AINV$\mathcal{K}$: a Class of Approximate Inverse Preconditioners based on Krylov-subspace methods, for Large Indefinite Linear Systems*

Giovanni Fasano
fasano@unive.it
Dipartimento di Management
Università Ca'Foscari Venezia

Massimo Roma
roma@dis.uniroma1.it
Dipartimento di Ingegneria Informatica,
Automatica e Gestionale "A. Ruberti"
SAPIENZA, Università di Roma

The Italian Ship Model Basin - INSEAN, CNR

Abstract. We propose a class of preconditioners for symmetric linear systems arising from numerical analysis and nonconvex optimization frameworks. Our preconditioners are specifically suited for *large* indefinite linear systems and may be obtained as *by-product* of Krylov-subspace solvers, as well as by applying L-BFGS updates. Moreover, our proposal is also suited for the solution of a sequence of linear systems, say $Ax = b_i$ or $A_i x = b_i$, where respectively the right-hand side changes or the system matrix slightly changes, too. Each preconditioner in our class is identified by setting the values of a pair of parameters and a scaling matrix, which are user-dependent, and may be chosen according to the structure of the problem in hand. We provide theoretical properties of our preconditioners, discussing the relation with the proposals in [19, 25]. In particular, we show that our preconditioners both shift some eigenvalues of the indefinite system matrix to ± 1, and are able to control the condition number of the preconditioned matrix. We study some structural properties of our class of preconditioners, and report the results on a comparative numerical experience with *LMP* preconditioners [19]. The experience is carried on first considering some relevant linear systems proposed in the literature. Then, we embed our preconditioners within a linesearch-based truncated Newton method, where sequences of linear systems (namely Newton's equations), are required to be solved. We perform an extensive numerical testing over the entire large scale unconstrained optimization test set of *CUTEr* collection [18], confirming the efficiency of our proposal.

Keywords: Preconditioners, large indefinite linear systems, large scale nonconvex optimization, Krylov-subspace methods.

1 Introduction

We study a class of preconditioners for the solution of the symmetric indefinite linear system

$$Ax = b, \qquad A \in \mathbb{R}^{n \times n}, \quad A = A^T,$$

where n is *large* and we do not assume any sparsity pattern for the system matrix A. The solution of large linear systems is sought in a variety of real applications and in different contexts. Moreover, the use of preconditioning is an essential issue to improve the efficiency of iterative solvers. Within different frameworks of complex systems, the solution of sequences of large linear systems often comes up, too. E.g., we often encounter sequences like

$$Ax = b_i \qquad \text{or} \qquad A_i x = b_i, \tag{1.1}$$

where $\{b_i\}$ are possibly arbitrary and the matrices $\{A_i\}$ are *slowly* varying with the index 'i'. Numerical Analysis and Optimization give plenty of frameworks where the solution of a sequence of large linear systems is sought. Truncated Newton methods in unconstrained optimization, KKT systems, interior point methods, and

* G.Fasano wishes to thank the Italian Ship Model Basin, CNR - INSEAN institute, for the indirect support.

PDE-constrained optimization are just some examples. Similarly, several real applications, ranging from power systems networks to economic models and queuing systems, involve the solution of large linear systems.

Typically, up to one decade ago, the specialized literature was keen on privileging the use of direct methods when n was moderately small, in view to their reasonable cost, since $O(n^3)$ might be unaffordable for large n. However, we have more recently observed an increasing blurred use of techniques, in both sparse direct methods and iterative algorithms, in order to efficiently solve linear systems (see e.g. [2, 4]). Observe that for linear systems where the matrix A is block-diagonal or banded, which typically arise when solving discretized PDEs, specific solvers from the literature can be used [21], which require to include effective preconditioning strategies, too.

In this paper we focus on the use of iterative methods to solve linear systems: the iterative techniques are also used to provide sufficient information on the system matrix, in order to generate the preconditioners.
We propose a general class of preconditioners, which uses information collected by any Krylov-subspace method or possibly using L-BFGS updates, in order to capture the structural properties of the system matrix.
In particular, we iteratively construct our preconditioners either by using (but not performing) a factorization of the system matrix (see, e.g. [13, 17, 34]), obtained as by product of Krylov-subspace methods, or performing a Jordan Canonical decomposition on a *very small size* matrix. We prove theoretical properties for such preconditioners, and describe results which indicate how to possibly select the parameters involved in the definition of our class of preconditioners. Our proposal is able to shift some eigenvalues of the preconditioned system matrix to either -1 or $+1$. Moreover, we show that we can partially control the condition number of the preconditioned matrix, with respect to the system matrix, by introducing some care when choosing specific parameters.

The basic idea of our approach draws its inspiration from *Approximate Inverse Preconditioners*, which have proved in general to be remarkably robust and efficient in practice [2, 3]. These methods claim that in principle, an approximate inverse of A should be computed and used as a preconditioner. Though in practice it might be difficult to ensure that the approximate inverse is sparse, suitable factorizations of matrix A can be fruitfully exploited, in order to build the approximate inverse preconditioner. In particular, a generalization of the Gram-Schmidt process can be used to provide a triangular factorization of A^{-1}, where the triangular matrices are in general dense. This is the basic idea of AINV preconditioner (see [2], Section 5.1.2).

In this paper we apply any Krylov-subspace method to generate a triangular factorization of A^{-1}. The latter is then used to build our preconditioners, namely the AINV$\mathcal{K}$ class, needing to store just a few vectors, without requiring any matrix storage and any product of matrices. As we collect information from Krylov-subspace methods, we assume that the entries of the system matrix are not stored at once and the necessary information is gained by simply using a routine, which computes the product of the system matrix times a vector. Note that, typically, the product of a matrix times a vector allows fast parallel computing, which is another possible advantage of our approach, in large scale settings.

AINV$\mathcal{K}$ can be naturally extended to the solution of a sequence of large linear systems. When sequences of systems are tackled, we generate the preconditioner $\mathcal{P}$, for the solution of the first linear system in the sequence, i.e. $Ax = b_1$ or $A_1x = b_1$. Then, we apply $\mathcal{P}$ for solving either $Ax = b_i$ or $A_ix = b_i$, $i = 2, 3, \ldots$ Thus, the cost of computing $\mathcal{P}$, for $i = 1$, is repaid by accelerating the solution for $i = 2, 3, \ldots$; a similar strategy was proposed in [25]. The latter approach might be strongly advantageous in numerical analysis and optimization frameworks, where the cost for computing the preconditioner is relatively small, with respect to solving each linear system in the sequence. Furthermore, when a Krylov-subspace method is adopted to compute the preconditioner, the full storage of system matrix is never required. On the other hand, the same Krylov-subspace method might be used also to compute the solution of the linear system (see also [32, 33]).

We experience our preconditioners AINV$\mathcal{K}$ on test problems from both numerical analysis and nonconvex optimization. In particular, we first test them on significant linear systems, from both the literature and real applications. Then, we focus on the so called *Newton–Krylov methods*, also known as (Hessian-free) truncated Newton methods (see e.g. [23] and [28] for a survey on the importance of preconditioning in truncated Newton methods). In the latter case we tackle a sequence of linear systems as in (1.1). In this context, both positive definite and indefinite linear systems are considered (for preconditioning indefinite linear systems in Interior Point methods see also [5]).

Unlike following the idea early developed in [30], where a full-memory quasi-Newton formula is adopted for the preconditioner, we show that a few iterations of any Krylov-subspace method can be used, in order to provide information for building our preconditioners. Even if the resulting matrices are not sparse, they allow to consider preconditioning also for large scale problems, by simply storing k vectors, with $k \ll n$.

We recall that in case the optimization problem in hand is nonconvex, i.e. the Hessian matrix of the objective function is possibly indefinite and at least one eigenvalue is negative, the solution of Newton's equations within truncated Newton schemes may claim for some cares. Indeed, the Krylov-subspace method used to solve

Newton's equation should be suitably applied considering that, unlike in numerical analysis, optimization frameworks require the definition of *gradient-related descent directions*, which have to satisfy additional properties (see e.g. [9, 29]). In this regard our proposal provides a tool, in order to preserve the latter properties.

As regards a comparison with the current literature, we show in the paper that, to a large extent, AINV$\mathcal{K}$ has several similarities with the efficient *LMP* preconditioners, which were recently introduced in [19]. In particular, the latter schemes encompass also the proposal in [25], and are based on the use of L-BFGS updates, in order to provide an approximate inverse of the system matrix. We first highlight some analogies between the AINV$\mathcal{K}$ and *LMP* classes. Then, we report a theoretical comparison between them, and carry on an extensive numerical experience comparing their performance. Observe that with respect to *LMP*, our proposal needs a reduced memory storage and involves a reduced computational burden. In addition, on one hand when the system matrix is positive definite *LMP*s show some slightly stronger theoretical properties, in order to control the condition number of the preconditioned matrix. On the other hand, the numerical experience proves the strong likelyhood of AINV$\mathcal{K}$ and *LMP* classes.
Moreover, AINV$\mathcal{K}$ retains great generality, since it may be applied also when the system matrix is indefinite. Further generality is also provided by AINV$\mathcal{K}$ through the dependency on several user-dependent parameters. Finally, we recall that in place of Krylov-subspace methods, in principle also L-BFGS updates can be used to build our preconditioners, so that they may be easily embedded within different numerical frameworks.

The paper is organized as follows: Section 2 reports some preliminaries and Section 3 contains the definition of the AINV$\mathcal{K}$ class, along with some of its spectral properties. Section 4 completes the properties of our proposal, while in Section 5 we carry on a theoretical comparison between AINV$\mathcal{K}$ and *LMP* preconditioners. In Section 6 we study more in depth the condition number of the preconditioned matrix MA, where M belongs to the class AINV$\mathcal{K}$. Finally, in Section 7 we report the results of an extensive numerical experience and Section 8 adds some conclusions.

As regards the notations, for a $n \times n$ real matrix A we denote by $\Lambda[A]$ the spectrum of A. I_k is the identity matrix of order k. With $C \succ 0$ we indicate that the matrix C is positive definite, $tr[C]$, $rk[C]$ and $det[C]$ are the *trace*, the *rank* and the *determinant* of C, respectively, while $\kappa(C)$ indicates the condition number of C. Finally, $\|\cdot\|$ denotes the Euclidean norm, e_h is the $h-$th unit vector and $O(\cdot)$ is used for the standard 'big O' notation.

