Comput Optim Appl (2007) 36: 249–270
DOI 10.1007/s10589-006-9004-x

Using constraint preconditioners with regularized saddle-point problems

H.S. Dollar · N.I.M. Gould · W.H.A. Schilders · A.J. Wathen

Published online: 22 February 2007

Abstract The problem of finding good preconditioners for the numerical solution of a certain important class of indefinite linear systems is considered. These systems are of a 2 by 2 block (KKT) structure in which the (2,2) block (denoted by $-C$) is assumed to be nonzero.

In *Constraint preconditioning for indefinite linear systems*, SIAM J. Matrix Anal. Appl. 21 (2000), Keller, Gould and Wathen introduced the idea of using constraint preconditioners that have a specific 2 by 2 block structure for the case of C being zero. We shall give results concerning the spectrum and form of the eigenvectors when a preconditioner of the form considered by Keller, Gould and Wathen is used but the system we wish to solve may have $C \neq 0$. In particular, the results presented here

H.S. Dollar (✉) · N.I.M. Gould
Computational Science and Engineering Department, Rutherford Appleton Laboratory, Chilton, Oxfordshire, OX11 0QX, England, UK
e-mail: s.dollar@rl.ac.uk

N.I.M. Gould
e-mail: n.i.m.gould@rl.ac.uk

N.I.M. Gould · A.J. Wathen
Numerical Analysis Group, Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK
e-mail: nick.gould@comlab.ox.ac.uk

A.J. Wathen
e-mail: andy.wathen@comlab.ox.ac.uk

W.H.A. Schilders
Design Methods and Solutions, NXP Semiconductors, High Tech Campus–48, 5656 AE Eindhoven, The Netherlands
e-mail: wil.schilders@nxp.com

W.H.A. Schilders
Department of Mathematics and Computer Science, Technische Universiteit Eindhoven, PO Box 513, 5600 MB Eindhoven, The Netherlands

indicate clustering of eigenvalues and, hence, faster convergence of Krylov subspace iterative methods when the entries of C are small; such a situations arise naturally in interior point methods for optimization and we present results for such problems which validate our conclusions.

Keywords Preconditioning · Indefinite linear systems · Krylov subspace methods

1 Introduction

The solution of systems of the form

$$\underbrace{\begin{bmatrix} A & B^T \\ B & -C \end{bmatrix}}_{\mathcal{A}_C} \begin{bmatrix} x \\ y \end{bmatrix} = \underbrace{\begin{bmatrix} c \\ d \end{bmatrix}}_{b} \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{m\times m}$ are symmetric and $B \in \mathbb{R}^{m\times n}$, is often required in optimization and other various fields, Sect. 1.1. We shall assume that $0 < m \leq n$ and B is of full rank. Various preconditioners which take the general form

$$\mathcal{P}_C = \begin{bmatrix} G & B^T \\ B & -C \end{bmatrix} \tag{1.2}$$

where $G \in \mathbb{R}^{n\times n}$ is some symmetric matrix, have been considered (for example, see [3–5, 8, 18, 23].) When $C = 0$, (1.2) is commonly known as a constraint preconditioner [2, 16, 17, 19]. In practice C is often positive semi-definite (and frequently diagonal).

As we will observe in Sect. 1.1, in interior point methods for constrained optimization a sequence of such problems are solved with the entries in C generally becoming small as the optimization iteration progresses. That is, the regularization is successively reduced as the iterates get closer to the minimum. For the Stokes problem, the entries of C are generally small since they scale with the underlying mesh size and so reduce for finer grids. This motivates us to look at the spectral properties of $\mathcal{P}^{-1}\mathcal{A}_C$, where

$$\mathcal{P} = \begin{bmatrix} G & B^T \\ B & 0 \end{bmatrix}, \tag{1.3}$$

but $C \neq 0$ in (1.1), Sect. 2. We will analyze both the cases of C having full rank and C being rank deficient. We note that when there are equality constraints in the nonlinear programming problem, the corresponding diagonal of C will be identically zero, and thus C will be (trivially) rank deficient.

The obvious advantage in being able to use such a constraint preconditioner is as follows: if B remains constant in each system of the form (1.1), and we choose G in our preconditioner to remain constant, then the preconditioner $\mathcal{P}$ will be unchanged. Any factorizations required to carry out the preconditioning steps in a Krylov subspace iteration will only need to be done once and then used during each execution of the chosen Krylov subspace iteration, instead of carrying out the factorizations at the beginning of each execution.

For symmetric (and in general normal) matrix systems, the convergence of an applicable iterative method is determined by the distribution of the eigenvalues of the coefficient matrix. It is often desirable for the number of distinct eigenvalues to be small so that the rate of convergence is rapid. For non-normal systems the convergence is not so readily described, see [14, page 6].

1.1 Applications requiring the solution of regularized saddle-point problems

In this section we indicate two application areas that require the solution of a regularized saddle-point problems. A comprehensive list of further applications can be found in [2].

Example 1.1 (Nonlinear Programming) Consider the convex nonlinear optimization problem

$$\text{minimize } f(x) \quad \text{such that } c(x) \geq 0, \tag{1.4}$$

where $x \in \mathbb{R}^n$, and $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $-c : \mathbb{R}^n \mapsto \mathbb{R}^{\widehat{m}}$ are convex and twice differentiable. Primal–dual interior point methods [24] for this problem aim to track solutions to the (perturbed) optimality conditions

$$\nabla f(x) = B^T(x)y \quad \text{and} \quad Yc(x) = \mu e \tag{1.5}$$

where y are Lagrange multipliers (dual variables), e is the vector of ones,

$$B(x) = \nabla c(x) \quad \text{and} \quad Y = \text{diag}\{y_1, y_2, \ldots, y_{\widehat{m}}\},$$

as the positive scalar parameter μ is decreased to zero. The Newton correction $(\Delta x, \Delta y)$ to the solution estimate (x, y) of (1.5) satisfy the equation [3]:

$$\begin{bmatrix} A(x,y) & -B^T(x) \\ YB(x) & C(x) \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} -\nabla f(x) + B^T(x)y \\ -Yc(x) + \mu e \end{bmatrix}$$

where

$$A(x,y) = \nabla_{xx} f(x) - \sum_{i=1}^{\widehat{m}} y_i \nabla_{xx} c_i(x) \quad \text{and} \quad C(x) = \text{diag}\{c_1(x), c_2(x), \ldots, c_{\widehat{m}}(x)\}.$$

It is common to eliminate the variables Δy from the Newton system. Since this may introduce unwarranted ill conditioning, it is often better [11] to isolate the effects of poor conditioning by partitioning the constraints so that the values of those indexed by $\mathcal{I}$ are "large" while those indexed by $\mathcal{A}$ are "small", and instead to solve

$$\begin{bmatrix} A + B_{\mathcal{I}}^T C_{\mathcal{I}}^{-1} Y_{\mathcal{I}} B_{\mathcal{I}} & B_{\mathcal{A}}^T \\ B_{\mathcal{A}} & -C_{\mathcal{A}} Y_{\mathcal{A}}^{-1} \end{bmatrix} \begin{bmatrix} \Delta x \\ -\Delta y_{\mathcal{A}} \end{bmatrix} = \begin{bmatrix} -\nabla f + B_{\mathcal{A}}^T y_{\mathcal{A}} + \mu B_{\mathcal{I}}^T C_{\mathcal{I}}^{-1} e \\ -c_{\mathcal{A}} + \mu Y_{\mathcal{A}}^{-1} e \end{bmatrix}$$

where, for brevity, we have dropped the dependence on x and y. The matrix $C_{\mathcal{A}} Y_{\mathcal{A}}^{-1}$ is symmetric and positive definite; as the iterates approach optimality, the entries of this matrix become small. The entries of $B_{\mathcal{I}}^T C_{\mathcal{I}}^{-1} Y_{\mathcal{I}} B_{\mathcal{I}}$ also become small when close to optimality.

Example 1.2 (Stokes) Mixed finite element (and other) discretizations of the Stokes equations

$$\begin{aligned} -\nabla^2 \vec{u} + \nabla p &= \vec{f} \quad \text{in } \Omega \\ \nabla \cdot \vec{u} &= 0 \quad \text{in } \Omega, \end{aligned}$$

for the fluid velocity $\vec{u}$ and pressure p in the domain $\Omega \subset \mathbb{R}^2$ or $\mathbb{R}^3$ yields linear systems in the saddle-point form (1.1) (for derivation and the following properties of this example see [7]). The symmetric block A arises from the diffusion terms $-\nabla^2 \vec{u}$ and B^T represents the discrete gradient operator whilst B represents its adjoint, the (negative) divergence. When (inf-sup) stable mixed finite element spaces are employed, $C = 0$, however for equal order and other spaces which are not inherently stable, stabilized formulations yield symmetric and positive semi-definite matrices C which typically have a large-dimensional kernel—for example for the famous $\boldsymbol{Q}_1$–$\boldsymbol{P}_0$ element which has piecewise bilinear velocities and piecewise constant pressures in 2-dimensions, C typically has a kernel of dimension $m/4$.

