^k$. Since all eigenvalues of $\mathcal{A}^k$ are real and positive (cf. Corollary 3.2),

$$\begin{aligned} 0 < \lambda_{\min}(\mathcal{A}^k) \ &\leq \ \frac{(\mathcal{A}^k u, u)_{\mathcal{M}(\gamma)}}{(u, u)_{\mathcal{M}(\gamma)}} \ = \ \frac{(\mathcal{M}(\gamma)\mathcal{A}^k u, u)}{(u, u)_{\mathcal{M}(\gamma)}} \\ &\leq \lambda_{\max}(\mathcal{A}^k) \quad \text{for all } u \neq 0. \end{aligned}$$

Therefore $(\mathcal{M}(\gamma)\mathcal{A}^k u, u) > 0$ for all $u \neq 0$, which shows that $\mathcal{M}(\gamma)\mathcal{A}^k$ is symmetric positive definite. □

We now derive expressions for α_i and β_{i+1} in Algorithm 4.1, in case $\mathcal{M}(\gamma)\mathcal{A}$ is positive definite and chosen as the inner product matrix.

Lemma 4.4 *Suppose that the (symmetric) matrix $\mathcal{M}(\gamma)$ is positive definite. Then Algorithm 4.1 is well defined for $M = \mathcal{A}$ and $G = \mathcal{M}(\gamma)\mathcal{A}$, and (until convergence) the scalars α_i and β_{i+1}, can be computed as*

$$\alpha_i = \frac{(r_i, r_i)_{\mathcal{M}(\gamma)}}{(\mathcal{A}p_i, p_i)_{\mathcal{M}(\gamma)}}\,, \tag{4.6}$$

$$\beta_{i+1} = \frac{(r_{i+1}, r_{i+1})_{\mathcal{M}(\gamma)}}{(r_i, r_i)_{\mathcal{M}(\gamma)}}\,. \tag{4.7}$$

Proof Since $\mathcal{M}(\gamma)$ is positive definite, $\mathcal{M}(\gamma)\mathcal{A}^k$ is symmetric positive definite for all $k \geq 0$ (cf. Lemma 4.3). In particular, $\mathcal{M}(\gamma)\mathcal{A}$ is symmetric positive definite and hence defines an inner product. Moreover, $\mathcal{M}(\gamma)\mathcal{A}^2 = (\mathcal{M}(\gamma)\mathcal{A})\mathcal{A}$ is symmetric positive definite, which means that $\mathcal{A}$ is both $\mathcal{M}(\gamma)\mathcal{A}$-symmetric and $\mathcal{M}(\gamma)\mathcal{A}$-definite. Therefore, the assumptions of Theorem 4.1 are satisfied, and Algorithm 4.1 with $M = \mathcal{A}$ and $G = \mathcal{M}(\gamma)\mathcal{A}$ is well defined.

It is easy to see that for $i \geq 0$ the denominator of α_i in (4.1) is equal to $(\mathcal{A}p_i, p_i)_{\mathcal{M}(\gamma)}$. For $i = 0$, the numerator of α_i is equal to

$$(x - x_0, p_0)_{\mathcal{M}(\gamma)\mathcal{A}} \;=\; (\mathcal{A}(x - x_0), r_0)_{\mathcal{M}(\gamma)} \;=\; (r_0, r_0)_{\mathcal{M}(\gamma)},$$

showing that (4.6) holds for $i = 0$. For $i \geq 1$ we use (4.5) and the orthogonality relation in item (3) of Theorem 4.1 to obtain

$$\begin{aligned}(x - x_i, p_i)_{\mathcal{M}(\gamma)\mathcal{A}} &= (x - x_i, r_i + \beta_i p_{i-1})_{\mathcal{M}(\gamma)\mathcal{A}} \\ &= (x - x_i, r_i)_{\mathcal{M}(\gamma)\mathcal{A}} + \beta_i (x - x_i, p_{i-1})_{\mathcal{M}(\gamma)\mathcal{A}} \\ &= (\mathcal{A}(x - x_i), r_i)_{\mathcal{M}(\gamma)} \\ &= (r_i, r_i)_{\mathcal{M}(\gamma)},\end{aligned}$$

which proves (4.6) for $i \geq 1$. Note that $\alpha_i \neq 0$ for $i \geq 0$.