2 Preliminaries

In this section we first introduce some preliminaries, then we propose our class of preconditioners. Consider the *indefinite* linear system

$$Ax = b, \tag{2.1}$$

where $A \in \mathbb{R}^{n\times n}$ is *symmetric*, n is *large* and $b \in \mathbb{R}^n$. Some real contexts where the latter system requires efficient solvers are detailed in Section 1. Suppose any Krylov-subspace method is used for the solution of (2.1), e.g. the Lanczos process (SYMMLQ, MINRES [31]) or the Conjugate Gradient (CG) method [17, 20] (but Planar-CG methods [11, 20] may be also an alternative choice). Here, we refer to the 'Lanczos process' as to the generalized Gram-Schmidt method which provides a reduction of (2.1) to the tridiagonal system $Tx = \bar{b}$, where T possibly admits a factorization. As well known, the Lanczos process and the CG are equivalent as long as $A \succ 0$, whereas the CG, though cheaper, in principle may not cope with the indefinite case.

2.1 The matrix factorization we use

In the next Assumption 2.1 we consider that a finite number of steps, say $h \ll n$, of the Krylov-subspace method adopted have been performed. With reference to the definition in [16, 35], we say that a symmetric indefinite matrix T is *factorizable* if the diagonal (or 2×2 block diagonal) matrix D and the unit lower triangular matrix L exist such that $T = LDL^T$.

Assumption 2.1 [Factorization] *Let us consider any Krylov-subspace method to solve the symmetric linear system (2.1). Suppose at step h of the Krylov-subspace method, with $h \leq n-1$, the matrices $R_h \in \mathbb{R}^{n\times h}$, $T_h \in \mathbb{R}^{h\times h}$ and the vector $u_{h+1} \in \mathbb{R}^n$ are generated, such that*

$$AR_h = R_h T_h + \rho_{h+1} u_{h+1} e_h^T, \qquad \rho_{h+1} \in \mathbb{R}. \tag{2.2}$$

Suppose that for the matrix T_h there exists the following decomposition:

$$T_h = \begin{cases} L_h D_h L_h^T, & \text{if } T_h \text{ is factorizable} \\ \\ V_h B_h V_h^T, & \text{if } T_h \text{ is not factorizable,} \end{cases} \tag{2.3}$$

where

$R_h = (u_1 \cdots u_h)$, $u_i^T u_j = 0$, $\|u_i\| = 1$, $1 \le i \ne j \le h+1$,

T_h *is tridiagonal, irreducible, nonsingular, with eigenvalues not all coincident,*

D_h *is* 1×1 *or* 2×2 *block diagonal,* L_h *is unit lower bidiagonal,*

$B_h = diag_{1 \le i \le h}\{\mu_i\}$, $V_h = (v_1 \cdots v_h) \in \mathbb{R}^{h \times h}$ *orthogonal,* (μ_i, v_i) *is eigenpair of* T_h.

To have a better intuition on the reason for which any Krylov-subspace method can be adopted to satisfy Assumption 2.1, we remark that they are essentially all based on the generation of orthogonal vectors (the Lanczos vectors), used to transform the system (2.1) into a tridiagonal one. Then, they substantially differ only in the way the resulting tridigonal system is solved. It is indeed the iterative reduction process to a tridiagonal form to be essential for our analysis. On the contrary, for our analysis the hypothesis of having a tridiagonal matrix is not really indispensable, though it helps to simplify the computation.

Remark 2.1 Note that the commonest Krylov-subspace methods satisfy (2.2) and the CG/Lanczos process satisfies also (2.3) with $T_h = L_h D_h L_h^T$, when T_h is factorizable.
In particular, also observe that from (2.2) we have $T_h = R_h^T A R_h$, so that whenever $A \succ 0$ then $T_h \succ 0$. Since the Jordan Canonical form of T_h in (2.3) is required only when the factorization $L_h D_h L_h^T$ of T_h is unstable, it is important to check whenever the latter condition occurs, without computing the eigenpairs of T_h if unnecessary. On this purpose, note that for instance the CG method, the Lanczos process and the Planar-CG methods either provide relation $T_h = L_h D_h L_h^T$, with D_h block diagonal (blocks can be 1×1 or 2×2 at most) or they detect that the latter factorization becomes unstable (see also [13, 32, 34] and the algorithm in LAPACK routine `ssysv`, by Bunch and Kaufman [7]). Thus, checking the eigenvalues of D_h will suggest if the Jordan Canonical form $T_h = V_h B_h V_h^T$ is really needed for T_h.

To complete this remark, it is worth to highlight that also L-BFGS quasi-Newton scheme may provide information in order to satisfy Assumption 2.1 and build our preconditioners. Indeed, according with [19], and using the correspondence between BFGS and the CG when A is positive definite, in solving (2.1) a set of h conjugate directions $p_1, \ldots, p_h$ (and the vectors $Ap_1, \ldots, Ap_h$) can easily be computed after h iterations of L-BFGS. Now, following the guidelines in [34], it is not difficult to see that after a brief computation, the vectors

$$r_1 = p_1$$
$$r_{i+1} = r_i - \frac{p_i^T r_i}{p_i^T A p_i} A p_i, \qquad i = 1, \ldots h-1$$

yield a set or orthogonal vectors, which can be used to provide T_h and both relations (2.2) and (2.3). Thus, in practice any iterative method commonly used for solving (2.1) may give, as by product, the information necessary to satisfy Assumption 2.1.

Remark 2.2 The Krylov-subspace method adopted may, in general, perform $m \ge h$ iterations, generating the orthonormal vectors $u_1, \ldots, u_m$. Then, we can set $R_h = (u_{\ell_1}, \ldots, u_{\ell_h})$, where $\{\ell_1, \ldots, \ell_h\} \subseteq \{1, \ldots, m\}$, and change relations (2.2)-(2.3) accordingly; i.e. Assumption 2.1 may hold selecting any h out of the m vectors (among $u_1, \ldots, u_m$) computed by the Krylov-subspace method. We are going to further study the latter issue, after giving some properties of our class of preconditioners.

Remark 2.3 For relatively small values of the parameter h in Assumption 2.1 (say $h \le 20$, as often suffices in most of the applications), the *worst-case* computation of the eigenpairs (μ_i, v_i), $i = 1, \ldots, h$, of T_h (i.e. whenever T_h is not factorizable) may be extremely fast, by using standard codes. E.g. if the CG is the Krylov-subspace method used in Assumption 2.1 to solve (2.1), then the Matlab [24] (general) function `eigs()` requires as low as $\approx 10^{-4}$ seconds to fully compute all the eigenpairs of T_h, for $h = 20$, on a commercial laptop. In the latter case indeed, the matrix T_h is tridiagonal. Nonetheless, in the separate paper [12] we consider a special case where the request (2.3) on T_h may be considerably weakened under mild assumptions.

Observe also that from Assumption 2.1 the parameter ρ_{h+1} may be possibly nonzero, i.e. the subspace $span\{u_1, \ldots, u_h\}$ is possibly not an invariant subspace under the transformation by matrix A. Thus, in this paper we consider a more general case with respect to [1].

3 Our class of preconditioners AINV$\mathcal{K}$: basic spectral properties

On the basis of Assumption 2.1, we can now define our preconditioners and show their properties. To this aim, suppose $T_h = L_h D_h L_h^T$ in (2.3) (i.e. T_h is factorizable), where $D_h = diag_{1\le j\le m}\{E_j^h\}$, with $E_j^h \in \mathbb{R}$ or $E_j^h \in \mathbb{R}^{2\times 2}$. Moreover, if $E_j^h \in \mathbb{R}^{2\times 2}$, assume that we compute the decomposition

$$E_j^h = U_j D_j^h U_j^T, \tag{3.1}$$

with $D_j^h = diag\{d_{j_1}, d_{j_2}\}$ and $U_j^T U_j = I_2$. On the other hand, if $E_j^h \in \mathbb{R}$ for an index j, for the sake of notation we again assume that (3.1) holds, setting

$$D_j^h \equiv d_{j_1} \equiv E_j^h \qquad \text{and} \qquad U_j = 1.$$

Then, we define (see also [15]) the matrices $|D_h| = \ diag_{1\le j\le m}\left\{U_j \cdot diag\{|d_{j_1}|, |d_{j_2}|\} \cdot U_j^T\right\}$, $|B_h| = \ diag_{1\le i\le h}\{|\mu_i|\}$ and

$$|T_h| \stackrel{\text{def}}{=} \begin{cases} L_h|D_h|L_h^T, & \text{if } T_h \text{ is factorizable} \\ \\ V_h|B_h|V_h^T, & \text{if } T_h \text{ is not factorizable.} \end{cases}$$

Observe that of course $|T_h| = T_h$ in case T_h is positive definite. Furthermore, it is easily seen that $|T_h|$ is positive definite, for any h, and $|T_h|^{-1}T_h^2|T_h|^{-1} = I_h$ whenever $T_h \succ 0$ or T_h is not factorizable.

As a consequence, if T_h is factorizable we have $T_h|T_h|^{-1} = (|T_h|^{-1}T_h)^T = L_h\hat{I}_h L_h^{-1}$, and if T_h is not factorizable we have $T_h|T_h|^{-1} = |T_h|^{-1}T_h = V_h\hat{I}_h V_h^T$, where

$$\hat{I}_h = diag_{1\le i\le h}\{\sigma_i\}, \qquad \sigma_i \in \{-1, +1\}. \tag{3.2}$$

Now let us introduce the following $n \times n$ matrix, which *depends on the parameter* 'a':

$$\begin{aligned} M_h &= (I_n - R_h R_h^T) + R_h|T_h|R_h^T + a\left(u_{h+1}u_h^T + u_h u_{h+1}^T\right), \qquad h \le n-1, \\ &= [R_h \mid u_{h+1} \mid R_{n,h+1}] \left[\begin{array}{c|c} \left(\begin{array}{c|c} |T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right) & 0 \\ \hline 0 & I_{n-(h+1)} \end{array}\right] \left[\begin{array}{c} R_h^T \\ \hline u_{h+1}^T \\ \hline R_{n,h+1}^T \end{array}\right] \end{aligned} \tag{3.3}$$

$$M_n = (I_n - R_n R_n^T) + R_n|T_n|R_n^T \ = \ R_n|T_n|R_n^T, \tag{3.4}$$

where R_h and T_h satisfy relations (2.2)-(2.3), $a \in \mathbb{R}$, the matrix $R_{n,h+1} \in \mathbb{R}^{n\times[n-(h+1)]}$ satisfies relation $R_{n,h+1}^T R_{n,h+1} = I_{n-(h+1)}$ and $[R_h \mid u_{h+1} \mid R_{n,h+1}]$ is orthogonal. By (3.3), when $h \le n-1$, the matrix M_h is the sum of three terms. It is easily seen that $I_n - R_h R_h^T$ represents a projector onto the subspace $\mathcal{S}_{n-h}$ orthogonal to the range of matrix R_h, so that $M_h v = v + a(u_{h+1}^T v)u_h$, for any $v \in \mathcal{S}_{n-h}$. Thus, for any $v \in \mathcal{S}_{n-h}$, when either $u_{h+1}^T v = 0$ or $a = 0$, then $M_h v = v$ (or equivalently if M_h is nonsingular $M_h^{-1}v = v$), i.e. the vector v is unaltered by applying M_h (or M_h^{-1}). As a result, if either $a = 0$ or $u_{h+1}^T v = 0$ then M_h behaves as the identity matrix for any vector $v \in \mathcal{S}_{n-h}$.