2 Preconditioning $\mathcal{A}_C$ by $\mathcal{P}$

Suppose that we precondition $\mathcal{A}_C$ by $\mathcal{P}$, where $\mathcal{P}$ is defined in (1.3). The decision to investigate this form of preconditioner is motivated in Sect. 1. We shall use the following assumptions in our theorems:

A1 $B \in \mathbb{R}^{m \times n}$ $(m \leq n)$ has full rank,
A2 C has rank $p > 0$ and is factored as EDE^T, where $E \in \mathbb{R}^{m \times p}$ and has orthonormal columns, and $D \in \mathbb{R}^{p \times p}$ is non-singular,
A3 If $p < m$, then $F \in \mathbb{R}^{m \times (m-p)}$ is such that its columns form a basis for the nullspace of C and $N \in \mathbb{R}^{n \times (n-m+p)}$ is such that its columns form a basis of the nullspace of $F^T B$,
A4 If $p = m$, then $N = I \in \mathbb{R}^{n \times n}$.

Theorem 2.1 *Assume that* **A1**–**A4** *hold, then the matrix* $\mathcal{P}^{-1}\mathcal{A}_C$ *has*:

- *at least* $2(m - p)$ *eigenvalues at* 1,
- *its non-unit eigenvalues defined by the finite* (*and non-unit*) *eigenvalues of the quadratic eigenvalue problem*

$$\begin{aligned} 0 = {} & \lambda^2 N^T B^T E D^{-1} E^T B N w_{n1} - \lambda N^T (G + 2B^T E D^{-1} E^T B) N w_{n1} \\ & + N^T (A + B^T E D^{-1} E^T B) N w_{n1}. \end{aligned}$$

Proof We shall consider the cases of $p = m$ and $0 < p < m$ separately.

Case $p = m$. The generalized eigenvalue problem takes the form

$$\begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} G & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{2.1}$$

Expanding this out we obtain

$$Ax + B^T y = \lambda Gx + \lambda B^T y, \tag{2.2}$$

$$Bx - Cy = \lambda Bx. \tag{2.3}$$

From (2.3) we deduce that either $\lambda = 1$ and $y = 0$, or $\lambda \neq 1$. If the former holds, then (2.2) implies that x must satisfy

$$Ax = Gx.$$

Thus, the associated eigenvectors will take the form

$$[x^T \quad 0^T]^T,$$

where $x \neq 0$ satisfies $Ax = Gx$. There is no guarantee that such an eigenvector will exist, and therefore no guarantee that there are any unit eigenvalues.

If $\lambda \neq 1$, then Eq. (2.3) and the non-singularity of C gives

$$y = (1 - \lambda)C^{-1}Bx, \quad x \neq 0.$$

By substituting this into (2.2) and rearranging we obtain the quadratic eigenvalue problem

$$(\lambda^2 B^T C^{-1} B - \lambda(G + 2B^T C^{-1} B) + A + B^T C^{-1} B)x = 0. \tag{2.4}$$

The non-unit eigenvalues of (2.1) are therefore defined by the finite (non-unit) eigenvalues of (2.4).

Now, assumption **A2** implies that

$$C^{-1} = E D^{-1} E^T,$$

and, hence, letting $w_{n1} = x$ we complete our proof for the case $p = m$.

Case $0 < p < m$. Any $y \in \mathbb{R}^m$ can be written as $y = Ey_e + Fy_f$. Substituting this into (2.1) and premultiplying the resulting generalized eigenvalue problem by

$$\begin{bmatrix} I & 0 \\ 0 & E^T \\ 0 & F^T \end{bmatrix},$$

we obtain

$$\left[\begin{array}{cc|c} A & B^T E & B^T F \\ E^T B & -D & 0 \\ \hline F^T B & 0 & 0 \end{array}\right] \begin{bmatrix} x \\ y_e \\ y_f \end{bmatrix} = \lambda \left[\begin{array}{cc|c} G & B^T E & B^T F \\ E^T B & 0 & 0 \\ \hline F^T B & 0 & 0 \end{array}\right] \begin{bmatrix} x \\ y_e \\ y_f \end{bmatrix}. \tag{2.5}$$

Noting that the (3,3) block has dimension $(m - p) \times (m - p)$ and is a zero matrix in both coefficient matrices, we can apply Theorem 2.1 from [16] to obtain:

- $\mathcal{P}^{-1}\mathcal{A}_C$ has an eigenvalue at 1 with multiplicity $2(m - p)$,

- the remaining $n - m + 2p$ eigenvalues are defined by the generalized eigenvalue problem

$$\overline{N}^T \begin{bmatrix} A & B^T E \\ E^T B & -D \end{bmatrix} \overline{N} w_n = \lambda \overline{N}^T \begin{bmatrix} G & B^T E \\ E^T B & 0 \end{bmatrix} \overline{N} w_n \tag{2.6}$$

  where $\overline{N}$ is an $(n + p) \times (n - m + 2p)$ basis for the nullspace of $[F^T B\ 0]$.

One choice for $\overline{N}$ is

$$\overline{N} = \begin{bmatrix} N & 0 \\ 0 & I \end{bmatrix}.$$

Substituting this into (2.6) we obtain the generalized eigenvalue problem

$$\begin{bmatrix} N^T A N & N^T B^T E \\ E^T B N & -D \end{bmatrix} \begin{bmatrix} w_{n1} \\ w_{n2} \end{bmatrix} = \lambda \begin{bmatrix} N^T G N & N^T B^T E \\ E^T B N & 0 \end{bmatrix} \begin{bmatrix} w_{n1} \\ w_{n2} \end{bmatrix}. \tag{2.7}$$

This generalized eigenvalue problem resembles that of (2.1) in the first case considered in this proof. Therefore, the non-unit eigenvalues of $\mathcal{P}^{-1}\mathcal{A}_C$ are equal to the finite (and non-unit) eigenvalues of the quadratic eigenvalue problem

$$\begin{aligned} 0 = {} & \lambda^2 N^T B^T E D^{-1} E^T B N w_{n1} - \lambda N^T (G + 2B^T E D^{-1} E^T B) N w_{n1} \\ & + N^T (A + B^T E D^{-1} E^T B) N w_{n1}. \end{aligned} \tag{2.8}$$

Since $N^T B^T E D^{-1} E^T B N$ has a nullspace of dimension $n - m$, this quadratic eigenvalue problem has $2(n - m + p) - (n - m) = n - m + 2p$ finite eigenvalues [22]. □

The following numerical examples illustrate how the rank of C dictates a lower bound on the number of unit eigenvalues. In particular, Example 2.2 demonstrates that there is no guarantee that the preconditioned matrix has unit eigenvalues when C is nonsingular.

Example 2.2 (C nonsingular) Consider the matrices

$$\mathcal{A}_C = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}, \qquad \mathcal{P} = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

so that $m = p = 1$ and $n = 2$. The preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}_C$ has eigenvalues at $\frac{1}{2}$, $2 - \sqrt{2}$ and $2 + \sqrt{2}$. The corresponding eigenvectors are $[0\ 1\ 0]^T$, $[1\ 0\ (\sqrt{2} - 1)]^T$ and $[1\ 0\ -(\sqrt{2} + 1)]^T$ respectively. The preconditioned system $\mathcal{P}^{-1}\mathcal{A}_C$ has all non-unit eigenvalues, but this does not go against Theorem 2.1 because $m - p = 0$. With our choices of $\mathcal{A}_C$ and $\mathcal{P}$, and setting $D = [1]$ and $E = [1]$ ($C = EDE^T$), the quadratic eigenvalue problem (2.8) is

$$\left(\lambda^2 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} + \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 0.$$

This quadratic eigenvalue problem has three finite eigenvalues which are $\lambda = \frac{1}{2}$, $\lambda = 2 - \sqrt{2}$ and $\lambda = 2 + \sqrt{2}$.

Example 2.3 (C semidefinite) Consider the matrices

$$\mathcal{A}_C = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \qquad \mathcal{P} = \begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

so that $m = 2$, $n = 2$ and $p = 1$. The preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}_C$ has two unit eigenvalues and a further two at $\lambda = 2 - \sqrt{2}$ and $\lambda = 2 + \sqrt{2}$. There is just one linearly independent eigenvector associated with the unit eigenvector; specifically this is $[0\ 0\ 1\ 0]^T$. For the non-unit eigenvalues, the eigenvectors are $[0\ 1\ 0\ (\sqrt{2} - 1)]^T$ and $[0\ 1\ 0\ -(\sqrt{2} + 1)]^T$ respectively.