Next, we consider the numerator of β_{i+1}, $i \geq 0$, in (4.4). Here we use (4.3) and again the orthogonality relation in item (3) of Theorem 4.1 to obtain

$$\begin{aligned}(r_{i+1}, p_i)_{\mathcal{M}(\gamma)\mathcal{A}} &= (x - x_{i+1}, \mathcal{A}p_i)_{\mathcal{M}(\gamma)\mathcal{A}} \\ &= \big(x - x_{i+1}, \alpha_i^{-1}(r_i - r_{i+1})\big)_{\mathcal{M}(\gamma)\mathcal{A}} \\ &= \alpha_i^{-1}(x - x_{i+1}, r_i)_{\mathcal{M}(\gamma)\mathcal{A}} - \alpha_i^{-1}(x - x_{i+1}, r_{i+1})_{\mathcal{M}(\gamma)\mathcal{A}} \\ &= -\alpha_i^{-1}(r_{i+1}, r_{i+1})_{\mathcal{M}(\gamma)}.\end{aligned}$$

Therefore,

$$\beta_{i+1} = -\frac{(r_{i+1}, p_i)_{\mathcal{M}(\gamma)\mathcal{A}}}{(p_i, p_i)_{\mathcal{M}(\gamma)\mathcal{A}}} = \alpha_i^{-1} \frac{(r_{i+1}, r_{i+1})_{\mathcal{M}(\gamma)}}{(\mathcal{A}p_i, p_i)_{\mathcal{M}(\gamma)}} = \frac{(r_{i+1}, r_{i+1})_{\mathcal{M}(\gamma)}}{(r_i, r_i)_{\mathcal{M}(\gamma)}},$$

which completes the proof. □

Remark 4.5 The formulas for the scalars α_i and β_{i+1} in Lemma 4.4 are identical to those in the classical CG method of Hestenes and Stiefel [11], with the exception that there the inner product is the standard Euclidean one. Hence, a CG method for $\mathcal{A}$ can also be derived by starting with the Hestenes and Stiefel CG method, with the Euclidean inner product replaced by the $\mathcal{M}(\gamma)$-inner product, and then showing conditions when this algorithm is well defined. In this approach, the subtle point that both $\mathcal{M}(\gamma)$ and $\mathcal{M}(\gamma)\mathcal{A}$ must be positive definite to make the method well defined is easily overlooked, because the formulas (4.6) and (4.7) seem to suggest that positive definiteness of $\mathcal{M}(\gamma)$ is sufficient.

4.3 An efficient implementation

In step i of Algorithm 4.1 with $M = \mathcal{A}$ and $G = \mathcal{M}(\gamma)\mathcal{A}$, we can compute the scalars α_i and β_{i+1} as shown in (4.6) and (4.7), respectively. We will now show how to replace the $\mathcal{M}(\gamma)$-inner products by $\mathcal{J}$-bilinearforms. Since $\mathcal{M}(\gamma) = \mathcal{J}\mathcal{A} - \gamma\mathcal{J}$, cf. (2.2), we have for all vectors $u, v \in \mathbb{R}^{n+m}$,

$$(u, v)_{\mathcal{M}(\gamma)} = (\mathcal{A}u, v)_{\mathcal{J}} - \gamma(u, v)_{\mathcal{J}}\,.$$

Therefore,

$$(r_i, r_i)_{\mathcal{M}(\gamma)} = (\mathcal{A}r_i, r_i)_{\mathcal{J}} - \gamma(r_i, r_i)_{\mathcal{J}}, \tag{4.8}$$

and, since $\mathcal{A}$ is $\mathcal{M}(\gamma)$-symmetric,

$$(\mathcal{A}p_i, p_i)_{\mathcal{M}(\gamma)} = (p_i, \mathcal{A}p_i)_{\mathcal{M}(\gamma)} = (\mathcal{A}p_i, \mathcal{A}p_i)_{\mathcal{J}} - \gamma(p_i, \mathcal{A}p_i)_{\mathcal{J}}. \tag{4.9}$$

Hence to compute α_i and β_{i+1}, the main work lies in evaluating the bilinear form $(u, v)_{\mathcal{J}}$, which is not more expensive than evaluating the Euclidean inner product (u, v). Note that we need to have available both $\mathcal{A}r_i$ and $\mathcal{A}p_i$. To avoid the necessity of computing both matrix-vector products in every step, we store two additional vectors, namely $y_i = \mathcal{A}r_i$ and $w_i = \mathcal{A}p_i$. The former is computed by multiplying $\mathcal{A}$

against r_i. The latter is computed via an additional recursion in the following way: Multiplying (4.5) by $\mathcal{A}$ yields

$$\underbrace{\mathcal{A}p_{i+1}}_{=w_{i+1}} = \underbrace{\mathcal{A}r_{i+1}}_{=y_{i+1}} + \beta_{i+1} \underbrace{\mathcal{A}p_i}_{=w_i} .$$

The complete algorithm looks as follows.