The term $R_h|T_h|R_h^T$ in (3.3) may be interpreted as a suitable approximation of matrix $|A|$ over the Krylov-subspace $span\{u_1, \ldots, u_h\}$. Indeed, multiplying (2.2) on the left by $R_h^T A^{-1}$ we also have

$$I_h = (R_h^T A^{-1} R_h)T_h + \rho_{h+1} R_h^T A^{-1} u_{h+1} e_h^T,$$

and when $\rho_{h+1} \approx 0$ we have $T_h \approx (R_h^T A^{-1} R_h)^{-1}$. In particular, when $h = n$ then $\rho_{h+1} = 0$, $u_{h+1} = 0$ and the matrix R_h is orthogonal, so that $T_n = (R_n^T A^{-1} R_n)^{-1}$.

Considering the matrix M_h in (3.3)-(3.4) we are now ready to introduce the following parameter dependent class of preconditioners

$$\begin{aligned} M_h^\sharp(a, \delta, D) &\stackrel{\text{def}}{=} D\left[I_n - (R_h \mid u_{h+1})(R_h \mid u_{h+1})^T\right] D^T \\ &\quad + (R_h \mid Du_{h+1}) \left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right)^{-1} (R_h \mid Du_{h+1})^T, \quad h \le n-1, \end{aligned} \tag{3.5}$$

$$M_n^\sharp(a, \delta, D) \stackrel{\text{def}}{=} R_n|T_n|^{-1}R_n^T, \tag{3.6}$$

where $\delta \in \mathbb{R}$ and $D \in \mathbb{R}^{n\times n}$ is nonsingular.

Theorem 3.1 [Basic Properties] *Consider any Krylov-subspace method to solve the symmetric linear system (2.1), where A is indefinite. Suppose that Assumption 2.1 holds and the Krylov-subspace method performs $h \leq n$ iterations. Let $a \in \mathbb{R}$, $\delta \neq 0$, and let the matrix $D \in \mathbb{R}^{n\times n}$ be such that $[R_h \mid Du_{h+1} \mid DR_{n,h+1}]$ is nonsingular, where $R_{n,h+1}R_{n,h+1}^T = I_n - (R_h \mid u_{h+1})(R_h \mid u_{h+1})^T$. Then, we have the following properties:*

a) the matrix $M_h^\sharp(a,\delta,D)$ is symmetric. Furthermore,

- *when $h \leq n-1$, for any $a \in \mathbb{R} \setminus \{\pm\delta(e_h^T|T_h|^{-1}e_h)^{-1/2}\}$, $M_h^\sharp(a,\delta,D)$ is nonsingular. In addition, if $D = I_n$ then*
$$\det\left(M_h^\sharp(a,\delta,I_n)\right) = \delta^{-2h}\det(|T_h|^{-1})\left(1 - \frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h\right)^{-1};$$
- *when $h = n$ the matrix $M_h^\sharp(a,\delta,D)$ is nonsingular. In addition, if $D = I_n$ then*
$$\det\left(M_n^\sharp(a,\delta,I_n)\right) = \det(|T_h|^{-1});$$

b) setting $D = I_n$ and $\delta = 1$ the matrix $M_h^\sharp(a,1,I_n)$ coincides with M_h^{-1};

c) for $|a| < |\delta|(e_h^T|T_h|^{-1}e_h)^{-1/2}$ the matrix $M_h^\sharp(a,\delta,D)$ is positive definite. Moreover, if $D = I_n$ the spectrum $\Lambda[M_h^\sharp(a,\delta,I_n)]$ is given by
$$\Lambda[M_h^\sharp(a,\delta,I_n)] = \Lambda\left[\left(\begin{array}{c|c}\delta^2|T_h| & ae_h \\ \hline ae_h^T & 1\end{array}\right)^{-1}\right] \cup \Lambda\left[I_{n-(h+1)}\right];$$

d) when $h \leq n-1$, $D = I_n$ and either $T_h \succ 0$ or T_h is not factorizable

- *then $M_h^\sharp(a,\delta,I_n)A$ has at least $(h-3)$ singular values equal to $+1/\delta^2$;*
- *if $a = 0$ then the matrix $M_h^\sharp(0,\delta,I_n)A$ has at least $(h-2)$ singular values equal to $+1/\delta^2$;*

e) when $h = n$, then $M_n^\sharp(a,\delta,D) = M_n^{-1}$, $\Lambda[M_n] = \Lambda[|T_n|]$ and $\Lambda[M_n^{-1}A] = \Lambda[AM_n^{-1}] \subseteq \{-1,+1\}$, i.e. the n eigenvalues of the preconditioned matrix $M_h^\sharp(a,\delta,D)A$ are either $+1$ or -1.

Proof: See the Appendix. □

3.1 Further spectral properties of our proposal

Most of the relevant results in the previous section refer to the *singular values* of the unsymmetric matrix $M_h^\sharp(a,\delta,D)A$, $h \geq 1$. Here we want to generalize and strengthen item *d)* of Theorem 3.1, so that more precise indications can be given also for the *eigenvalues* of $M_h^\sharp(a,\delta,D)A$. Indeed, we recall that if $\sigma_1 \leq \cdots \leq \sigma_n$ and $|\lambda_1| \leq \cdots \leq |\lambda_n|$ are respectively the singular values and the eigenvalues of $M_h^\sharp(a,\delta,D)A$, then by Browne's theorem (see e.g. [22]) we have $\sigma_1 \leq |\lambda_1|$ and $\sigma_n \geq |\lambda_n|$. Now we can easily prove the following technical lemma.

Lemma 3.2 *Given the symmetric matrices $M, A \in \mathbb{R}^{n\times n}$, with $M = N\mathcal{L}^T\mathcal{L}N^T$ nonsingular and $N, \mathcal{L} \in \mathbb{R}^{n\times n}$, then MA has the same eigenvalues of the matrix $\mathcal{L}N^TAN\mathcal{L}^T$.*

Proof: If (v,λ) is an eigenpair of MA we simply have
$$MAv = \lambda v \iff N\mathcal{L}^T\mathcal{L}N^TAv = \lambda v \iff \mathcal{L}N^TAN\mathcal{L}^T(\mathcal{L}^{-T}N^{-1}v) = \lambda(\mathcal{L}^{-T}N^{-1}v),$$
so that λ is also an eigenvalue of the matrix $\mathcal{L}N^TAN\mathcal{L}^T$. □

The previous lemma yields a technical condition which is used in the next proposition, in order to strongly reinforce item *d)* of Theorem 3.1. In the next proposition we both give conditions in terms of *eigenvalues* of $M_h^\sharp(a,\delta,D)A$, in place of *singular values*, and give indications on the *condition number* of the preconditioned matrix $M_h^\sharp(a,\delta,D)A$.

Theorem 3.3 [Advanced Properties] *Let $\lambda_M(A)$ [$\lambda_m(A)$] be the largest [smallest] eigenvalue of the indefinite matrix A, and let $\lambda_i(M_h^\sharp(0,\delta,I_n)A)$ be the $i-th$ eigenvalue of $M_h^\sharp(0,\delta,I_n)A$. In the hypotheses of Theorem 3.1, let $D = I_n$, $\delta \neq 0$, $|a| < |\delta|(e_h^T|T_h|^{-1}e_h)^{-1/2}$ and T_h factorizable. Then, we have the following properties:*

i) for any value of $a \in \mathbb{R}$ the matrix $M_h^\sharp(a,\delta,I_n)A$ has at least $(h-2)$ eigenvalues equal to $\pm 1/\delta^2$. Moreover, if $E_m^h \in \mathbb{R}$ in (3.1) (i.e. the last diagonal block of D_h is 1×1) then $M_h^\sharp(a,\delta,I_n)A$ has at least $(h-1)$ eigenvalues equal to $\pm 1/\delta^2$;

ii) for any value of $a \in \mathbb{R}$, if D_h is diagonal then the matrix $M_h^\sharp(a,\delta,I_n)A$ has at least $(h-1)$ eigenvalues equal to $\pm 1/\delta^2$;

iii) for any value of $a \in \mathbb{R}$, if $A \succ 0$ then the matrix $M_h^\sharp(a,\delta,I_n)A$ has at least $(h-1)$ eigenvalues equal to $+1/\delta^2$;

iv) if $A \succ 0$ and $a = \delta^2\rho_{h+1}$ then the matrix $M_h^\sharp(a,\delta,I_n)A$ has at least h eigenvalues equal to $+1/\delta^2$;

v) denoting (see (3.2)) $\sigma_{\min} = \min_{1\le i\le h}\{\sigma_i\}$ and $\sigma_{\max} = \max_{1\le i\le h}\{\sigma_i\}$, if $a = 0$ then

$$\begin{cases} \lambda_i(M_h^\sharp(0,\delta,I_n)A) \ge \min\left\{\frac{\sigma_{\min}}{\delta^2}, \lambda_m(A)\right\} - O\left(\left[\rho_{h+1}^2+\rho_{h+2}^2\right]^{1/2}\right) \\ \\ \lambda_i(M_h^\sharp(0,\delta,I_n)A) \ \le \ \max\left\{\frac{\sigma_{\max}}{\delta^2}, \lambda_M(A)\right\} + O\left(\left[\rho_{h+1}^2+\rho_{h+2}^2\right]^{1/2}\right). \end{cases} \tag{3.7}$$

Proof: From (2.3) since T_h is factorizable we have $|T_h| = L_h|D_h|L_h^T$. As regards *i)*, let $h \le n-1$ and observe that setting $N = [R_h \mid u_{h+1} \mid R_{n,h+1}]$ the matrix $M_h^\sharp(a,\delta,I_n)A$ may be rewritten as

$$M_h^\sharp(a,\delta,D) = N \left[\begin{array}{c|c|c} \left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right)^{-1} & 0 \\ \hline 0 & I_{n-(h+1)} \end{array}\right] N^T, \qquad h \le n-1.$$