Since $2(m - p) = 2$, we correctly expected there to be at least two unit eigenvalues, Theorem 2.1. The remaining eigenvalues will be defined by the finite eigenvalues of the quadratic eigenvalue problem (2.8):

$$\left(\lambda^2 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\right) \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 0$$

where $D = [1]$ and $E = [0\ 1]^T$ are used as factors of C. This quadratic eigenvalue problem has three finite eigenvalues which are $\lambda = 2 - \sqrt{2}$ and $\lambda = 2 + \sqrt{2}$; the corresponding eigenvectors have $u_1 = 0$.

2.1 Analysis of the quadratic eigenvalue problem

We note that the quadratic eigenvalue problem (2.8) can have negative and complex eigenvalues, see [22]. The following theorem gives sufficient conditions for general quadratic eigenvalue problems to have real and positive eigenvalues.

Theorem 2.4 *Consider the quadratic eigenvalue problem*

$$(\lambda^2 K - \lambda L + M)x = 0, \tag{2.9}$$

where $M, L \in \mathbb{R}^{n\times n}$ are symmetric positive definite, and $K \in \mathbb{R}^{n\times n}$ is symmetric positive semidefinite. Define $\gamma(M, L, K)$ to be

$$\gamma(M, L, K) = \min\{(x^T Lx)^2 - 4(x^T Mx)(x^T Kx) : \|x\|_2 = 1\}.$$

If $\gamma(M, L, K) > 0$, then the eigenvalues λ are real and positive, and there are n linearly independent eigenvectors associated with the n largest (n smallest) eigenvalues.

Proof From [22, Sect. 1] we know that under our assumptions the quadratic eigenvalue problem

$$(\mu^2 M + \mu L + K)x = 0$$

has real and negative eigenvalues. Suppose we divide this equation by μ^2 and set $\lambda = -1/\mu$. The quadratic eigenvalue problem (2.9) is obtained, and since μ is real and negative, λ is real and positive. □

We would like to be able to use the above theorem to show that, under suitable assumptions, all the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}_C$ are real and positive. Let

$$\widetilde{D} = N^T B^T E D^{-1} E^T B N \tag{2.10}$$

where D and E are as defined in assumption **A2**, and N is as defined in assumption **A3**. If we assume that $N^T A N + \widetilde{D}$ is positive definite, then we may write $N^T A N + \widetilde{D} = R^T R$ for some nonsingular matrix R. If we premultply-multiply the quadratic eigenvalue problem (2.8) by R^{-T} and substitute in $z = R w_{n1}$, then we find that it is similar to the quadratic eigenvalue problem

$$\big(\lambda^2 R^{-T}\widetilde{D}R^{-1} - \lambda R^{-T}(N^T G N + 2\widetilde{D})R^{-1} + I\big)z = 0.$$

Thus, if we assume that $N^T A N + \widetilde{D}$, $N^T G N + 2\widetilde{D}$ are positive definite and $\widetilde{D}$ is positive semi-definite, and can show that

$$\gamma\big(I, R^{-T}(N^T G N + 2\widetilde{D})R^{-1}, R^{-T}\widetilde{D}R^{-1}\big) > 0,$$

where $\gamma(\cdot,\cdot,\cdot)$ is as defined in Theorem 2.4, then we can apply the above theorem to show that (2.8) has real and positive eigenvalues.

Let us assume that $\|z\|_2 = 1$, then

$$\begin{aligned}
&\big(z^T R^{-T}\big(N^T G N + 2\widetilde{D}\big)R^{-1}z\big)^2 - 4z^T z z^T R^{-T}\widetilde{D}R^{-1}z \\
&\quad = \big(z^T R^{-T} N^T G N R^{-1} z + 2z^T R^{-T}\widetilde{D}R^{-1}z\big)^2 - 4z^T R^{-T}\widetilde{D}R^{-1}z \\
&\quad = (z^T R^{-T} N^T G N R^{-1} z)^2 \\
&\qquad + 4z^T R^{-T}\widetilde{D}R^{-1}z\big(z^T R^{-T} N^T G N R^{-1} z + z^T R^{-T}\widetilde{D}R^{-1}z - 1\big) \\
&\quad = (w_{n1}^T N^T G N w_{n1})^2 + 4w_{n1}^T\widetilde{D}w_{n1}\big(w_{n1}^T N^T G N w_{n1} + w_{n1}^T\widetilde{D}w_{n1} - 1\big)
\end{aligned} \tag{2.11}$$

where $1 = \|z\|_2 = \|R w_{n1}\|_2 = \|w_{n1}\|_{N^T A N + \widetilde{D}}$. Clearly, we can guarantee that (2.11) is positive if

$$w_{n1}^T N^T G N w_{n1} + w_{n1}^T\widetilde{D}w_{n1} > 1 \quad \text{for all } w_{n1} \text{ such that } \|w_{n1}\|_{N^T A N + \widetilde{D}} = 1,$$

that is

$$\frac{w_{n1}^T N^T G N w_{n1} + w_{n1}^T\widetilde{D}w_{n1}}{w_{n1}^T(N^T A N + \widetilde{D})w_{n1}} > \frac{w_{n1}^T(N^T A N + \widetilde{D})w_{n1}}{w_{n1}^T(N^T A N + \widetilde{D})w_{n1}} \quad \text{for all } w_{n1} \neq 0.$$

Rearranging we find that we require

$$w_{n1}^T N^T G N w_{n1} > w_{n1}^T N^T A N w_{n1}$$

for all $w_{n1} \neq 0$. Thus we need only scale any positive definite G such that $w_{n1}^T N^T G N w_{n1}/(w_{n1}^T N^T N w_{n1}) > \|A\|_2^2$ for all $N w_{n1} \neq 0$ to guarantee that (2.11) is positive for all w_{n1} such that $\|w_{n1}\|_{N^T A N + \widetilde{D}} = 1$. For example, we could choose $G = \alpha I$, where $\alpha > \|A\|_2^2$.

Using the above in conjunction with Theorem 2.1 we obtain the following result:

Theorem 2.5 *Suppose that* **A1–A4** *hold and* $\widetilde{D}$ *is as defined in* (2.10). *Further, assume that* $A + \widetilde{D}$ *and* $G + 2\widetilde{D}$ *are symmetric positive definite,* $\widetilde{D}$ *is symmetric positive semidefinite and*

$$\min\left\{(z^T G z)^2 + 4(z^T \widetilde{D} z)(z^T G z + z^T \widetilde{D} z - 1) : \|z\|_{A+\widetilde{D}} = 1\right\} > 0, \tag{2.12}$$

then all the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}_C$ *are real and positive. The matrix* $\mathcal{P}^{-1}\mathcal{A}_C$ *also has* $m - p + i + j$ *linearly independent eigenvectors. There are*

1. $m - p$ *eigenvectors of the form* $[0^T \ y_f^T]^T$ *that correspond to the case* $\lambda = 1$,
2. i $(0 \leq i \leq n)$ *eigenvectors of the form* $[w^T \ 0^T \ y_f^T]^T$ *arising from* $Aw = Gw$ *for which the* i *vectors* w *are linearly independent, and* $\lambda = 1$, *and*
3. j $(0 \leq j \leq n - m + 2p)$ *eigenvectors of the form* $[0^T \ w_{n1}^T \ w_{n2}^T \ y_f^T]^T$ *corresponding to the eigenvalues of* $\mathcal{P}^{-1}\mathcal{A}_C$ *not equal to* 1, *where the components* w_{n1} *arise from the quadratic eigenvalue problem*

$$\begin{aligned} 0 = {} & \lambda^2 N^T B^T E D^{-1} E^T B N w_{n1} - \lambda N^T (G + 2B^T E D^{-1} E^T B) N w_{n1} \\ & + N^T (A + B^T E D^{-1} E B) N w_{n1}, \end{aligned}$$

with $\lambda \neq 1$, *and* $w_{n2} = (1 - \lambda) D^{-1} E^T B N w_{n1}$.

Proof It remains for us to prove the form of the eigenvectors and that they are linearly independent. We will consider the case $p = m$ and $0 < p < m$ separately.

Case $p = m$. From the proof of Theorem 2.1, when $\lambda = 1$ the eigenvectors must take the form $[x^T \ 0^T]^T$, where $Ax = \sigma G x$ for which the i vectors x are linearly independent, $\sigma = 1$. Hence, any eigenvectors corresponding to a unit eigenvalue fall into the second statement of the theorem and there are i $(0 \leq i \leq n)$ such eigenvectors which are linearly independent. The proof of Theorem 2.1 also shows that the eigenvectors corresponding to $\lambda \neq 1$ take the form $[x^T \ y^T]^T$, where x corresponds to the quadratic eigenvalue problem (2.4) and $y = (1 - \lambda) C^{-1} B x = (1 - \lambda) D^{-1} E B N x$ (since we can set $D = C$ and $E = I$). Clearly, there are at most $n + m$ such eigenvectors. By our assumptions, all of the vectors x defined by the quadratic eigenvalue problem (2.4) are linearly independent. Also, if x is associated with two eigenvalues, then these eigenvalues must be distinct [22]. By setting $w_{n1} = x$ and $w_{n2} = y$ we obtain j $(0 \leq j \leq n + m)$ eigenvectors of the form given in statement 3 of the proof.