Algorithm 4.2 (CG method for $\mathcal{A}$)
Input: System matrix $\mathcal{A}$, right hand side b, real parameter γ, initial guess x_0.
Initialize: $r_0 = b - \mathcal{A}x_0$, $p_0 = r_0$, $y_0 = \mathcal{A}r_0$, $w_0 = y_0$
For $i = 0, 1, \ldots$ until convergence:

$$\alpha_i = \frac{(y_i, r_i)_{\mathcal{J}} - \gamma(r_i, r_i)_{\mathcal{J}}}{(w_i, w_i)_{\mathcal{J}} - \gamma(p_i, w_i)_{\mathcal{J}}} \tag{4.10}$$

$$x_{i+1} = x_i + \alpha_i p_i \tag{4.11}$$

$$r_{i+1} = r_i - \alpha_i w_i \tag{4.12}$$

$$y_{i+1} = \mathcal{A}r_{i+1} \tag{4.13}$$

$$\beta_{i+1} = \frac{(y_{i+1}, r_{i+1})_{\mathcal{J}} - \gamma(r_{i+1}, r_{i+1})_{\mathcal{J}}}{(y_i, r_i)_{\mathcal{J}} - \gamma(r_i, r_i)_{\mathcal{J}}} \tag{4.14}$$

$$p_{i+1} = r_{i+1} + \beta_{i+1} p_i \tag{4.15}$$

$$w_{i+1} = y_{i+1} + \beta_{i+1} w_i \tag{4.16}$$

Algorithm 4.2 requires five vectors of storage: x_i, r_i, p_i, y_i, and w_i are required in step i. Since the denominator of β_{i+1} is equal to the numerator of α_i, this quantity only has to be evaluated once in every step. When these scalars are stored, each step requires four evaluations of the bilinear form $(u, v)_{\mathcal{J}}$, i.e. four dot products. In addition, one matrix-vector product and four vector updates (DAXPY's) have to be performed. The following table compares this cost with the cost of the MINRES algorithm for symmetric indefinite linear systems [14], in the two-term recurrence implementation given in [16, Fig. 1, p. 728]:

	Algorithm 4.2	MINRES from [16]
vectors to be stored	5	8
matrix-vector products	1	1
vector updates	4	5
dot products	4	2

It is possible to implement MINRES using just six vectors of storage, when three-term recurrences are used (see, e.g., [7, Algorithm 6.1]).

In any case, the computational cost of Algorithm 4.2 compares well with the standard MINRES method.

We summarize the theoretical requirements and properties of Algorithm 4.2 in the following result.

Corollary 4.6 *If the (symmetric) matrix $\mathcal{M}(\gamma)$ is positive definite, then for $M = \mathcal{A}$ and $G = \mathcal{M}(\gamma)\mathcal{A}$ the assumptions of Theorem 4.1 are satisfied, and Algorithm 4.2 is a well defined CG method for solving $\mathcal{A}x = b$.*

5 Practical issues

In this section we discuss some practical issues concerning our CG method in Algorithm 4.2.

5.1 The condition number of $\mathcal{M}(\gamma)$

While Algorithm 4.2 is based on the inner product defined by $\mathcal{M}(\gamma)\mathcal{A}$, we actually compute the scalars α_i and β_{i+1} using the inner product defined by $\mathcal{M}(\gamma)$, cf. (4.6) and (4.7). Therefore it is of interest to estimate $\kappa(\mathcal{M}(\gamma))$, the condition number of $\mathcal{M}(\gamma)$, in order to assess the numerical stability of the method. The following result is a generalization of [5, Corollary 3.2].

Lemma 5.1 *Suppose that (2.3) and (3.1) hold, so that the matrix $\mathcal{M}(\gamma)$ is symmetric positive definite. Let $\xi \equiv (\lambda_{\min}(A) - \gamma)(\gamma - \lambda_{\max}(C)) - \|B\|^2$, then*

$$\kappa(\mathcal{M}(\gamma)) \;<\; \frac{4}{\xi}\,(\lambda_{\max}(A) - \gamma)(\lambda_{\min}(A) - \gamma)\,. \tag{5.1}$$

Proof By assumption, the matrix $\mathcal{M}(\gamma)$ permits the factorization

$$\begin{bmatrix} (A-\gamma I)^{1/2} & 0 \\ 0 & (\gamma I - C)^{1/2} \end{bmatrix} \begin{bmatrix} I & X^T \\ X & I \end{bmatrix} \begin{bmatrix} (A-\gamma I)^{1/2} & 0 \\ 0 & (\gamma I - C)^{1/2} \end{bmatrix}, \tag{5.2}$$

where $X \equiv (\gamma I - C)^{-1/2} B (A - \gamma I)^{-1/2}$, so that

$$\|X\| \le \frac{\|B\|}{(\lambda_{\min}(A) - \gamma)^{1/2}(\gamma - \lambda_{\max}(C))^{1/2}}\,.$$