Recalling that L_h is unit lower bidiagonal, which implies that $L_h^{-1}e_h = e_h$ and $e_h^T L_h^T = e_h^T$, using (A.5) to obtain the factorization

$$\begin{aligned} \left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right)^{-1} \quad = \quad & \left(\begin{array}{c|c} \frac{1}{\delta}L_h^{-T}|D_h|^{-1/2} & \frac{\omega}{\delta^2}\left(1-\frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h\right)^{1/2}|T_h|^{-1}e_h \\ \hline 0 & \left(1-\frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h\right)^{-1/2} \end{array}\right) \cdot \\ & \cdot \left(\begin{array}{c|c} \frac{1}{\delta}|D_h|^{-1/2}L_h^{-1} & 0 \\ \hline \frac{\omega}{\delta^2}\left(1-\frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h\right)^{1/2} e_h^T|T_h|^{-1} & \left(1-\frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h\right)^{-1/2} \end{array}\right), \end{aligned}$$

where

$$\omega = -\frac{a}{1-\frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h},$$

we have

$$\begin{aligned} M_h^\sharp(a,\delta,I_n) \quad = \quad & N \left[\begin{array}{c|c|c} \frac{1}{\delta}L_h^{-T}|D_h|^{-1/2} & \frac{\omega}{\delta^2}\left(1-\frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h\right)^{1/2}|T_h|^{-1}e_h & 0 \\ \hline 0 & \left(1-\frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h\right)^{-1/2} & 0 \\ \hline 0 & 0 & I_{n-(h+1)} \end{array}\right] \cdot \\ & \cdot \left[\begin{array}{c|c|c} \frac{1}{\delta}|D_h|^{-1/2}L_h^{-1} & 0 & 0 \\ \hline \frac{\omega}{\delta^2}\left(1-\frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h\right)^{1/2} e_h^T|T_h|^{-1} & \left(1-\frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h\right)^{-1/2} & 0 \\ \hline 0 & 0 & I_{n-(h+1)} \end{array}\right] N^T \\ = \quad & N\bar{\mathcal{L}}_h^T\bar{\mathcal{L}}_h N^T, \end{aligned}$$

with $\bar{\mathcal{L}}_h$ being the matrix in square brackets. Moreover, by (2.2) and (A.9)

$$\begin{aligned} N^TAN \quad = \quad & \left[\begin{array}{c|c|c} R_h^TAR_h & R_h^TAu_{h+1} & R_h^TAR_{n,h+1} \\ \hline u_{h+1}^TAR_h & u_{h+1}^TAu_{h+1} & u_{h+1}^TAR_{n,h+1} \\ \hline R_{n,h+1}^TAR_h & R_{n,h+1}^TAu_{h+1} & R_{n,h+1}^TAR_{n,h+1} \end{array}\right] \\ = \quad & \left[\begin{array}{c|c|c} T_h & \rho_{h+1}e_h & 0 \\ \hline \rho_{h+1}e_h^T & t_{h+1,h+1} & \rho_{h+2}e_1^T \\ \hline 0 & \rho_{h+2}e_1 & R_{n,h+1}^TAR_{n,h+1} \end{array}\right]. \end{aligned} \tag{3.8}$$

Finally, let

$$\Delta = \left(1 - \frac{a^2}{\delta^2} e_h^T |T_h|^{-1} e_h\right)^{1/2}$$

and note that

$$|D_h|^{-1/2} L_h^{-1} T_h = |D_h|^{-1/2} L_h^{-1} L_h D_h L_h^T = |D_h|^{-1/2} D_h L_h^T$$

$$|D_h|^{-1/2} L_h^{-1} e_h = |D_h|^{-1/2} e_h$$

$$e_h^T |T_h|^{-1} T_h = e_h^T L_h^{-T} |D_h|^{-1} L_h^{-1} L_h D_h L_h^T = e_h^T L_h^{-T} \hat{I}_h L_h^T$$

$$e_h^T |T_h|^{-1} e_h = e_h^T L_h^{-T} |D_h|^{-1} L_h^{-1} e_h = e_h^T |D_h|^{-1} e_h$$

$$|T_h|^{-1} e_h = L_h^{-T} |D_h|^{-1} L_h^{-1} e_h = L_h^{-T} |D_h|^{-1} e_h$$

$$|D_h|^{-1/2} D_h |D_h|^{-1/2} = D_h |D_h|^{-1} = \hat{I}_h$$

$$|D_h|^{-1/2} D_h |D_h|^{-1} = |D_h|^{-1} D_h |D_h|^{-1/2}.$$

Then, by Lemma 3.2, λ is an eigenvalue of $M_h^\sharp(a, \delta, I_n)A$ if and only if λ is an eigenvalue of $\bar{\mathcal{L}}_h N^T A N \bar{\mathcal{L}}_h^T$, i.e. (after some computation) λ is an eigenvalue of

$$\bar{\mathcal{L}}_h N^T A N \bar{\mathcal{L}}_h^T \quad =$$

$$\left[\begin{array}{c|c|c} \frac{1}{\delta^2}\hat{I}_h & \frac{1}{\delta}\left[\frac{\omega\Delta}{\delta^2}\hat{I}_h + \frac{\rho_{h+1}}{\Delta} I_h\right] |D_h|^{-1/2} e_h & 0 \\ \hline \frac{1}{\delta} e_h^T |D_h|^{-1/2} \left[\frac{\omega\Delta}{\delta^2}\hat{I}_h + \frac{\rho_{h+1}}{\Delta} I_h\right] & \frac{\omega}{\delta^2} e_h^T \left[\frac{\omega\Delta^2}{\delta^2}\hat{I}_h |D_h|^{-1} + 2\rho_{h+1}|D_h|^{-1}\right] e_h + \frac{t_{h+1,h+1}}{\Delta^2} & \frac{\rho_{h+2}}{\Delta} e_1^T \\ \hline 0 & \frac{\rho_{h+2}}{\Delta} e_1 & R_{n,h+1}^T A R_{n,h+1} \end{array}\right] \tag{3.9}$$

Now, observe that from (3.1) we have

$$e_h^T |D_h|^{-1/2} = \begin{cases} \alpha e_h^T & \text{if } E_m^h \in \mathbb{R} \\ \\ \alpha e_{h-1}^T + \beta e_h^T & \text{if } E_m^h \in \mathbb{R}^{2\times 2}. \end{cases} \tag{3.10}$$

Thus, if $E_m^h \in \mathbb{R}$ the matrix $M_h^\sharp(a, \delta, I_n)A$ has at least $(h-1)$ eigenvalues equal to $\pm 1/\delta^2$ (corresponding to the eigenvectors $e_1, \ldots, e_{h-1}$), else if $E_m^h \in \mathbb{R}^{2\times 2}$ the matrix $M_h^\sharp(a, \delta, I_n)A$ has at least $(h-2)$ eigenvalues equal to $\pm 1/\delta^2$ (corresponding to the eigenvectors $e_1, \ldots, e_{h-2}$).

As regards *ii)*, we have that $|T_h| = L_h |D_h| L_h^T$ and since D_h is diagonal then $E_m^h \in \mathbb{R}$. Thus, *ii)* follows from *i)*.

As regards *iii)*, since $A \succ 0$ we surely have that T_h is factorizable, with $|T_h| = L_h D_h L_h^T$ and D_h diagonal. As a consequence, *ii)* yields *iii)*.

The item *iv)* is a special case of *iii)*, so that by the choice of the parameter a, the matrix in (3.9) reduces to

$$\left[\begin{array}{c|c|c} \frac{1}{\delta^2} I_h & 0 & 0 \\ \hline 0 & \frac{\omega}{\delta^2}\left[\frac{\omega\Delta^2}{\delta^2} e_h^T D_h^{-1} e_h + 2\rho_{h+1} e_h^T D_h^{-1} e_h\right] + \frac{t_{h+1,h+1}}{\Delta^2} & \frac{\rho_{h+2}}{\Delta} e_1^T \\ \hline 0 & \frac{\rho_{h+2}}{\Delta} e_1 & R_{n,h+1}^T A R_{n,h+1} \end{array}\right].$$

Thus, the matrix $M_h^\sharp(a, \delta, I_n)A$ has at least h eigenvalues equal to $+1/\delta^2$ (corresponding to the eigenvectors $e_1, \ldots, e_h$).

As regards item *v)*, by a generalization of the *monotonicity theorem (or Weyls inequality)* (see e.g. [6] Theorems 8.4.9 and 8.4.11), we have that if matrices B, C are symmetric then

$$\lambda_i(B) + \lambda_{\min}(C) \leq \lambda_i(B + C) \leq \lambda_i(B) + \lambda_{\max}(C),$$

where $\lambda_i(B)$ is the $i-$th eigenvalue of B and $\lambda_{\min}(C)$ [$\lambda_{\max}(C)$] is the smallest [largest] eigenvalue of C. Thus, in case $a = 0$ (i.e. $\Delta = 1$) the matrix $\bar{\mathcal{L}}_h N^T A N \bar{\mathcal{L}}_h^T$ in (3.9) reduces to

$$\left[\begin{array}{c|c|c} \frac{1}{\delta^2}\hat{I}_h & \frac{\rho_{h+1}}{\delta}|D_h|^{-1/2} e_h & 0 \\ \hline \frac{\rho_{h+1}}{\delta} e_h^T |D_h|^{-1/2} & t_{h+1,h+1} & \rho_{h+2} e_1^T \\ \hline 0 & \rho_{h+2} e_1 & R_{n,h+1}^T A R_{n,h+1} \end{array}\right].$$

Setting

$$B = \left[\begin{array}{c|c|c} \frac{1}{\delta^2}\hat{I}_h & 0 & 0 \\ \hline 0 & t_{h+1,h+1} & 0 \\ \hline 0 & 0 & R_{n,h+1}^T A R_{n,h+1} \end{array}\right],$$

$$C = \left[\begin{array}{c|c|c} 0 & \frac{\rho_{h+1}}{\delta}|D_h|^{-1/2}e_h & 0 \\ \hline \frac{\rho_{h+1}}{\delta}e_h^T|D_h|^{-1/2} & 0 & \rho_{h+2}e_1^T \\ \hline 0 & \rho_{h+2}e_1 & 0 \end{array}\right],$$

from *Gershgorin Circle Theorem* and from (3.10) we have

$$\lambda_{\min}(C) = -O\left(\left[\rho_{h+1}^2 + \rho_{h+2}^2\right]^{1/2}\right)$$

$$\lambda_{\max}(C) = +O\left(\left[\rho_{h+1}^2 + \rho_{h+2}^2\right]^{1/2}\right),$$

so that

$$\left\{\begin{array}{l} \lambda_i(M_h^\sharp(a,\delta,I_n)A) \;\geq\; \min\left\{\frac{\sigma_{\min}}{\delta^2}, t_{h+1,h+1}, \lambda_{\min}(R_{n,h+1}^T A R_{n,h+1})\right\} - O\left(\left[\rho_{h+1}^2 + \rho_{h+2}^2\right]^{1/2}\right) \\ \\ \lambda_i(M_h^\sharp(a,\delta,I_n)A) \;\leq\; \max\left\{\frac{\sigma_{\max}}{\delta^2}, t_{h+1,h+1}, \lambda_{\max}(R_{n,h+1}^T A R_{n,h+1})\right\} + O\left(\left[\rho_{h+1}^2 + \rho_{h+2}^2\right]^{1/2}\right). \end{array}\right.$$

By *Poincare's separation theorem* (see also Corollary 4.3.16 of [22]) the $i-$th eigenvalue $\lambda_i(R_{n,h+1}^T A R_{n,h+1})$ of the matrix $R_{n,h+1}^T A R_{n,h+1}$ satisfies

$$\lambda_m(A) \leq \lambda_i(R_{n,h+1}^T A R_{n,h+1}) \leq \lambda_M(A),$$

so that (3.7) holds. □

Observe that the results in Theorem 3.3 hold provided that T_h is factorizable. In the latter case we might have either $T_h \succ 0$ or T_h indefinite, so that the results in Theorem 3.3 are partially more general than the results in Lemma 3.3 and Theorem 3.4 of [19]. Conversely, in case $A \succ 0$ the results in item *v)* of Theorem 3.3 are weaker than those in [19].