It remains for us to prove that the $i + j$ eigenvectors defined above are linearly independent. Hence, we need to show that

$$\begin{bmatrix} x_1^{(1)} & \dots & x_i^{(1)} \\ 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} a_1^{(1)} \\ \vdots \\ a_i^{(1)} \end{bmatrix} + \begin{bmatrix} x_1^{(2)} & \dots & x_j^{(2)} \\ y_1^{(2)} & \dots & y_j^{(2)} \end{bmatrix} \begin{bmatrix} a_1^{(2)} \\ \vdots \\ a_j^{(2)} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2.13}$$

implies that the vectors $a^{(1)}$ and $a^{(2)}$ are zero vectors. Multiplying (2.13) by $\mathcal{P}^{-1}\mathcal{A}_C$, and recalling that in the previous equation the first matrix arises from $\lambda_l = 1$ ($l = 1, \dots, i$) and the second matrix from $\lambda_l \neq 1$ ($l = 1, \dots, j$) gives

$$\begin{bmatrix} x_1^{(1)} & \dots & x_i^{(1)} \\ 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} a_1^{(1)} \\ \vdots \\ a_i^{(1)} \end{bmatrix} + \begin{bmatrix} x_1^{(2)} & \dots & x_j^{(2)} \\ y_1^{(2)} & \dots & y_j^{(2)} \end{bmatrix} \begin{bmatrix} \lambda_1^{(2)} a_1^{(2)} \\ \vdots \\ \lambda_j^{(2)} a_j^{(2)} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.14}$$

Subtracting (2.13) from (2.14) we obtain

$$\begin{bmatrix} x_1^{(2)} & \dots & x_j^{(2)} \\ y_1^{(2)} & \dots & y_j^{(2)} \end{bmatrix} \begin{bmatrix} (\lambda_1^{(2)} - 1)a_1^{(2)} \\ \vdots \\ (\lambda_j^{(2)} - 1)a_j^{(2)} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.15}$$

Some of the eigenvectors x defined by the quadratic eigenvalue problem (2.4) will be associated with two (non-unit) eigenvalues; let us assume that there are k such eigenvectors. By our assumptions, these eigenvalues must be distinct. Without loss of generality, assume that $x_l^{(2)} = x_{k+l}^{(2)}$ for $l = 1, \dots, k$. The vectors $x_l^{(2)}$ ($l = k + 1, \dots, j$) are linearly independent and $\lambda_l^{(2)} \neq 1$ ($l = 2k + 1, \dots, j$), which gives rise to $a_l^{(2)} = 0$ for $l = 2k + 1, \dots, j$. Equation (2.15) becomes

$$\begin{bmatrix} x_1^{(2)} & \dots & x_k^{(2)} & x_1^{(2)} & \dots & x_k^{(2)} \\ y_1^{(2)} & \dots & y_k^{(2)} & y_{k+1}^{(2)} & \dots & y_{2k}^{(2)} \end{bmatrix} \begin{bmatrix} (\lambda_1^{(2)} - 1)a_1^{(2)} \\ \vdots \\ (\lambda_j^{(2)} - 1)a_{2k}^{(2)} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.16}$$

The vectors $x_l^{(2)}$ ($l = 1, \dots, k$) are linearly independent. Hence

$$(\lambda_l^{(2)} - 1)a_l^{(2)} x_l^{(2)} + (\lambda_{l+k}^{(2)} - 1)a_{l+k}^{(2)} x_l^{(2)} = 0, \quad l = 1, \dots, k,$$

and

$$a_l^{(2)} = -a_{l+k}^{(2)} \frac{1 - \lambda_{l+k}^{(2)}}{1 - \lambda_l^{(2)}}, \quad l = 1, \dots, k.$$

Now $y_l^{(2)} = (1 - \lambda_l^{(2)})C^{-1}Bx_l^{(2)}$ for $l = 1, \dots, 2k$. Hence, we require

$$(\lambda_l^{(2)} - 1)^2 a_l^{(2)} C^{-1}Bx_l^{(2)} + (\lambda_{l+k}^{(2)} - 1)^2 a_{l+k}^{(2)} C^{-1}Bx_l^{(2)} = 0, \quad l = 1, \dots, k.$$

Substituting in $a_l^{(2)} = -a_{l+k}^{(2)}(1-\lambda_{l+k}^{(2)})/(1-\lambda_l^{(2)})$ and rearranging gives $(\lambda_l^{(2)} - 1)a_l^{(2)} = (\lambda_{l+k}^{(2)} - 1)a_{l+k}^{(2)}$ for $l = 1, \ldots, k$. Since these eigenvalues are non-unit and $\lambda_l^{(2)} \neq \lambda_{l+k}^{(2)}$ for $l = 1, \ldots, k$, we conclude that $a_l^{(2)} = 0$ $(l = 1, \ldots, j)$.

We also have linear independence of $x_l^{(1)}$ $(l = 1, \ldots, i)$, which implies that $a_l^{(1)} = 0$ $(l = 1, \ldots, i)$.

Case $0 < p < m$. From the proof of Theorem 2.1, the generalized eigenvalue problem can be expressed as

$$\left[\begin{array}{cc|c} A & B^T E & B^T F \\ E^T B & -D & 0 \\ \hline F^T B & 0 & 0 \end{array}\right] \begin{bmatrix} x \\ y_e \\ y_f \end{bmatrix} = \lambda \left[\begin{array}{cc|c} G & B^T E & B^T F \\ E^T B & 0 & 0 \\ \hline F^T B & 0 & 0 \end{array}\right] \begin{bmatrix} x \\ y_e \\ y_f \end{bmatrix}. \tag{2.17}$$

The first part of the proof for this case follows similarly to that of Theorem 2.3 in [16]. Let $[\overline{M}\ \overline{N}][\overline{R}^T\ 0]^T$ be an orthogonal factorization of $[F^T B\ 0]$, where $R \in \mathbb{R}^{(m-p)\times(m-p)}$ is upper triangular, $\overline{M} \in \mathbb{R}^{(n+p)\times(m-p)}$, and $\overline{N} \in \mathbb{R}^{(n+p)\times(n-m+2p)}$ is a basis for the nullspace of $[F^T B\ 0]$. Premultiplying (2.17) by the nonsingular and square matrix

$$\begin{bmatrix} \overline{M}^T & 0 \\ \overline{N}^T & 0 \\ 0 & I \end{bmatrix},$$

substituting in

$$\begin{bmatrix} x \\ y_e \\ y_f \end{bmatrix} = \begin{bmatrix} \overline{M} & \overline{N} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} w_m \\ w_n \\ y_f \end{bmatrix},$$

and expanding out gives

$$\overline{M}^T \widehat{A}\,\overline{M} w_m + \overline{M}^T \widehat{A}\,\overline{N} w_n + \overline{R} y_f = \lambda\big[\overline{M}^T \widehat{G}\,\overline{M} w_m + \overline{M}^T \widehat{G}\,\overline{N} w_n + \overline{R} y_f\big], \tag{2.18}$$

$$\overline{N}^T \widehat{A}\,\overline{M} w_m + \overline{N}^T \widehat{A}\,\overline{N} w_n = \lambda\big[\overline{N}^T \widehat{G}\,\overline{M} w_m + \overline{N}^T \widehat{G}\,\overline{N} w_n\big], \tag{2.19}$$

$$\overline{R}^T w_m = \lambda \overline{R}^T w_m \tag{2.20}$$

where

$$\widehat{A} = \begin{bmatrix} A & B^T E \\ E^T B & -D \end{bmatrix} \quad \text{and} \quad \widehat{G} = \begin{bmatrix} G & B^T E \\ E^T B & 0 \end{bmatrix}.$$

From (2.20), it may be deduced that either $\lambda = 1$ or $w_m = 0$. In the former case, (2.18) and (2.19) may be simplified to

$$Q^T \begin{bmatrix} A & B^T E \\ E^T B & -D \end{bmatrix} Q w = Q^T \begin{bmatrix} G & B^T E \\ E^T B & 0 \end{bmatrix} Q w \tag{2.21}$$

where $Q = [\overline{M}\ \overline{N}]$ and $w = [w_m^T\ w_n^T]$. Since Q is orthogonal, the general eigenvalue problem (2.21) is equivalent to considering

$$\begin{bmatrix} A & B^T E \\ E^T B & -D \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \sigma \begin{bmatrix} G & B^T E \\ E^T B & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \tag{2.22}$$

where $[w_1^T\ w_2^T]^T \neq 0$ if and only if $\sigma = 1$, and $w_1 \in \mathbb{R}^n$, $w_2 \in \mathbb{R}^p$. As in the first case of this proof, nonsingularity of D and $\sigma = 1$ implies that $w_2 = 0$. There are $m - p$ linearly independent eigenvectors $[0^T\ 0^T\ u_f^T]^T$ corresponding to $w_1 = 0$, and a further i $(0 \leq i \leq n)$ linearly independent eigenvectors corresponding to $w_1 \neq 0$ and $\sigma = 1$.