Since (3.1) holds we have $\xi > 0$, and hence the right hand side is less than one, giving $1 + \|X\| < 2$. Moreover, elementary calculations using the definition of ξ show that

$$1 - \|X\|^2 \;\geq\; \frac{\xi}{(\lambda_{\min}(A) - \gamma)\,(\gamma - \lambda_{\max}(C))}\,,$$

which will be used below.

For any congruence $M = FHF^T$,

$$\kappa(M) \;\leq\; \|F\|^2\,\|H\|\,\|F^{-1}\|^2\,\|H^{-1}\| \;=\; \kappa(F^2)\kappa(H)\,.$$

Using this in (5.2) yields

$$\begin{aligned}
\kappa(\mathcal{M}(\gamma)) &\leq \kappa\left(\begin{bmatrix} A - \gamma I & 0 \\ 0 & \gamma I - C \end{bmatrix}\right)\,\kappa\left(\begin{bmatrix} I & X^T \\ X & I \end{bmatrix}\right) \\
&= \frac{\lambda_{\max}(A) - \gamma}{\gamma - \lambda_{\max}(C)}\,\frac{1 + \|X\|}{1 - \|X\|} \\
&= \frac{\lambda_{\max}(A) - \gamma}{\gamma - \lambda_{\max}(C)}\,\frac{(1 + \|X\|)^2}{1 - \|X\|^2} \\
&< 4\,\frac{\lambda_{\max}(A) - \gamma}{\gamma - \lambda_{\max}(C)}\,\frac{(\lambda_{\min}(A) - \gamma)(\gamma - \lambda_{\max}(C))}{\xi}\,,
\end{aligned}$$

which concludes the proof. □

The bound (5.1) indicates a relation between $\kappa(\mathcal{M}(\gamma))$ and the sufficient condition (3.1) for positive definiteness of $\mathcal{M}(\gamma)$: With larger ξ, the bound on the condition number of $\mathcal{M}(\gamma)$ becomes smaller, and vice versa.

The best choice of γ is the one that minimizes $\kappa(\mathcal{M}(\gamma))$, but $\gamma = \widehat{\gamma}$ as in Corollary 3.1 is a more accessible substitute. For this choice, and the corresponding value of $\xi = \widehat{\xi}$, (5.1) implies that

$$\kappa(\mathcal{M}(\widehat{\gamma})) \;<\; \frac{(2\lambda_{\max}(A) - \lambda_{\max}(C))\,(\lambda_{\min}(A) - \lambda_{\max}(C))}{\widehat{\xi}}\,. \qquad (5.3)$$

In the special case $C = 0$, (5.3) simplifies to

$$\kappa(\mathcal{M}(\widehat{\gamma})) \;<\; \frac{2}{\widehat{\xi}}\lambda_{\max}(A)\,\lambda_{\min}(A) \;<\; \frac{8\lambda_{\max}(A)}{\lambda_{\min}(A) - 2\|B\|}\,.$$

In this case, Benzi and Simoncini [5, Corollary 3.2] have estimated $\kappa(\mathcal{M}(\widehat{\gamma}))$ as

$$\kappa(\mathcal{M}(\widehat{\gamma})) \;\approx\; \frac{4\lambda_{\max}(A)}{\lambda_{\min}(A) - 4\lambda_{\max}(BA^{-1}B^T)}\,.$$

From (3.6), it is easy to see that the denominator on the right hand side is bounded from below by $\lambda_{\min}(A) - 2\|B\|$, so that the resulting estimate on the right hand side corresponds to our bound up to a constant factor of two.

5.2 Error bounds

When the matrix $\mathcal{M}(\gamma)$ is positive definite, the CG method in Algorithm 4.2 is optimal in the $(\mathcal{M}(\gamma)\mathcal{A})$-norm, see item (4) in Theorem 4.1. We know that $\mathcal{A}$ is $(\mathcal{M}(\gamma)\mathcal{A})$-symmetric (cf. Lemma 4.3), and hence $\mathcal{A}$ has a complete set of eigenvectors that are orthonormal with respect to the inner product defined by $\mathcal{M}(\gamma)\mathcal{A}$ (cf. the proof of Corollary 3.2). Hence we may write

$$\mathcal{A} \;=\; \mathcal{Y}\Lambda\mathcal{Y}^{-1}, \quad \text{where} \quad \mathcal{Y}^T\left(\mathcal{M}(\gamma)\mathcal{A}\right)\mathcal{Y} \;=\; I\,.$$