Remark 3.1 We urge to complete this section, highlighting that the right-hand side of (3.7) may be easily simplified when $A \succ 0$. Indeed, suppose we compute $\frac{\tau}{\delta^2}Ax = \frac{\tau}{\delta^2}b$, with $\tau = \|w\|^2/w^TAw$ and $\tau \in \mathbb{R}\setminus\{0\}$, where w is *any vector* such that $w^TAw \neq 0$. Then, as showed also in [19], we obtain $\lambda_m(\frac{\tau}{\delta^2}A) \leq w^T(\frac{\tau}{\delta^2}A)w/\|w\|^2 = \frac{1}{\delta^2} \leq \lambda_M(\frac{\tau}{\delta^2}A)$. Thus, given the linear system (2.1), when $A \succ 0$ and $a = 0$ we can always multiply it for the scalar $\frac{\tau}{\delta^2}$ such that (3.7) becomes

$$\lambda_m(A) - O\left(\left[\rho_{h+1}^2 + \rho_{h+2}^2\right]^{1/2}\right) \;\leq\; \lambda_i(M_h^\sharp(0,\delta,I_n)A) \;\leq\; \lambda_M(A) + O\left(\left[\rho_{h+1}^2 + \rho_{h+2}^2\right]^{1/2}\right).$$

Recalling that when the Krylov-subspace method adopted is approaching the solution, then both $\rho_{h+1} \to 0$ and $\rho_{h+2} \to 0$, we soon realize that when $A \succ 0$, $a = 0$, $\delta = 1$, $D = I_n$ and the Krylov-subspace method approaches the solution of the linear system, then we can always obtain the bound

$$\frac{\max_i[\lambda_i(M_h^\sharp(0,1,I_n)A)]}{\min_i[\lambda_i(M_h^\sharp(0,1,I_n)A)]} \;\leq\; \frac{\lambda_M(A) + O\left(\left[\rho_{h+1}^2 + \rho_{h+2}^2\right]^{1/2}\right)}{\lambda_m(A) - O\left(\left[\rho_{h+1}^2 + \rho_{h+2}^2\right]^{1/2}\right)}, \qquad \text{when } \;\rho_{h+1}, \rho_{h+2} \to 0.$$

The latter result is weaker than the analogous result in Lemma 3.3 and Theorem 3.4 of [19]. However, we are going to show that the practical performance of our proposal is definitely comparable with that in [19]; moreover, we will show that our preconditioners require less memory storage and involve a reduced computational cost.

4 Some additional features

This section summarizes some very general properties associated with the preconditioners (3.5)-(3.6). First note that of course the matrix $R_{n,h+1}$ in the statement of Theorem 3.1 always exists, such that $[R_h \mid u_{h+1} \mid R_{n,h+1}]$

is orthogonal. However, $R_{n,h+1}$ is *neither built nor used* in (3.5)-(3.6), and it is introduced only for theoretical purposes. Furthermore, it is easy to see that since $[R_h \mid u_{h+1} \mid R_{n,h+1}]$ is orthogonal, any nonsingular diagonal matrix D may be used in order to satisfy the hypotheses of Theorem 3.1.

More precisely, we first address the issue of determining the nonsingular matrix D such that the matrix $[R_h \mid Du_{h+1} \mid DR_{n,h+1}]$ is nonsingular, considering that the columns of $[R_h \mid u_{h+1} \mid R_{n,h+1}]$ are orthogonal. To this aim, in order to maintain all the columns of $[R_h \mid Du_{h+1} \mid DR_{n,h+1}]$ similarly scaled (i.e. $\|u_i\| \approx 1$, for $i = 1, \ldots, n$), we can require the matrix D to be orthogonal. Thus, from [6] Fact 5.14.14, we know that D is orthogonal *if and only if*

$$D = \prod_{i=1}^{m} P(\theta_i, j_i, k_i), \tag{4.1}$$

with $1 \leq m \leq n(n-1)/2$, $\theta_i \in \mathbb{R}$, $j_i,\ k_i \in \{1, \ldots, n\}$, and the matrix $P(\theta_i, j_i, k_i)$ is the *Givens rotation* defined by

$$P(\theta_i, j_i, k_i) \stackrel{\text{def}}{=} I_n + [(\cos\theta) - 1]\,(E_{j,j} + E_{k,k}) + (\sin\theta)\,(E_{j,k} - E_{k,j})\,,$$

where $E_{p,q} \in \mathbb{R}^{n\times n}$ with all zero entries except 1 at position (p, q). By (4.1), setting $0 \leq |\theta_i| < \pi/2$ or $\pi/2 < |\theta_i| \leq \pi$, with $i = 1, \ldots, m$, we can potentially build any orthogonal matrix D such that $[R_h \mid Du_{h+1} \mid DR_{n,h+1}]$ is nonsingular.

Remark 4.1 Observe that the introduction of the nonsingular matrix D in (3.5) addresses a very general structure for the preconditioner $M_h^\sharp(a, \delta, D)$. As an example, setting $h = 0$ we have $M_0^\sharp(a, \delta, D) = DD^T \succ 0$, so that the preconditioner $M_0^\sharp(a, \delta, D)$ will encompass several classes of preconditioners from the literature (e.g. diagonal banded and block diagonal preconditioners [32]), even though no information is provided by the Krylov-subspace method. Indeed, in the latter case DD^T might be considered as a *first-level preconditioner* for the current linear system.
Furthermore, with the choice $D = I_n$ and $\delta = 1$ the preconditioner $M_h^\sharp(a, 1, I_n)$ can be regarded as an approximate inverse preconditioner [32], without any scaling. Finally, though the choice $\delta = 1$ in (3.5) seems the most obvious, numerical reasons related to formula (3.9) and to the condition number of $M_h^\sharp(a, \delta, D)A$ may suggest other values for the parameter 'δ' (see also Remark 4.3).

On the other hand, in order to let the statement of Theorem 3.1 unchanged, it is worth to highlight that the projector $\mathcal{P} = I_n - (R_h|u_{h+1})(R_h|u_{h+1})^T$ in (3.5) might in principle be replaced by the more general expression $\mathcal{P} = I_n - (R_h|u_{h+1})\Omega(R_h|u_{h+1})^T$, with $\Omega \in \mathbb{R}^{(h+1)\times(h+1)}$ positive definite. However, the latter new expression for $\mathcal{P}$ (such that $\mathcal{P}$ is no more a projector) requires to determine the matrix $V_{n,h+1}$, which satisfies the conditions

$$V_{n,h+1}V_{n,h+1}^T = I_n - (R_h \mid u_{h+1})\Omega(R_h \mid u_{h+1})^T$$

$$(R_h \mid u_{h+1})^T V_{n,h+1} = 0,$$

and allows the identity matrix $I_{n-(h+1)}$ in (A.3) in the Appendix to be replaced by a positive definite matrix. As a consequence, in this paper we set $\Omega = I_{h+1}$.

Proposition 4.1 [Invariance - Scalability] *Suppose (2.2) holds, with the matrix A possibly indefinite. Let $P \in \mathbb{R}^{h\times h}$, with P orthogonal. Then, the preconditioners $M_h^\sharp(0, \delta, D)$ are invariant under the transformation $R_h = Q_h P$, $Q_h = (q_1 \cdots q_h)$, $q_i^T q_j = 0$ and $\|q_i\| = 1$, for $1 \leq i \neq j \leq h$, and $Q_h^T u_{h+1} = 0$. Moreover, considering the scaled system $(\varepsilon A)x = (\varepsilon b)$ then*

$$\begin{aligned} M_h^\sharp(a, \delta, D) &= D\left[I_n - (R_h \mid u_{h+1})\,(R_h \mid u_{h+1})^T\right] D^T \\ &\quad + (R_h \mid Du_{h+1}) \left(\begin{array}{c|c} \delta^2\varepsilon|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right)^{-1} (R_h \mid Du_{h+1})^T \qquad h \leq n-1, \\ M_n^\sharp(a, \delta, D) &= \frac{1}{\varepsilon} R_n |T_n|^{-1} R_n^T, \qquad h = n. \end{aligned} \tag{4.2}$$

Proof: From (2.2) and condition $R_h = Q_h P$ we have

$$T_h = P^T Q_h^T A Q_h P = P^T \tilde{T}_h P,$$

where $\tilde{T}_h$ is possibly not tridiagonal. Moreover, we can easily prove that

$$|P^T \tilde{T}_h P| = P^T |\tilde{T}_h| P,$$

so that, setting $R_h = Q_h P$, we have for $h \leq n-1$

$$\begin{aligned}
M_h^\sharp(0,\delta,D) &= D\left[I_n - (R_h \mid u_{h+1})(R_h \mid u_{h+1})^T\right] D^T \\
&\quad + (R_h \mid Du_{h+1}) \left(\begin{array}{c|c} \delta^2 |P^T Q_h^T A Q_h P| & 0 \cdot e_h \\ \hline 0 \cdot e_h^T & 1 \end{array}\right)^{-1} (R_h \mid Du_{h+1})^T \\
&= D\left[I_n - (Q_h \mid u_{h+1})(Q_h \mid u_{h+1})^T\right] D^T + \frac{1}{\delta^2} Q_h P |P^T Q_h^T A Q_h P|^{-1} P^T Q_h^T + Du_{h+1}u_{h+1}^T D^T \\
&= D\left[I_n - Q_h Q_h^T\right] D^T + \frac{1}{\delta^2} Q_h P P^T |Q_h^T A Q_h|^{-1} P P^T Q_h^T \\
&= D\left[I_n - (Q_h \mid u_{h+1})(Q_h \mid u_{h+1})^T\right] D^T + (Q_h \mid Du_{h+1}) \left(\begin{array}{c|c} \delta^2 |\tilde{T}_h| & 0 \\ \hline 0 & 1 \end{array}\right)^{-1} (Q_h \mid Du_{h+1})^T,
\end{aligned}$$

which coincides with (3.5), setting $a = 0$, replacing R_h with Q_h and considering that $|T|^{-1}$ is likely dense, regardless of the sparsity of T. The previous result holds also for $h = n$, after a trivial computation.