Now suppose that $\lambda \neq 1$, in which case $w_m = 0$. Equations (2.18) and (2.19) yield

$$\overline{N}^T \begin{bmatrix} A & B^T E \\ E^T B & -D \end{bmatrix} \overline{N} w_n = \lambda \overline{N}^T \begin{bmatrix} G & B^T E \\ E^T B & 0 \end{bmatrix} \overline{N} w_n, \tag{2.23}$$

$$\overline{M}^T \begin{bmatrix} A & B^T E \\ E^T B & -D \end{bmatrix} \overline{N} w_n + \overline{R} y_f = \lambda \left[\overline{M}^T \begin{bmatrix} G & B^T E \\ E^T B & 0 \end{bmatrix} \overline{N} w_n + \overline{R} y_f \right]. \tag{2.24}$$

The generalized eigenvalue problem (2.24) defines $n - m + 2p$ eigenvalues, where j $(0 \leq j \leq n - m)$ of these are not equal to 1 and for which two cases have to be distinguished. If $w_n = 0$, then (2.23) and $\lambda \neq 1$ imply that $y_f = 0$. In this case no extra eigenvalues arise. Suppose that $w_n \neq 0$, then, from the proof of Theorem 2.1, the eigenvalues are equivalently defined by (2.8) and

$$w_n = \begin{bmatrix} w_{n1} \\ (1 - \lambda) D^{-1} E^T B N w_{n1} \end{bmatrix}.$$

Hence, the j $(0 \leq j \leq n - m + 2l)$ eigenvectors corresponding to the non-unit eigenvalues of $\mathcal{P}^{-1}\mathcal{A}_C$ take the form $[0^T\ w_{n1}^T\ w_{n2}^T\ y_f^T]^T$.

Proof of the linear independence of these eigenvectors follows similarly to the case of $p = m$. □

Observing that the coefficient matrices in (2.5) are of the form of those considered by Gould, Hribar and Nocedal [12], we could apply a projected preconditioned conjugate gradient method to solve (1.1) if all the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}_C$ are real and positive and we have a decomposition of C as in **A2**. Theorem 2.5 therefore gives conditions which allow us to use such a method. Dollar gives a variant of this method in which no decomposition of C is required, see [6, Sect. 5.5]. The derivation of such a method bears close resemblance to that of a nullspace method. The nullspace N is required in the derivation but, as in [12], we can rewrite the algorithm in such a manner that there is no need for N to be known explicitly.

3 Convergence

In the context of this paper, the convergence of an iterative method under preconditioning is not only influenced by the spectral properties of the coefficient matrix, but

also by the relationship between m, n and p. We can determine an upper bound on the number of iterations of an appropriate Krylov subspace method by considering minimum polynomials of the coefficient matrix.

Definition 3.1 Let $\mathcal{A} \in \mathbb{R}^{(n+m)\times(n+m)}$. The monic polynomial f of minimum degree such that $f(\mathcal{A}) = 0$ is called the minimum polynomial of $\mathcal{A}$.

Krylov subspace theory states that iteration with any method with an optimality property, e.g. GMRES, will terminate when the degree of the minimum polynomial is attained, [21]. In particular, the degree of the minimum polynomial is equal to the dimension of the corresponding Krylov subspace (for general b), [20, Proposition 6.1].

Theorem 3.2 *Suppose that the assumptions of Theorem 2.5 hold. The dimension of the Krylov subspace* $\mathcal{K}(\mathcal{P}^{-1}\mathcal{A}_C, b)$ *is at most* $\min\{n - m + 2p + 2, n + m\}$.

Proof Suppose that $0 < p < m$. As in the proof to Theorem 2.1, the generalized eigenvalue problem can be written as

$$\left[\begin{array}{ccc|c} A & B^TE & B^TF & \\ E^TB & -D & 0 & \\ \hline F^TB & 0 & 0 & \end{array}\right]\begin{bmatrix} x \\ y_e \\ y_f \end{bmatrix} = \lambda \left[\begin{array}{ccc|c} G & B^TE & B^TF & \\ E^TB & 0 & 0 & \\ \hline F^TB & 0 & 0 & \end{array}\right]\begin{bmatrix} x \\ y_e \\ y_f \end{bmatrix}. \tag{3.1}$$

Hence, the preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}_C$ can be written as

$$\widehat{\mathcal{P}}^{-1}\widehat{\mathcal{A}_C} = \begin{bmatrix} \Theta_1 & 0 \\ \Theta_2 & I \end{bmatrix}, \tag{3.2}$$

where the precise forms of $\Theta_1 \in \mathbb{R}^{(n+p)\times(n+p)}$ and $\Theta_2 \in \mathbb{R}^{(m-p)\times(n+p)}$ are irrelevant.

From the earlier eigenvalue derivation, it is evident that the characteristic polynomial of the preconditioned linear system (3.2) is

$$(\mathcal{P}^{-1}\mathcal{A}_C - I)^{2(m-p)} \prod_{i=1}^{n-m+2p} (\mathcal{P}^{-1}\mathcal{A}_C - \lambda_i I).$$

In order to prove the upper bound on the Krylov subspace dimension, we need to show that the order of the minimum polynomial is less than or equal to $\min\{n - m + 2p + 2, n + m\}$. Expanding the polynomial $(\mathcal{P}^{-1}\mathcal{A}_C - I)\prod_{i=1}^{n-m+2p}(\mathcal{P}^{-1}\mathcal{A}_C - \lambda_i I)$ of degree $n - m + 2p + 1$, we obtain

$$\begin{bmatrix} (\Theta_1 - I)\prod_{i=1}^{n-m+2p}(\Theta_1 - \lambda_i I) & 0 \\ \Theta_2 \prod_{i=1}^{n-m+2p}(\Theta_1 - \lambda_i I) & 0 \end{bmatrix}.$$

Since the assumptions of Theorem 2.5 hold, Θ_1 has a full set of linearly independent eigenvectors and is diagonalizable. Hence, $(\Theta_1 - I)\prod_{i=1}^{n-m+2p}(\Theta_1 - \lambda_i I) = 0$.

We therefore obtain

$$(\mathcal{P}^{-1}\mathcal{A}_C - I)\prod_{i=1}^{n-m+2p}(\mathcal{P}^{-1}\mathcal{A}_C - \lambda_i I) = \begin{bmatrix} 0 & 0 \\ \Theta_2 \prod_{i=1}^{n-m+2p}(\Theta_1 - \lambda_i I) & 0 \end{bmatrix}. \quad (3.3)$$

If $\Theta_2 \prod_{i=1}^{n-m+2p}(\Theta_1 - \lambda_i I) = 0$, then the order of the minimum polynomial of $\mathcal{P}^{-1}\mathcal{A}_C$ is less than or equal to $\min\{n-m+2p+1, n+m\}$. If $\Theta_2 \prod_{i=1}^{n-m+2p}(\Theta_1 - \lambda_i I) = 0$, then the dimension of $\mathcal{K}(\mathcal{P}^{-1}\mathcal{A}_C, c)$ is at most $\min\{n-m+2p+2, n+m\}$ since multiplication of (3.3) by another factor $(\mathcal{P}^{-1}\mathcal{A}_C - I)$ gives the zero matrix.

If $p = m$, then trivially $\mathcal{K}(\mathcal{P}^{-1}\mathcal{A}_C, b)$ has dimension at most $\min\{n - m + 2p + 2, n + m\}$. □

3.1 Clustering of eigenvalues when $\|C\|$ is small

When using interior-point methods to solve optimization problems, the matrix C is generally diagonal and of full rank. In this case, Theorem 3.2 would suggest that there is little advantage of using a constraint preconditioner of the form $\mathcal{P}$ over any other preconditioner. However, in interior-point methods the entries of C also become small as we get close to optimality and, hence, $\|C\|$ is small. In the following we shall assume that the norm considered is the ℓ_2 norm, but the results can be generalized to other norms.