Suppose that x_0 is an initial guess for the solution of $\mathcal{A}x = b$, and write the initial error as $x - x_0 = \mathcal{Y}v$, for some vector $v \in \mathbb{R}^{n+m}$. Then the $(\mathcal{M}(\gamma)\mathcal{A})$-norm of the ith error satisfies (cf. item (4) in Theorem 4.1)

$$\begin{aligned} \|x - x_i\|_{\mathcal{M}(\gamma)\mathcal{A}} &= \min_{p\in\pi_i} \|p(\mathcal{A})(x - x_0)\|_{\mathcal{M}(\gamma)\mathcal{A}} \\ &= \min_{p\in\pi_i} \left(v^T\, p(\Lambda)^2\, v\right)^{1/2} \\ &\le \|x - x_0\|_{\mathcal{M}(\gamma)\mathcal{A}} \min_{p\in\pi_i} \max_{\lambda\in\Lambda(\mathcal{A})} |p(\lambda)|\,. \qquad (5.4) \end{aligned}$$

Here $\Lambda(\mathcal{A})$ denotes the (real and positive) spectrum of $\mathcal{A}$. The bound (5.4) and its derivation is completely analogous to the standard convergence bound for the classical CG method in case of a symmetric positive definite system matrix M and error minimization in the M-norm. Estimation of the quantity $\min_{p\in\pi_i} \max_{\lambda\in\Lambda(\mathcal{A})} |p(\lambda)|$ using the eigenvalue distribution of $\mathcal{A}$ has been exhaustively done in the literature, see, e.g., [3, Chapter 13]. The important information given by (5.4) is that when the spectrum of $\mathcal{A}$ is clustered away from the origin, then fast convergence can be expected. This gives some indication on how to choose a preconditioner.

To get a computable estimate on the error that is minimized in every step, we use that $\mathcal{A}$ is $(\mathcal{M}(\gamma)\mathcal{A})$-symmetric and $(\mathcal{M}(\gamma)\mathcal{A})$-definite. In this case [1, Corollary 5.2] applies, and shows that

$$\left(\widehat{\kappa}(\mathcal{A})^{-1}\,\frac{(r_i, r_i)_{\mathcal{M}(\gamma)}}{(b, b)_{\mathcal{M}(\gamma)}}\right)^{1/2} \le \frac{\|x - x_i\|_{\mathcal{M}(\gamma)\mathcal{A}}}{\|x\|_{\mathcal{M}(\gamma)\mathcal{A}}}$$

$$\leq \left(\widehat{\kappa}(\mathcal{A}) \, \frac{(r_i, r_i)_{\mathcal{M}(\gamma)}}{(b, b)_{\mathcal{M}(\gamma)}} \right)^{1/2} , \tag{5.5}$$

where $(r_i, r_i)_{\mathcal{M}(\gamma)}$ is the numerator of α_i (and thus available in every step at no extra cost), cf. (4.6), and

$$\widehat{\kappa}(\mathcal{A}) \;\equiv\; \frac{\max_{\lambda \in \Lambda(\mathcal{A})} \lambda}{\min_{\lambda \in \Lambda(\mathcal{A})} \lambda}.$$

Bendixon's Theorem [12, p. 69] yields

$$\min \{\lambda_{\min}(A), \lambda_{\min}(C)\} \;\leq\; \lambda \;\leq\; \max \{\lambda_{\max}(A), \lambda_{\max}(C)\}$$

for all $\lambda \in \Lambda(\mathcal{A})$, so that

$$\widehat{\kappa}(\mathcal{A}) \;\leq\; \frac{\max \{\lambda_{\max}(A), \lambda_{\max}(C)\}}{\min \{\lambda_{\min}(A), \lambda_{\min}(C)\}} \, .$$

Of course, when C is singular, this estimate is useless. Close estimates for the eigenvalues of $\mathcal{A}$ may be obtained using parameters computed by the CG method itself. This gives a convergence bound that becomes tighter during the run of the method. We will not discuss this approach here, and refer the interested reader to [1, Section 7].