Furthermore, observe that the matrix R_h in (3.5)-(3.6) is invariant under the scale factor ε in $(\varepsilon A)x = (\varepsilon b)$, and replacing A with εA we have $T_h(\varepsilon) = \varepsilon R_h^T A R_h = \varepsilon T_h$. Thus, (4.2) trivially holds. □

Broadly speaking, as for *LMP* preconditioners (see [27] and Section 5 for a further comparison), the preconditioners (3.5)-(3.6) cannot be independent of the scale parameter ε. Indeed, as we can soon realize, when $h = n$ and $A \succ 0$ the matrix $M_h^\sharp(a,\delta,D)$ is the inverse of the system matrix εA, so that

$$M_n^\sharp(a,\delta,D) \cdot (\varepsilon A) = \left[\frac{1}{\varepsilon} R_n T_n^{-1} R_n^T\right] \left[\varepsilon R_n T_n R_n^T\right] = I_n.$$

In Section 5 we are going to consider again the case $a = 0$, which yields a strong relation between our proposal and *LMP* preconditioners.

Remark 4.2 Recalling Remark 2.2, several strategies for building our preconditioners, by selecting ℓ vectors among $\{u_1, \ldots, u_m\}$, with $\ell \leq m$, can be adopted (see also [25]). However, the reader is warned that depending on the resulting strategy adopted, the properties in Theorem 3.1 should be suitably restated. On this guideline, now we want to analyze the strategies corresponding to choose either the *first* ℓ vectors $\{u_1, \ldots, u_\ell\}$, or the *last* $m - \ell$ vectors $\{u_{\ell+1}, \ldots, u_m\}$. To this purpose, considering (2.1) suppose a Krylov-subspace method was adopted to generate the recurrence

$$AR_m = R_m T_m + \rho_{m+1} u_{m+1} e_m^T. \tag{4.3}$$

From Assumption 2.1 the matrix T_m is *tridiagonal*. Thus, since $R_m = (R_\ell \mid R_{m,\ell+1})$, where $R_{m,\ell+1} = (u_{\ell+1} \mid \cdots \mid u_m)$, setting for the tridiagonal matrix T_m the decomposition

$$T_m = \left(\begin{array}{c|c} T_\ell & \frac{0}{\sigma e_1^T} \\ \hline \frac{\sigma e_\ell^T}{0} & T_{m,\ell+1} \end{array}\right), \qquad \text{for some } \sigma \in \mathbb{R},$$

from (4.3) we have

$$\begin{aligned}
AR_m &= A(R_\ell \mid R_{m,\ell+1}) = (R_\ell \mid R_{m,\ell+1}) T_m + \rho_{m+1} u_{m+1} e_m^T \\
&= \left(R_\ell T_\ell + \sigma u_{\ell+1} e_\ell^T \mid \sigma u_\ell e_1^T + R_{m,\ell+1} T_{m,\ell+1}\right) + \rho_{m+1} u_{m+1} e_m^T,
\end{aligned}$$

which is equivalent to the following pair of conditions

$$AR_\ell = R_\ell T_\ell + \sigma u_{\ell+1} e_\ell^T \tag{4.4}$$

$$AR_{m,\ell+1} = R_{m,\ell+1} T_{m,\ell+1} + \sigma u_\ell e_1^T + \rho_{m+1} u_{m+1} e_m^T. \tag{4.5}$$

Thus, if only the first ℓ vectors $u_1, \ldots, u_\ell$ are used to build our preconditioners, then relation (4.4) must be adopted in place of (2.2). On the other hand, if the last $m - \ell$ vectors $u_{\ell+1}, \ldots, u_m$ are used, relation (2.2) must

be replaced by (4.5). However, in the latter case the statements of Theorem 3.1 and Theorem 3.3 should be slightly modified, accordingly.

Finally, other possible strategies to select vectors among $\{u_1, \ldots, u_m\}$ can be considered, which may require a more consistent reformulation of the statements of Theorem 3.1 and Theorem 3.3.

It is possible to show that trying to introduce a slightly more general structure of $M_h^\sharp(a, \delta, D)$, where the parameter 'δ' is replaced by a scaling (diagonal) matrix $\Delta \in \mathbb{R}^{h\times h}$ (used to *balance* the rows of matrix $|T_h|$), the item d) of Theorem 3.1 may not be fulfilled. The next result summarizes the properties of our class of preconditioners, for a very simple and opportunistic choice of the parameters 'a', 'δ' and matrix 'D'.

Corollary 4.2 [Standard Case] *Consider any Krylov-subspace method to solve the symmetric linear system (2.1). Suppose that Assumption 2.1 holds and the Krylov-subspace method performs $h \leq n$ iterations. Then, setting $a = 0$, $\delta = 1$ and $D = I_n$ in Theorem 3.1 the preconditioner*

$$M_h^\sharp(0,1,I_n) = \left[I_n - (R_h \mid u_{h+1})\,(R_h \mid u_{h+1})^T\right] + (R_h \mid u_{h+1}) \left(\begin{array}{c|c} |T_h| & 0 \\ \hline 0 & 1 \end{array}\right)^{-1} (R_h \mid u_{h+1})^T \tag{4.6}$$

$$M_n^\sharp(0,1,I_n) = R_n |T_n|^{-1} R_n^T, \tag{4.7}$$

is such that

a) the matrix $M_h^\sharp(0,1,I_n)$ is symmetric and nonsingular for any $h \leq n$;

b) the matrix $M_h^\sharp(0,1,I_n)$ coincides with M_h^{-1} (where M_h is defined in (3.3)-(3.4)), for any $h \leq n$;

c) the matrix $M_h^\sharp(0,1,I_n)$ is positive definite. Moreover, its spectrum $\Lambda[M_h^\sharp(0,1,I_n)]$ is given by

$$\Lambda[M_h^\sharp(0,1,I_n)] = \Lambda\left[|T_h|^{-1}\right] \cup \Lambda\left[I_{n-h}\right];$$

d) when $h \leq n-1$ the matrix $M_h^\sharp(0,1,I_n)A$ has at least $(h-1)$ eigenvalues in the set $\{-1,+1\}$;

e) when $h = n$ then $\Lambda[M_n] = \Lambda[|T_n|]$ and $\Lambda[M_n^\sharp(0,1,I_n)A] = \Lambda[M_n^{-1}A] = \Lambda[AM_n^{-1}] \subseteq \{-1,+1\}$, i.e. the n eigenvalues of $M_h^\sharp(0,1,I_n)A$ are either $+1$ or -1.

Proof: The result is directly obtained from (3.3)-(3.4), Theorem 3.1 and Theorem 3.3, setting $a = 0$, $\delta = 1$ and $D = I_n$. □

Remark 4.3 The choice of the parameters 'δ' and 'a', and the matrix 'D' is problem dependent. In particular, 'δ' and 'a' may be set in order to impose conditions like the following (which tend to force the clustering of the eigenvalues of matrix $H_{(h+1)\times(h+1)}$ or $H_{h\times h}$ in (A.12)-(A.13) in the Appendix, near $+1$ or near -1):

$$\begin{array}{ll} \det\left[H_{(h+1)\times(h+1)}\right] = 1, & tr\left[H_{(h+1)\times(h+1)}\right] = h+1, \\ \det\left[H_{h\times h}\right] = 1, & tr\left[H_{h\times h}\right] = h. \end{array}$$

Finally, observe that depending on the quantities in the expressions (A.13)-(A.14), there may be real values of the parameters 'δ' and 'a' such that $\eta = 0$. Choosing the latter values for 'δ' and 'a' may reinforce the conclusions of item d) in Theorem 3.1.

4.1 Computational cost

First, observe that the case $h \approx n$ in Theorem 3.1 and Corollary 4.2 is of scarce interest for large scale problems. Indeed, in the literature of preconditioners for large scale problems the values of 'h' typically do not exceed $10 \div 20$ (see e.g. [25, 26]). Moreover, for small values of h (say $h \ll n$), in (3.5) the computation of the inverse matrix

$$\left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right)^{-1}, \tag{4.8}$$

may be cheaply performed both when T_h is factorizable or not factorizable. Indeed, recalling that T_h is nonsingular, relation (2.3) and relation (A.10) in the Appendix will provide the result. Thus, the overall cost (flops) for computing (4.8) is mostly due to the computational burden of $|T_h|^{-1}$. However, with a better insight and considering that our preconditioners are suited for large scale problems, observe that the application of our proposal only requires to compute the matrix (4.8) times a $(h+1)$-real vector. Indeed, the recursions of Krylov-subspace methods never use directly matrices. Thus, the computational core of computing the matrix (4.8) times a vector is the product $|T_h|^{-1}u$, where $u \in \mathbb{R}^h$. In this regard, we have the following characterization:

- if T_h is **factorizable** then $|T_h|^{-1}u = (L_h|D_h|L_h^T)^{-1}u = L_h^{-T}|D_h|^{-1}L_h^{-T}u$. Considering the results in Section 4 of [12], we have that the cost $\mathcal{C}(|T_h|^{-1}u)$ of computing the product $|T_h|^{-1}u$, is given by $\mathcal{C}(|T_h|^{-1}u) = O(h^2)$;
- if T_h is **not factorizable** (which is possibly expected to be a rare case, even when A is indefinite) then $|T_h|^{-1}u = (V_h|B_h|V_h^T)^{-1}u = V_h|B_h|^{-1}V_h^T u$, and recalling that B_h is diagonal, the cost $\mathcal{C}(|T_h|^{-1}u)$ of calculating the product $|T_h|^{-1}u$ (not including the cost to compute the Jordan Canonical form of T_h, which amounts to $O(h^3)$), is given by $\mathcal{C}(|T_h|^{-1}u) = O(h^2)$.

To sum up, setting $D = I_n$ (i.e. without introducing any scaling), from (3.5) the overall cost for (only) computing the matrix-vector product $M_h^\sharp(a, \delta, I_n)w$, for $w \in \mathbb{R}^n$, is given by

$$(h+1)n + n(h+1) + (h+1)^2 = 2(h+1)n + (h+1)^2, \tag{4.9}$$

regardless of the fact that A is possibly indefinite or positive definite. The quantity (4.9) is rather competitive with both *LMP* [19] and *PREQN* [25], considering that $h \ll n$.

Furthermore, with respect to *LMP*, our proposal requires to store only the $(h+1)$ vectors $u_1, \ldots, u_{h+1}$, in place of $2h$ vectors (in addition, when $A \succ 0$, *LMP* requires also to compute the latter $2h$ vectors with a procedure that needs h matrix-vector products, at an additional cost of $\approx 3h^2n$ flops [19]).