Theorem 3.3 *Let $\zeta > 0$, $\delta \geq 0$, $\varepsilon \geq 0$ and $\delta^2 + 4\zeta(\delta - \varepsilon) \geq 0$ then the roots of the quadratic equation*

$$\lambda^2\zeta - \lambda(\delta + 2\zeta) + \varepsilon + \zeta = 0$$

satisfy

$$\lambda = 1 + \frac{\delta}{2\zeta} \pm \mu, \qquad \mu \leq \sqrt{2}\max\left\{\frac{\delta}{2\zeta}, \sqrt{\frac{|\delta - \varepsilon|}{\zeta}}\right\}$$

Proof The roots of the quadratic equation satisfy

$$\begin{aligned}\lambda &= \frac{\delta + 2\zeta \pm \sqrt{(\delta + 2\zeta)^2 - 4\zeta(\varepsilon + \zeta)}}{2\zeta} \\ &= 1 + \frac{\delta}{2\zeta} \pm \frac{\sqrt{\delta^2 + 4\zeta(\delta - \varepsilon)}}{2\zeta} \\ &= 1 + \frac{\delta}{2\zeta} \pm \sqrt{\left(\frac{\delta}{2\zeta}\right)^2 + \frac{\delta - \varepsilon}{\zeta}}.\end{aligned}$$

If $\frac{\delta - \varepsilon}{\zeta} \geq 0$, then

$$\sqrt{\left(\frac{\delta}{2\zeta}\right)^2 + \frac{\delta - \varepsilon}{\zeta}} \leq \sqrt{2\max\left\{\left(\frac{\delta}{2\zeta}\right)^2, \frac{\delta - \varepsilon}{\zeta}\right\}} = \sqrt{2}\max\left\{\frac{\delta}{2\zeta}, \sqrt{\frac{\delta - \varepsilon}{\zeta}}\right\}.$$

If $\frac{\delta-\varepsilon}{\zeta} \leq 0$, then the assumption $\delta^2 + 4\zeta(\delta - \varepsilon) \geq 0$ implies that

$$\left(\frac{\delta}{2\zeta}\right)^2 \geq \frac{\varepsilon - \delta}{\zeta} \geq 0.$$

Hence,

$$\sqrt{\left(\frac{\delta}{2\zeta}\right)^2 + \frac{\delta-\varepsilon}{\zeta}} \leq \frac{\delta}{2\zeta} < \sqrt{2}\max\left\{\frac{\delta}{2\zeta}, \sqrt{\frac{\varepsilon-\delta}{\zeta}}\right\}. \qquad \square$$

Remark 3.4 The important point to notice is that if $\zeta \gg \delta$ and $\zeta \gg \varepsilon$, then $\lambda \approx 1$ in Theorem 3.3.

Theorem 3.5 *Assume that the assumptions of Theorem 2.5 hold, then the eigenvalues λ of (2.8) subject to $E^T BNu \neq 0$, will satisfy*

$$|\lambda - 1| = \mathcal{O}\big(\max\{\|C\|, \|G - A\|\sqrt{\|C\|}\}\big)$$

for small values of $\|C\|$.

Proof Suppose that $C = EDE^T$ is a reduced singular value decomposition of C, where the columns of $E \in \mathbb{R}^{m\times p}$ are orthogonal and $D \in \mathbb{R}^{p\times p}$ is diagonal with entries d_j that are non-negative and in non-increasing order.

In the following, $\|.\| = \|.\|_2$, so that

$$\|C\| = \|D\| = d_1.$$

Premultiplying the quadratic eigenvalue problem (2.8) by u^T gives

$$0 = \lambda^2 u^T \widetilde{D} u - \lambda(u^T N^T G N u + 2u^T \widetilde{D} u) + (u^T N^T A N u + u^T \widetilde{D} u). \tag{3.4}$$

Assume that $v = E^T BNu$ and $\|v\| = 1$, where u is an eigenvector of the above quadratic eigenvalue problem, then

$$\begin{aligned} u^T \widetilde{D} u &= v^T D^{-1} v \\ &= \frac{v_1^2}{d_1} + \frac{v_2^2}{d_2} + \cdots + \frac{v_m^2}{d_m} \geq \frac{v^T v}{d_1} \\ &= \frac{1}{\|C\|}. \end{aligned}$$

Hence,

$$\frac{1}{u^T \widetilde{D} u} \leq \|C\|.$$

Let $\zeta = u^T \widetilde{D} u$, $\delta = u^T N^T G N u$ and $\varepsilon = u^T N^T A N u$, then (3.4) becomes

$$\lambda^2 \zeta - \lambda(\delta + 2\zeta) + \varepsilon + \zeta = 0.$$

From Theorem 3.3, λ must satisfy

$$\lambda = 1 + \frac{\delta}{2\zeta} \pm \mu, \quad \mu \leq \sqrt{2}\max\left\{\frac{\delta}{2\zeta}, \sqrt{\frac{|\delta - \varepsilon|}{\zeta}}\right\}.$$

Now $\delta \leq c\|N^T G N\|$, $\varepsilon \leq c\|N^T A N\|$, where c is an upper bound on $\|u\|$ and u are eigenvectors of (2.8) subject to $\|E^T B N u\| = 1$. Hence, the eigenvalues of (2.8) subject to $E^T B N u \neq 0$ satisfy

$$|\lambda - 1| = \mathcal{O}\big(\max\big\{\|C\|, \|G - A\|\sqrt{\|C\|}\big\}\big)$$

for small values of $\|C\|$. □

The results of this theorem are not very surprising, but basic eigenvalue perturbation theorems such as Theorem 7.7.2 in [10] in conjunction with Theorem 2.3 of [16] are weaker than what we have established. Specifically, the structure of our coefficient matrix and preconditioner means that we are still guaranteed to have $2(m - p)$ unit eigenvalues, whereas the more general eigenvalue perturbation theorems would only imply that these eigenvalues will be close to 1.

Example 3.6 (C with small entries) Suppose that $\mathcal{A}_C$ and $\mathcal{P}$ are as in Example 2.2, but $C = [10^{-a}]$ for some positive real number a. Setting $D = [10^{-a}]$ and $E = [1]$ ($C = EDE^T$), the quadratic eigenvalue problem (2.8) is

$$\left(\lambda^2 \begin{bmatrix} 10^a & 0 \\ 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 2 + 2 \times 10^a & 0 \\ 0 & 2 \end{bmatrix} + \begin{bmatrix} 1 + 10^a & 0 \\ 0 & 1 \end{bmatrix}\right) \begin{bmatrix} x_y \\ x_z \end{bmatrix} = 0.$$

This quadratic eigenvalue problem has three finite eigenvalues: $\lambda = \frac{1}{2}$,

$$\lambda = 1 + 10^{-a} \pm 10^{-a}\sqrt{1 + 10^a}.$$

For large values of a, $\lambda \approx 1 + 10^{-a} \pm 10^{-\frac{a}{2}}$; the eigenvalues will be close to 1.

This clustering of part of the spectrum of $\mathcal{P}^{-1}\mathcal{A}_C$ will often translate into a speeding up of the convergence of a selected Krylov subspace method, [1, Sect. 1.3].

3.2 Numerical examples

We will carry out several numerical tests to verify that, in practice, our theoretical results translate to a speeding up in the convergence of a selected Krylov subspace method as the entries of C converge towards 0.

Example 3.7 The CUTEr test set [13] provides a set of quadratic programming problems. We shall use the problem CVXQP2_M in the following two examples. This problem has $n = 1000$ and $m = 250$. "Barrier" penalty terms (in this case α, where α is defined below) are added to the diagonal of A to simulate systems that might arise during an iteration of an interior-point method for such problems. We shall set

$G = \text{diag}(A)$ (ignoring the additional penalty terms), and $C = \alpha I$, where α is a positive, real parameter that we will change.

All tests were performed on a dual Intel Xeon 3.20 GHz machine with hyper-threading and 2 GByte of RAM. It was running Fedora Core 2 (Linux kernel 2.6.8) with MATLAB® 7.0. We solve the resulting linear systems with unrestarted GMRES [10], the Projected Preconditioned Conjugate Gradient (PPCG) method [6, Algorithm 5.5.2] and the Simplified Quasi-Minimal Residual (SQMR) method [9].[1] We terminate the iteration when the value of residual is reduced by at least a factor of 10^{-8} and always use $\mathcal{P}$ and $\mathcal{P}_C$ as left preconditioners. We emphasize that for the PPCG method knowledge of the eigenvalues is all you need to describe convergence whereas Greenbaum, Pták and Strakoš show that this is not generally the case with GMRES [15].