We next relate the $(\mathcal{M}(\gamma)\mathcal{A})$-norm of the error to the Euclidean norm of the residual, which is often used as a stopping criterion for the CG method (even though this quantity is not minimized, and may strongly oscillate during the iteration). Since $\mathcal{M}(\gamma) = \mathcal{J}\mathcal{A} - \gamma\mathcal{J} > 0$, we have $u^T\mathcal{J}\mathcal{A}u - \gamma u^T\mathcal{J}u > 0$, or $-u^T\mathcal{J}\mathcal{A}u < -\gamma u^T\mathcal{J}u$, for all vectors $u \in \mathbb{R}^{n+m}$, so that

$$\begin{aligned}
\|x - x_i\|^2_{\mathcal{M}(\gamma)\mathcal{A}} &= (x - x_i)^T \mathcal{M}(\gamma)\mathcal{A}(x - x_i) \\
&= (x - x_i)^T (\mathcal{A}^T\mathcal{J} - \gamma\mathcal{J})\mathcal{A}(x - x_i) \\
&= r_i^T \mathcal{J} r_i - \gamma (x - x_i)^T \mathcal{J}\mathcal{A}(x - x_i) \\
&< r_i^T \mathcal{J} r_i - \gamma^2 (x - x_i)^T \mathcal{J}(x - x_i) \\
&= (r_i, r_i)_{\mathcal{J}} - \gamma^2 (x - x_i, x - x_i)_{\mathcal{J}} \\
&\leq \|r_i\|^2 + \gamma^2 \|x - x_i\|^2 \\
&= \|r_i\|^2 + \gamma^2 \|\mathcal{A}^{-1} r_i\|^2 \\
&\leq \|r_i\|^2 \left(1 + \frac{\gamma^2}{\sigma^2_{\min}(\mathcal{A})} \right) ,
\end{aligned}$$

where $\sigma_{\min}(\mathcal{A})$ denotes the smallest singular value of $\mathcal{A}$. In particular, for $\gamma = \widehat{\gamma}$ as in Corollary 3.1,

$$\|x - x_i\|_{\mathcal{M}(\widehat{\gamma})\mathcal{A}} \;<\; \|r_i\| \left(1 + \frac{(\lambda_{\min}(A) + \lambda_{\max}(C))^2}{4\sigma^2_{\min}(\mathcal{A})} \right)^{1/2} .$$

We see that if $\mathcal{A}$ is not too ill conditioned, then the Euclidean norm of the residual gives a reasonable bound on the $(\mathcal{M}(\gamma)\mathcal{A})$-norm of the error.

6 Numerical examples

In this section we present results of numerical experiments with test problems generated by the MATLAB [13] package Incompressible Flow Iterative Solution Software (IFISS), version 2.2 [15]. We use the driver `stokes_testproblem` of this code with default options to set up a stabilized discretization of a Stokes equations model problem[1], resulting in a linear system of the form

$$\begin{bmatrix} A & B^T \\ B & -\frac{1}{4}C \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{6.1}$$

where $A = A^T > 0$ is of order $n = 578$, $C = C^T \geq 0$ is of order $m = 256$, and, by construction, $\text{rank}(B) = m - 2 = 254$. To agree with the notation used in [7], we have written out the stabilization parameter $\frac{1}{4}$ explicitly. The system matrix is of order $n + m = 834$ and of rank $n + m - 1 = 833$. However, the system is consistent, and the singularity of the system matrix represents no difficulty for the iterative methods considered here. Other parameters relevant for our context, and computed using MATLAB's `eig` routine, are:

$$\lambda_{\max}(A) = 3.9493, \quad \lambda_{\min}(A) = 0.0764,$$
$$\lambda_{\max}(\tfrac{1}{4}C) = 0.0156, \quad \lambda_{\min}(\tfrac{1}{4}C) = 0.$$

The spectra of A and $\frac{1}{4}C$ are separated, so that (2.3) is satisfied. Since $\|B\| = 0.2476$, the sufficient condition (3.2) is not satisfied. Nevertheless, with

$$\widehat{\gamma} = \tfrac{1}{2}\,(\lambda_{\min}(A) + \lambda_{\max}(\tfrac{1}{4}C)) = 0.0460\,,$$

we set up the matrix $\mathcal{M}(\widehat{\gamma})$, which is positive definite and rather well conditioned:

$$\lambda_{\max}(\mathcal{M}(\widehat{\gamma})) = 3.9191, \;\; \lambda_{\min}(\mathcal{M}(\widehat{\gamma})) = 0.0118, \;\; \kappa(\mathcal{M}(\widehat{\gamma})) = 333.3771$$

[1] The default parameters are: lid driven cavity; cavity type: regularized; grid parameter 4 (16×16 grid); uniform grid; Q_1-P_0 elements; stabilization parameter 1/4; uniform streamlines. See the IFISS user guide or [7, Chapter 5] for a detailed description of this test problem.