5 Relation between our class of preconditioners and *LMP*

Limited Memory Preconditioners (*LMP*) are an interesting class of preconditioners recently introduced in [19], and they also encompass the preconditioner *PREQN* in [25]. They use L-BFGS quasi-Newton updates to collect information, in order to build an approximate inverse preconditioner for (2.1), when A is *positive definite*. In principle, since they adopt L-BFGS in place of Krylov-subspace methods, the idea behind their approach follows a different guideline with respect to our proposal. However, recalling the analogy of the search directions computed by L-BFGS and the conjugate directions computed by the CG, when A is positive definite, it is not surprising that the two approaches show similarities.

In particular, assuming $A \succ 0$ (the latter assumption holds along the whole section) the expression of *LMP* preconditioners is given by

$$\mathcal{M}_h = \left[I_n - S_h(S_h^T A S_h)^{-1} S_h^T A\right] \mathcal{M}_0 \left[I_n - A S_h(S_h^T A S_h)^{-1} S_h^T\right] + S_h(S_h^T A S_h)^{-1} S_h^T, \tag{5.1}$$

where $S_h \in \mathbb{R}^{n\times h}$ is provided by L-BFGS and $\mathcal{M}_0 \succ 0$.
Now, assuming that L-BFGS and CG are performed for h steps starting from the same initial point, due to the above analogy between L-BFGS and CG, we can set (in exact arithmetics)

$$S_h = (s_1 \cdots s_h), \qquad\qquad s_i^T A s_j = 0, \quad 1 \le i \ne j \le h,$$

where $s_1, \ldots, s_h$ are the conjugate directions computed by the CG.

Intuitively speaking, in (3.5) the matrix DD^T plays the role of scaling matrix, similar to that of $\mathcal{M}_0$ in (5.1), since $[I_n - (R_h \mid u_{h+1})(R_h \mid u_{h+1})^T]$ is a projector and we can write in (3.5)

$$\begin{aligned} D\left[I_n - (R_h \mid u_{h+1})(R_h \mid u_{h+1})^T\right] D^T \quad = \\ D\left[I_n - (R_h \mid u_{h+1})(R_h \mid u_{h+1})^T\right]\left[I_n - (R_h \mid u_{h+1})(R_h \mid u_{h+1})^T\right] D^T, \end{aligned}$$

so that when $A \succ 0$ the quantity $S_h(S_h^T A S_h)^{-1} S_h^T$ in (5.1) resembles the second term in (3.5), after setting $\delta = 1$ and $a = 0$.

Observe that in (5.1) $\mathcal{M}_h$ is invariant under any scaling of the vectors $s_1, \ldots, s_h$. Thus, if $r_1, \ldots, r_h$ are the corresponding residuals computed by the CG up to step h, we can equivalently set in (5.1)

$$S_h = \left(\frac{s_1}{\|r_1\|} \cdots \frac{s_h}{\|r_h\|}\right).$$

Moreover, thanks to the relation between the residuals and the conjugate directions when applying the CG (see

also [34]), we have $S_h = R_h L_h^{-T}$, where

$$R_h = \left(\frac{r_1}{\|r_1\|} \cdots \frac{r_h}{\|r_h\|} \right) \qquad \text{and} \qquad L_h = \begin{pmatrix} 1 & & & & \\ -\frac{\|r_2\|}{\|r_1\|} & 1 & & & \\ & -\frac{\|r_3\|}{\|r_2\|} & 1 & & \\ & & \ddots & \ddots & \\ & & & -\frac{\|r_h\|}{\|r_{h-1}\|} & 1 \end{pmatrix}. \tag{5.2}$$

Thus, recalling Assumption 2.1, relation (5.1) becomes

$$\begin{aligned} \mathcal{M}_h &= \left[I_n - R_h L_h^{-T}(L_h^{-1}R_h^T A R_h L_h^{-T})^{-1} L_h^{-1} R_h^T A\right] \mathcal{M}_0 \left[I_n - A R_h L_h^{-T}(L_h^{-1}R_h^T A R_h L_h^{-T})^{-1} L_h^{-1} R_h^T\right] \\ &\quad + R_h L_h^{-T}(L_h^{-1}R_h^T A R_h L_h^{-T})^{-1} L_h^{-1} R_h^T \\ &= \left[I_n - R_h (R_h^T A R_h)^{-1} R_h^T A\right] \mathcal{M}_0 \left[I_n - A R_h (R_h^T A R_h)^{-1} R_h^T\right] + R_h (R_h^T A R_h)^{-1} R_h^T \\ &= \left[I_n - R_h T_h^{-1} \left(R_h T_h + \rho_{h+1} \frac{r_{h+1}}{\|r_{h+1}\|} e_h^T \right)^T\right] \mathcal{M}_0 \left[I_n - \left(R_h T_h + \rho_{h+1} \frac{r_{h+1}}{\|r_{h+1}\|} e_h^T \right) T_h^{-1} R_h^T\right] \\ &\quad + R_h T_h^{-1} R_h^T \\ &= \left[I_n - R_h R_h^T - \rho_{h+1} R_h T_h^{-1} e_h \frac{r_{h+1}^T}{\|r_{h+1}\|}\right] \mathcal{M}_0 \left[I_n - R_h R_h^T - \rho_{h+1} \frac{r_{h+1}}{\|r_{h+1}\|} e_h^T T_h^{-1} R_h^T\right] \\ &\quad + R_h T_h^{-1} R_h^T. \end{aligned} \tag{5.3}$$

By the latter relation, and using relation (3.5) with $\delta = 1$ and $a = 0$, we deduce that $\mathcal{M}_h$ and $M_h^\sharp(a, \delta, I_n)$ may be in general strongly different. However, recalling that $(I_n - R_h R_h^T)$ is idempotent, setting $\rho_{h+1} = 0$ and $\mathcal{M}_0 = I_n$, we obtain

$$\mathcal{M}_h \equiv M_h^\sharp(0, 1, I_n).$$

Thus, if $A \succ 0$, when the L-BFGS update stops (i.e. when correspondingly $\rho_{h+1} = 0$ applying the CG), *LMP* with $\mathcal{M}_0 = I_n$ and our proposal, with $\delta = 1$, $a = 0$, $D = I_n$, in exact arithmetics coincide. Moreover, when $\rho_{h+1} \neq 0$, there is no chance that the two preconditioners may coincide.

5.1 An explicit example

Now we want to be more specific and to explicitly compare the preconditioners $\mathcal{M}_h$ in (5.1) and $M_h^\sharp(0, 1, I_n)$, when $h \ll n$ and the CG is used to provide the conjugate directions $s_1, \ldots, s_h$, along with the sequences of scalars $\{\alpha_1, \ldots, \alpha_h\}$ and $\{\beta_1, \ldots, \beta_h\}$. In particular, suppose the CG has performed h steps *without stopping*, i.e. $\beta_h \neq 0$, with

$$r_{i+1} = r_i - \alpha_i A s_i, \qquad \alpha_i = \frac{\|r_i\|^2}{s_i^T A s_i}, \qquad i = 1, \ldots, h,$$

$$s_{i+1} = r_{i+1} + \beta_i s_i, \qquad \beta_i = \frac{\|r_{i+1}\|^2}{\|r_i\|^2}, \qquad i = 1, \ldots, h.$$

Then, relation (2.2) becomes more explicitly (see also [34])

$$A R_h = R_h T_h - \frac{\sqrt{\beta_h}}{\alpha_h} \frac{r_{h+1}}{\|r_{h+1}\|} e_h^T,$$

where $T_h = L_h D_h L_h^T$, $D_h = diag_{i=1,\ldots,h}\{1/\alpha_i\}$ and L_h is given in (5.2). Thus, since L_h^{-1} is unit lower triangular, then

$$R_h T_h^{-1} e_h = R_h L_h^{-T} D_h^{-1} L_h^{-1} e_h = R_h L_h^{-T} D_h^{-1} e_h = \alpha_h R_h L_h^{-T} e_h. \tag{5.4}$$

Moreover, by a simple check we have that

$$L_h^{-1} = \begin{pmatrix} 1 & & & & & \\ \sqrt{\beta_1} & 1 & & & & \\ \sqrt{\beta_1\beta_2} & \sqrt{\beta_2} & 1 & & & \\ \vdots & \vdots & \ddots & \ddots & & \\ \prod_{i=1}^{h-2}\sqrt{\beta_i} & \prod_{i=2}^{h-2}\sqrt{\beta_i} & \cdots & \sqrt{\beta_{h-2}} & 1 & \\ \prod_{i=1}^{h-1}\sqrt{\beta_i} & \prod_{i=2}^{h-1}\sqrt{\beta_i} & \cdots & \cdots & \sqrt{\beta_{h-1}} & 1 \end{pmatrix},$$

so that from (5.4)

$$R_h T_h^{-1} e_h = \alpha_h R_h \begin{pmatrix} \prod_{i=1}^{h-1}\sqrt{\beta_i} \\ \prod_{i=2}^{h-1}\sqrt{\beta_i} \\ \vdots \\ \sqrt{\beta_{h-1}} \\ 1 \end{pmatrix} = \alpha_h \|r_h\| R_h \begin{pmatrix} 1/\|r_1\| \\ 1/\|r_2\| \\ \vdots \\ \vdots \\ 1/\|r_{h-1}\| \\ 1/\|r_h\| \end{pmatrix} = \alpha_h \|r_h\| \sum_{i=1}^{h} \frac{r_i}{\|r_i\|^2}.$$