In Fig. 1 we compare the performance (in terms of iteration count) between using a preconditioner of the form $\mathcal{P}$ and one of the form $\mathcal{P}_C$, Eqs. (1.3) and (1.2) respectively for the three different iterative methods. Although the SQMR method doesn't have an optimality property as was assumed in Sect. 3, as α becomes smaller, we hope that the difference between the number of iterations required by the two preconditioners decreases. We observe that, for this example, once $\alpha \leq 10^{-4}$ there is little benefit in reproducing C in the preconditioner in any of the iterative methods tested. However, the SQMR method requires around 900 iterations when $\alpha \ll 1$, whilst PPCG and GMRES require just 500 iterations to reach the desired tolerance. We would expect the PPCG and GMRES methods to take around 500 iterations because the preconditioned system has 500 unit eigenvalues and a further 500 clustered about one when $\alpha \ll 1$; the remaining 500 eigenvalues lie away from the unit eigenvalues. The SQMR method does not satisfy an optimality condition and, in this and the following example, this results in substantially more than 500 iterations being required to reach the desired tolerance when $\alpha \ll 1$.

In this example, when $\alpha \approx 1$ and the preconditioned system $\mathcal{P}^{-1}\mathcal{A}_C$ has additional eigenvalues clustered around 1 above those $2m - p$ guaranteed to lie at 1. However, as α decreases, this eigenvalues move away from 1 which results in the number of iterations to increase.

Example 3.8 In this example we again use the CVXQP2_M problem from the CUTEr test set. The only difference to the above example is that we shall set $C = \alpha \times \text{diag}(0, \ldots, 0, 1, \ldots, 1)$, where $\text{rank}(C) = \lfloor m/2 \rfloor$.

In Fig. 2 we compare the performance (in terms of iteration count) between using a preconditioner of the form $\mathcal{P}$ and one of the form $\mathcal{P}_C$, Eqs. (1.3) and (1.2) respectively for our chosen iterative methods. We observe that if $\alpha \approx 1$, then fewer iterations are required in Fig. 2 than in Fig. 1 to reach the required tolerance — this is as we would expect because of there now being a guarantee of at least 250 unit eigenvalues in the preconditioned system compared to the possibility of none. However, as α approaches 0, the number of eigenvalues clustered around 1 will converge to be the same as in Example 3.7. We observe from Figs. 1 and 2 that the number of iterations to reach the required tolerance is, as expected, converging to be the same as $\alpha \to 0$.

[1] MATLAB® code for SQMR can be obtained from the MATLAB® Central File Exchange at http://www.mathworks.fr/matlabcentral/.

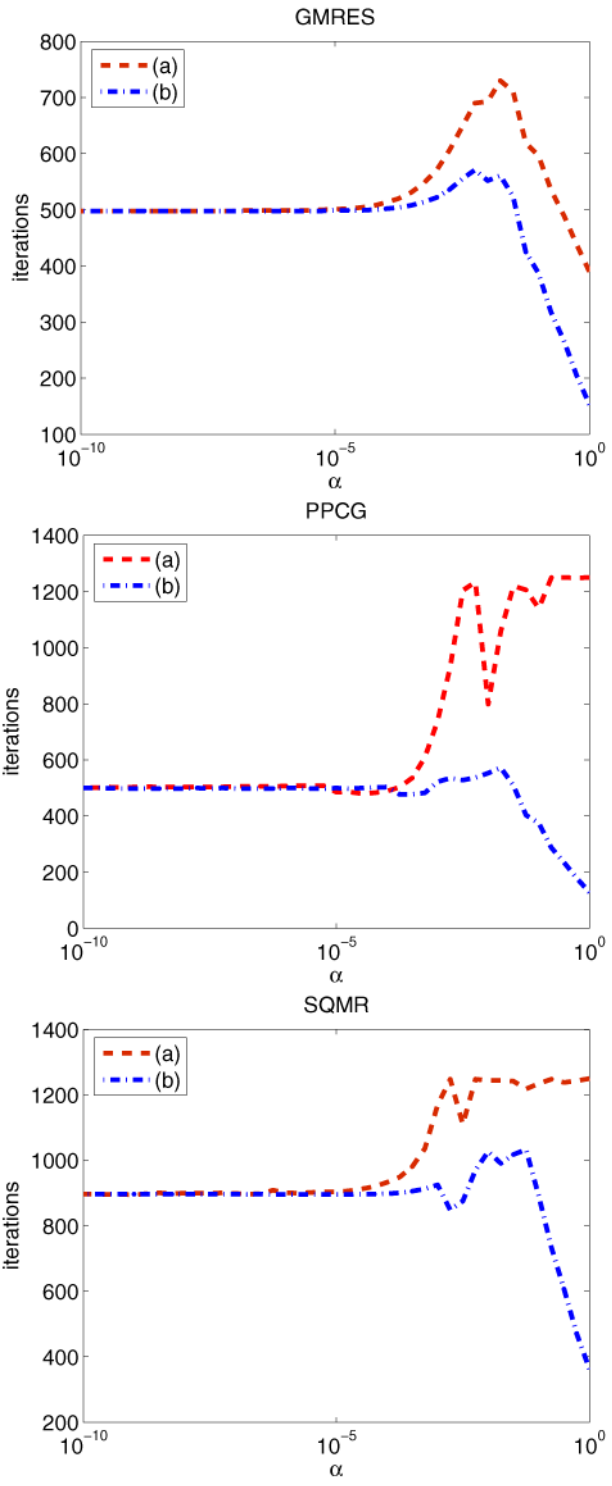

Fig. 1 Comparison of number of iterations required when either (**a**) $\mathcal{P}$ or (**b**) $\mathcal{P}_C$ are used as preconditioners for $C = \alpha I$ with GMRES, PPCG and SQMR on the CVXQP2_M problem

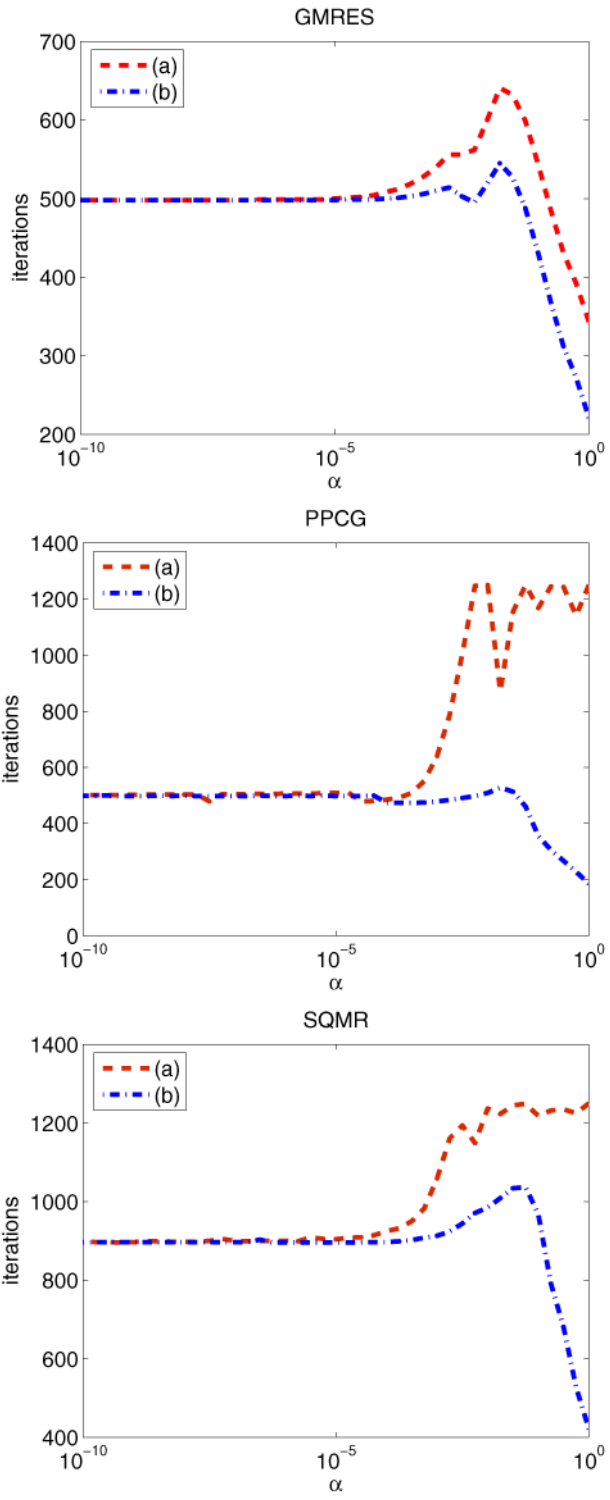

Fig. 2 Comparison of number of iterations required when either (**a**) $\mathcal{P}$ or (**b**) $\mathcal{P}_C$ are used as preconditioners for $C = \alpha \times \text{diag}(0, \ldots, 0, 1, \ldots, 1)$, where $\text{rank}\, C = \lfloor m/2 \rfloor$, with GMRES, PPCG and SQMR on the CVXQP2_M problem

Example 3.9 AUG2DQP is another test problem from the CUTEr test set. This problem has $n = 3280$ and $m = 1600$. "Barrier" penalty terms (in this case α, where α is defined below) are added to the diagonal of A to simulate systems that might arise during an iteration of an interior-point method for such problems. We shall set $G = \text{diag}(A)$ (ignoring the additional penalty terms), and $C = \alpha I$, where α is a positive, real parameter that we will change. In Fig. 3 we observe that once $\alpha \leq 10^{-4}$ there is little benefit in reproducing C in the preconditioner for the PPCG method. Similarly, when $C = \alpha \times \text{diag}(0, \ldots, 0, 1, \ldots, 1)$, where $\text{rank}(C) = \lfloor m/2 \rfloor$, there is little benefit in reproducing C in the preconditioner for the PPCG method when $\alpha \leq 10^{-4}$, Fig. 4.