(computed using MATLAB's `eig` and `cond`). We negate the second block row in (6.1) to obtain

$$\begin{bmatrix} A & B^T \\ -B & \frac{1}{4}C \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ -g \end{bmatrix}. \tag{6.2}$$

We now apply MATLAB's build-in MINRES algorithm to (6.1) and our Algorithm 4.2 to (6.2) (both with $x_0 = 0$). The resulting convergence characteristics are shown in Fig. 6.1 To compute the error norms, $x^T = [u^T, p^T]$ is obtained by solving the system (6.2) using the MATLAB backslash operator.

Clearly, Algorithm 4.2 is competitive with MINRES, which is optimal for the linear system (6.1) in the sense that it minimizes the Euclidean norm of the residual over the Krylov subspace generated by the system matrix and the right hand side. In fact, the convergence of the Euclidean residual norms of Algorithm 4.2 slightly outperforms those of MINRES. Note, however, that the Euclidean residual norms of Algorithm 4.2 are not monotonically decreasing; they do not satisfy a minimization property. On the other hand, the $(\mathcal{M}(\widehat{\gamma})\mathcal{A})$-norm of the error is monotonically decreasing, and in this example it is very close to the Euclidean residual norm. Moreover, a good estimate of this norm is given by $(r_i, r_i)_{\mathcal{M}(\widehat{\gamma})} \,/\, (b, b)_{\mathcal{M}(\widehat{\gamma})}$, a quantity that is available at no additional cost during the iteration (cf. (5.5)).

To obtain a larger test example we again use `stokes_testproblem`, but this time we choose the grid parameter 6 in IFISS, resulting in a 64×64 grid, and system dimensions $n = 8450$ and $m = 4096$. In this case the computation of the exact eigenvalues is rather expensive, and so we only compute *estimates* in MATLAB:

$$\begin{aligned} &\texttt{normest}(A) = 3.9965, \\ &\lambda_{\min}(A) \approx 0.0048 \ \ \text{(estimated using } \texttt{eigs} \text{ with maxit=20)}, \\ &\texttt{normest}(\tfrac{1}{4}C) = 9.7656e-004, \quad \texttt{normest}(B) = 0.0625. \end{aligned}$$

For these estimates (2.3) is satisfied, but again (3.2) is not. The run of Algorithm 4.2 is based on $\gamma = \frac{1}{2}(0.0048 + \texttt{normest}(\frac{1}{4}C)) = 0.0029$. In Fig. 6.2 we show the convergence characteristics of MINRES and Algorithm 4.2; the notation corresponds to the one of Fig. 6.1. Obviously, the qualitative behavior of the algorithms for both test problems is the same.

7 Concluding remarks

We have considered the idea of negating the second block row in a saddle point system, which leads to an unsymmetric but positive

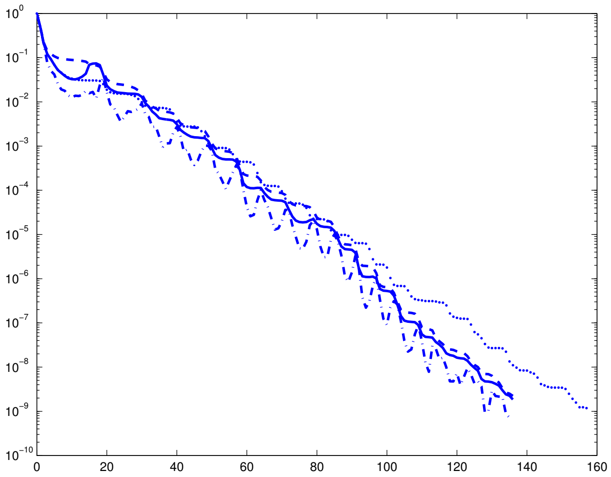

Fig. 6.1. Convergence characteristics for the default `stokes_testproblem`: $\|r_i\|/\|r_0\|$ of MATLAB's MINRES (dotted); $\|r_i\|/\|r_0\|$ (solid), $\|x-x_i\|_{\mathcal{M}(\widehat{\gamma})\mathcal{A}}/\|x-x_0\|_{\mathcal{M}(\widehat{\gamma})\mathcal{A}}$ (dashed), and $(r_i,r_i)_{\mathcal{M}(\widehat{\gamma})}/(b,b)_{\mathcal{M}(\widehat{\gamma})}$ (dashed-dotted) of Algorithm 4.2.