These examples suggest that during premultply-asymptotic iterations of an interior point method for a nonlinear programming problem, we may need to use a preconditioner of the form $\mathcal{P}_C$, but as the method proceeds there will be a point at which we will be able to swap to using a preconditioner of the form $\mathcal{P}$. From this point

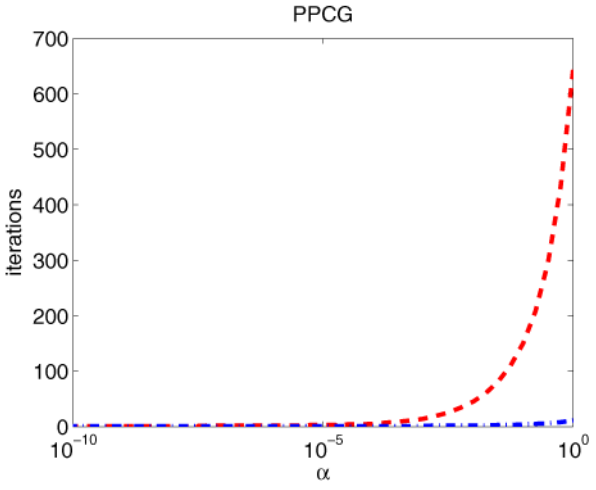

Fig. 3 Number of PPCG iterations when either (**a**) $\mathcal{P}$ or (**b**) $\mathcal{P}_C$ are used as preconditioners for $C = \alpha I$ on the AUG2DQP problem

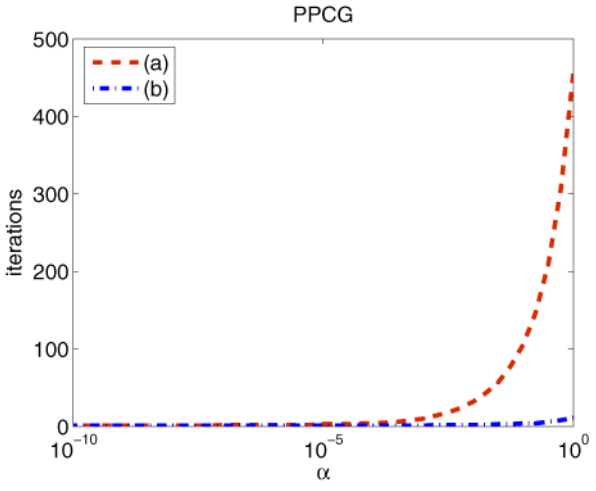

Fig. 4 Number of PPCG iterations when either (**a**) $\mathcal{P}$ or (**b**) $\mathcal{P}_C$ are used as preconditioners for $C = \alpha \times \text{diag}(0, \ldots, 0, 1, \ldots, 1)$, where rank $C = \lfloor m/2 \rfloor$, on the AUG2DQP problem

onwards, we'll be able to use the same preconditioner during each iterative solve of the resulting sequence of saddle-point problems.

4 Conclusion and further research

In this paper, we have investigated a class of preconditioners for indefinite linear systems that incorporate the (1,2) and (2,1) blocks of the original matrix. These blocks are often associated with constraints. We have shown that if C has rank $p > 0$, then the preconditioned system has at least $2(m - p)$ unit eigenvalues, regardless of the structure of G. In addition, we have shown that if the entries of C are very small, then we will expect an additional $2p$ eigenvalues to be clustered around 1 and, hence, for the number of iterations required by our chosen Krylov subspace method to be dramatically reduced. These later results are of particular relevance to interior point methods for optimization.

The practical implications of the analysis of this paper in the context of solving nonlinear programming problems will be the subject of a follow-up paper. We will investigate the point at which the user should switch from using a preconditioner of the form $\mathcal{P}_C$ to that of $\mathcal{P}$ during an interior point method, and how the sub-matrix G in the preconditioner should be chosen.

Acknowledgements The authors would like to thank the referees for their helpful comments.

References

1. Axelsson, O., Barker, V.A.: Finite Element Solution of Boundary Value Problems. Theory and Computation, Classics in Applied Mathematics, vol. 35. SIAM, Philadelphia (2001). Reprint of the 1984 original
2. Benzi, M., Golub, G.H., Liesen, J.: Numerical solution of saddle point problems. Acta Numer. **14**, 1–137 (2005)
3. Bergamaschi, L., Gondzio, J., Zilli, G.: Preconditioning indefinite systems in interior point methods for optimization. Comput. Optim. Appl. **28**, 149–171 (2004)
4. Cafieri, S., D'Apuzzo, M., De Simone, V., di Serafino, D.: On the iterative solution of KKT systems in potential reduction software for large-scale quadratic problems. Technical Report 09/2004, Dept. of Mathematics, Second University of Naples (December 2004). Comput. Optim. Appl. (to appear)
5. Dollar, H.S.: Constraint-style preconditioners for regularized saddle-point problems. Technical Report 3/2006, Dept. of Mathematics, University of Reading (March 2006). SIAM J. Matrix Anal. Appl. (to appear)
6. Dollar, H.S.: Iterative linear algebra for constrained optimization. Thesis of Doctor of Philosophy. Oxford University (2005)
7. Elman, H.C., Silvester, D.J., Wathen, A.J.: Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics. Oxford University Press, Oxford (2005)
8. Forsgren, A., Gill, P.E., Griffin, J.D.: Iterative solution of augmented systems arising in interior methods. Technical Report NA-05-03, University of California, San Diego (August 2005)
9. Freund, R.W., Nachtigal, N.M.: A new Krylov-subspace method for symmetric indefinite linear systems. In: Ames, W.F. (ed.) Proceedings of the 14th IMACS World Congress on Computational and Applied Mathematics, pp. 1253–1256. IMACS (1994)
10. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn., Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore (1996)
11. Gould, N.I.M.: On the accurate determination of search directions for simple differentiable penalty functions. IMA J. Numer. Anal. **6**, 357–372 (1986)

12. Gould, N.I.M., Hribar, M.E., Nocedal, J.: On the solution of equality constrained quadratic programming problems arising in optimization. SIAM J. Sci. Comput. **23**, 1376–1395 (2001)
13. Gould, N.I.M., Orban, D., Toint, P.L.: CUTEr and SifDec: a constrained and unconstrained testing environment, revisited. ACM Trans. Math. Software **29**, 373–394 (2003)
14. Greenbaum, A.: Iterative Methods for Solving Linear Systems. Frontiers in Applied Mathematics, vol. 17. SIAM, Philadelphia (1997)
15. Greenbaum, A., Pták, V., Strakoš, Z.: Any nonincreasing convergence curve is possible for GMRES. SIAM J. Matrix Anal. Appl. **17**, 465–469 (1996)
16. Keller, C., Gould, N.I.M., Wathen, A.J.: Constraint preconditioning for indefinite linear systems. SIAM J. Matrix Anal. Appl. **21**, 1300–1317 (2000)
17. Lukšan, L., Vlček, J.: Indefinitely preconditioned inexact Newton method for large sparse equality constrained non-linear programming problems. Numer. Linear Algebra Appl. **5**, 219–247 (1998)
18. Perugia, I., Simoncini, V.: Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations. Numer. Linear Algebra Appl. **7**, 585–616 (2000)
19. Rozložník, M., Simoncini, V.: Krylov subspace methods for saddle point problems with indefinite preconditioning. SIAM J. Matrix Anal. Appl. **24**, 368–391 (2002)
20. Saad, Y.: Iterative Methods for Sparse Linear Systems, 2nd edn. SIAM, Philadelphia (2003)
21. Saad, Y., Schultz, M.H.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **7**, 856–869 (1986)
22. Tisseur, F., Meerbergen, K.: The quadratic eigenvalue problem. SIAM Rev. **43**, 235–286 (2001)
23. Toh, K.-C., Phoon, K.-K., Chan, S.-H.: Block preconditioners for symmetric indefinite linear systems. Int. J. Numer. Methods Eng. **60**, 1361–1381 (2004)
24. Wright, S.J.: Primal–Dual Interior-Point Methods. SIAM, Philadelphia (1997)