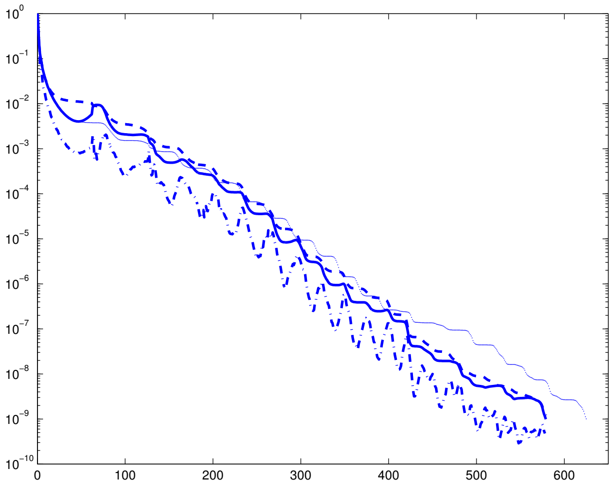

Fig. 6.2. Convergence characteristics for the larger `stokes_testproblem`. The curves correspond to those in Fig. 6.1.

(semi)definite, rather than symmetric but indefinite system matrix. We have generalized previous results on the definiteness and conditioning of the bilinear form with respect to which the unsymmetric saddle point matrix is symmetric. In particular, we have included the case of a general positive semidefinite block C.

We have derived an efficient CG method for solving the (unsymmetric positive definite) saddle point system. In numerical experiments we have seen that this method compares well with the MINRES method, which often is considered the standard solver for (symmetric indefinite) saddle point systems. We point out that, unlike our method, MINRES does not require the estimation of the parameter γ, which can be a significant practical advantage.

Our goal has been to present the theory and the algorithms in a clean and easily readable, rather than most general form. Rather than the end of the story, we consider our paper the starting point for further work. For example, an analysis is needed of the indefinite case, i.e. the practically relevant situations when $\|B\|$ or $\|C\|$ are too large, or γ has been chosen to yield an indefinite matrix $\mathcal{M}(\gamma)$. In very large scale applications, where only crude estimates of the relevant eigenvalues are available, the latter situation is not unlikely to occur. In addition, to make the CG method really useful in practical applications, preconditioning techniques must be studied, and its numerical stability must be understood.

Acknowledgements Part of the work of Jörg Liesen was done during his visit of Emory University in April 2006. He thanks Michele Benzi for his kind hospitality and for very helpful discussions and suggestions. Thanks also to Miro Rozložník, Valeria Simoncini, Petr Tichý, Andy Wathen, and an anonymous referee for their helpful comments.

References

1. S. F. Ashby, M. J. Holst, T. A. Manteuffel, and P. E. Saylor, *The role of the inner product in stopping criteria for conjugate gradient iterations*, BIT, 41 (2001), pp. 26–52.
2. S. F. Ashby, T. A. Manteuffel, and P. E. Saylor, *A taxonomy for conjugate gradient methods*, SIAM J. Numer. Anal., 27 (1990), pp. 1542–1568.
3. O. Axelsson, *Iterative solution methods*, Cambridge University Press, Cambridge, 1994.
4. M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numer., 14 (2005), pp. 1–137.
5. M. Benzi and V. Simoncini, *On the eigenvalues of a class of saddle point matrices*, Numer. Math., 103 (2006), pp. 173–196.
6. 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.
7. H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2005.

8. B. Fischer, A. Ramage, D. J. Silvester, and A. J. Wathen, *Minimum residual methods for augmented systems*, BIT, 38 (1998), pp. 527–543.

9. F. R. Gantmacher, *The theory of matrices. Vols. 1, 2*, Chelsea Publishing Co., New York, 1959.

10. W. Givens, *Fields of values of a matrix*, Proc. Amer. Math. Soc., 3 (1952), pp. 206–209.

11. M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436 (1953).

12. A. S. Householder, *The theory of matrices in numerical analysis*, Dover Publications Inc., New York, 1975. Reprint of 1964 edition.

13. The MathWorks Company, *Matlab, version 6.5.* http://www.mathworks.com.

14. C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

15. D. J. Silvester, H. C. Elman, and A. Ramage, *Incompressible Flow Iterative Solution Software (IFISS), version 2.2.* http://www.manchester.ac.uk/ifiss.

16. G. L. G. Sleijpen, H. A. van der Vorst, and J. Modersitzki, *Differences in the effects of rounding errors in Krylov solvers for symmetric indefinite linear systems*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 726–751.

17. M. Stoll and A. Wathen, *The Bramble-Pasciak preconditioner for saddle point problems*, Numerical Analysis Group Research Report NA-07/13, Oxford University, Computing Laboratory, 2007.

18. ———, *Combination preconditioning and self-adjointness in non-standard inner products with application to saddle point problems*, Numerical Analysis Group Research Report NA-07/11, Oxford University, Computing Laboratory, 2007.