Finally, by (5.3) with $\mathcal{M}_0 = I_n$ we obtain

$$\begin{aligned} \mathcal{M}_h &= (I_n - R_h R_h^T)(I_n - R_h R_h^T) + \frac{\sqrt{\beta_h}}{\alpha_h}\left[R_h T_h^{-1} e_h \frac{r_{h+1}^T}{\|r_{h+1}\|} + \frac{r_{h+1}}{\|r_{h+1}\|} e_h^T T_h^{-1} R_h^T\right] + \\ &\quad \frac{\beta_h}{\alpha_h^2}(R_h T_h^{-1} e_h)(R_h T_h^{-1} e_h)^T + R_h T_h^{-1} R_h^T \\ &= (I_n - R_h R_h^T) + R_h T_h^{-1} R_h^T + \frac{\sqrt{\beta_h}\alpha_h\|r_h\|}{\alpha_h\|r_{h+1}\|}\left[\left(\sum_{i=1}^{h}\frac{r_i}{\|r_i\|^2}\right) r_{h+1}^T + r_{h+1}\left(\sum_{i=1}^{h}\frac{r_i}{\|r_i\|^2}\right)^T\right] + \\ &\quad \frac{\beta_h \alpha_h^2 \|r_h\|^2}{\alpha_h^2}\left(\sum_{i=1}^{h}\frac{r_i}{\|r_i\|^2}\right)\left(\sum_{i=1}^{h}\frac{r_i}{\|r_i\|^2}\right)^T \\ &= M_h^\sharp(0,1,I_n) + \left[\left(\sum_{i=1}^{h}\frac{r_i}{\|r_i\|^2}\right) r_{h+1}^T + r_{h+1}\left(\sum_{i=1}^{h}\frac{r_i}{\|r_i\|^2}\right)^T\right] + \\ &\quad \|r_{h+1}\|^2\left(\sum_{i=1}^{h}\frac{r_i}{\|r_i\|^2}\right)\left(\sum_{i=1}^{h}\frac{r_i}{\|r_i\|^2}\right)^T, \end{aligned} \tag{5.5}$$

which explicitly shows that setting $\mathcal{M}_0 = I_n$ and

$$r_{sum} = \sum_{i=1}^{h}\frac{r_i}{\|r_i\|^2}$$

then $\mathcal{M}_h$ and $M_h^\sharp(0,1,I_n)$ differ for the symmetric rank-2 update

$$(r_{sum} \mid r_{h+1}) \begin{pmatrix} \|r_{h+1}\|^2 & 1 \\ 1 & 0 \end{pmatrix} (r_{sum} \mid r_{h+1})^T,$$

when the CG is used to construct both of them.
Of course, as long as $\rho_{h+1} \neq 0$ (i.e. $r_{h+1} \neq 0$) or we choose other settings for the parameters $\mathcal{M}_0$, a, δ and D, the two proposals $\mathcal{M}_h$ and $M_h^\sharp(a,\delta,D)$ may further differ. In any case it may be difficult to perform an exhaustive comparison, which is independent of the choice of the parameters.

6 Further issues on the condition number of $M_h^\sharp(a,\delta,D)A$

In this section we want to estimate the condition number $\kappa(M_h^\sharp(a,\delta,D)A)$ of the unsymmetric matrix $M_h^\sharp(a,\delta,D)A$ (where $M_h^\sharp(a,\delta,D)$ is computed as in (3.5)-(3.6)). We immediately have the general formula

$$\begin{aligned}\kappa(M_h^\sharp(a,\delta,D)A) &\stackrel{\text{def}}{=} \|M_h^\sharp(a,\delta,D)A\|\cdot\|(M_h^\sharp(a,\delta,D)A)^{-1}\| \\ &= \|M_h^\sharp(a,\delta,D)A\|\cdot\|A^{-1}(M_h^\sharp(a,\delta,D))^{-1}\|. \end{aligned} \tag{6.1}$$

Observe that (6.1) provides a more precise result with respect to (3.7) and formula (3.10) in [19]. Indeed (see also [22]), if $\lambda_1,\ldots,\lambda_n$ are the eigenvalues of $M_h^\sharp(a,\delta,I_n)A$, we have in general

$$\frac{\max_i|\lambda_i|}{\min_i|\lambda_i|} \le \kappa(M_h^\sharp(a,\delta,I_n)A).$$

Moreover, in this section we consider the case where A is indefinite, which is more general than the condition which yields (3.7). Now we prove the next technical lemma.

Lemma 6.1 *Let $C\in\mathbb{R}^{h\times h}$ be a symmetric and positive definite matrix. Let $0<\omega_1\le\cdots\le\omega_h$ be the ordered eigenvalues of C, with $\omega_1,\ldots,\omega_h$ not all coincident, and let $a\in\mathbb{R}$, $\delta\in\mathbb{R}$. Then, given the quantities*

$$\alpha = -\delta^2(h-1)\omega_1+\delta^2 tr(C)+1,$$

$$\beta = \frac{\delta^2 det(C)\left[1-\dfrac{a^2}{\delta^2}e_h^TC^{-1}e_h\right]}{(\omega_h)^{h-1}},$$

we have

$$\alpha^2-4\beta>0,$$

and

$$\frac{[tr(C)-(h-1)\omega_1]\,\omega_h^{h-1}}{det(C)}>1.$$

Proof: By the definition of α and β, and since $C\succ 0$, the condition $\alpha^2-4\beta\ge 0$ is satisfied if and only if

$$\delta^2(e_h^TC^{-1}e_h)^{-1}\left[1-\frac{\alpha^2(\omega_h)^{h-1}}{4\delta^2 det(C)}\right]\le\ a^2. \tag{6.2}$$

Now, observing that $\omega_1,\ldots,\omega_h$ are not all coincident, $\alpha>\delta^2\omega_h+1$ and for any $\omega_1\ge 0$ we have $(\delta^2\omega_1+1)^2\ge 4\delta^2\omega_1$, we obtain

$$\frac{\alpha^2(\omega_h)^{h-1}}{4\delta^2 det(C)}\ \ge\ \frac{\alpha^2}{4\delta^2\omega_1}\ >\ \frac{(\delta^2\omega_h+1)^2}{4\delta^2\omega_1}\ \ge\ \frac{(\delta^2\omega_1+1)^2}{4\delta^2\omega_1}\ \ge\ 1, \tag{6.3}$$

so that (6.2) holds for any choice of a, which also implies that $\alpha^2-4\beta\ge 0$. Also observe that by (6.3) $\alpha^2(\omega_h)^{h-1}/[4\delta^2 det(C)]>1$, so that (6.2) can never be satisfied as an equality, i.e. $\alpha^2-4\beta\ne 0$ for any value of the parameter a.

Finally, note that since $det(C)=\Pi_{i=1}^h\omega_i$ we have

$$\omega_h^{h-1}>\frac{det(C)}{tr(C)-(h-1)\omega_1}, \tag{6.4}$$

inasmuch as $\omega_1,\ldots,\omega_h$ are not all coincident, and

$$\frac{det(C)}{tr(C)-(h-1)\omega_1}\le\frac{det(C)}{\omega_h}=\prod_{i=1}^{h-1}\omega_i<\omega_h^{h-1}.$$

As a consequence, we have the condition

$$\frac{[tr(C)-(h-1)\omega_1]\,\omega_h^{h-1}}{det(C)}>1.$$

□

In the following result we provide an estimation of the condition number $\kappa(M_h^\sharp(a,\delta,D)A)$ in (6.1), which depends on the parameters 'δ' and 'a', and the matrix 'D' in (3.5). Note that for the sake of clarity (but with a little abuse of notation), in the sequel we directly indicate with $\mu_1,\dots,\mu_h$ the eigenvalues of $|T_h|$ and not the eigenvalues of T_h.

Proposition 6.2 [Condition Number] *Consider the matrix $M_h^\sharp(a,\delta,D)$ in (3.5)-(3.6), with $h \le n-1$, where $|T_h|$ satisfies Assumption 2.1. Let $\mu_1 \le \cdots \le \mu_h$ be the (ordered) eigenvalues of $|T_h|$, where $\mu_1,\dots,\mu_h$ are not all coincident. Then, if*

$$|a| \ < |\delta|(e_h^T|T_h|^{-1}e_h)^{-1/2}, \qquad \delta \neq 0 \tag{6.5}$$

we have

$$\kappa\left(M_h^\sharp(a,\delta,D)A\right) \ \le \ \xi_h \cdot \kappa(N)^2 \cdot \kappa(A), \tag{6.6}$$

with

$$\xi_h = \frac{\max\left\{1, \frac{\gamma_h+(\gamma_h^2-4\sigma_h)^{1/2}}{2}\right\}}{\min\left\{1, \frac{\gamma_h-(\gamma_h^2-4\sigma_h)^{1/2}}{2}\right\}} \ge 1, \tag{6.7}$$

$$N = [R_h \mid Du_{h+1} \mid DR_{n,h+1}]$$

and

$$\gamma_h = -\delta^2(h-1)\mu_1 + \delta^2 tr(|T_h|) + 1$$

$$\sigma_h = \frac{\delta^2 det(|T_h|)\left[1 - \frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h\right]}{(\mu_h)^{h-1}}.$$

In particular, when $D = I_n$ in (3.5), then $\kappa(M_h^\sharp(a,\delta,I_n)A) \le \xi_h \cdot \kappa(A)$.

Proof: Let $\lambda_1 \le \cdots \le \lambda_{h+1}$ be the (ordered) eigenvalues of the matrix

$$\left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right), \tag{6.8}$$

which is positive definite as long as condition (6.5) is fulfilled. Observe that by the identity

$$\left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right) = \left(\begin{array}{c|c} I_h & 0 \\ \hline \frac{a}{\delta^2}e_h^T|T_h|^{-1} & 1 \end{array}\right)\left(\begin{array}{c|c} \delta^2|T_h| & 0 \\ \hline 0 & 1-\frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h \end{array}\right)\left(\begin{array}{c|c} I_h & \frac{a}{\delta^2}|T_h|^{-1}e_h \\ \hline 0 & 1 \end{array}\right)$$

we have

$$det\left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right) \ = \ \delta^{2h} det(|T_h|)\left[1 - \frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h\right] \tag{6.9}$$

and $\delta^2|T_h|$ is the $h \times h$ upper left diagonal block of matrix (6.8). Therefore, by the Cauchy interlacing properties (Lemma 8.4.4 in [6]) between the sequences $\{\mu_j\}_{j=1,\dots,h}$ and $\{\lambda_i\}_{i=1,\dots,h+1}$ we have the relation

$$\lambda_1 \le \delta^2\mu_1 \le \lambda_2 \le \delta^2\mu_2 \le \cdots \le \lambda_h \le \delta^2\mu_h \le \lambda_{h+1}. \tag{6.10}$$

By (6.8), (6.9) and (6.10) we can immediately infer the following intermediate results:

i) $\delta^2\mu_1 \le \lambda_i \le \delta^2\mu_h, \qquad i = 2,\dots,h$

ii) $\sum_{i=1}^{h+1}\lambda_i = \delta^2 tr(|T_h|) + 1,$

iii) $\prod_{i=1}^{h+1}\lambda_i = \delta^{2h} det(|T_h|)\left[1 - \frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h\right].$

From i) we deduce that

$$\delta^2(h-1)\mu_1 \;\le\; \sum_{i=2}^{h} \lambda_i \;\le \delta^2(h-1)\mu_h,$$

so that from ii), iii), (6.10) and recalling that the matrix (6.8) is positive definite, we have

$$\max\left\{0, -\delta^2(h-1)\mu_h + \delta^2 tr(|T_h|) + 1\right\} \;\le\; \lambda_1 + \lambda_{h+1} \;\le\; -\delta^2(h-1)\mu_1 + \delta^2 tr(|T_h|) + 1$$

and

$$\frac{\delta^{2h} det(|T_h|)\left[1 - \frac{a^2}{\delta^2} e_h^T |T_h|^{-1} e_h\right]}{\delta^{2(h-1)}(\mu_h)^{h-1}} \;\le\; \lambda_1 \cdot \lambda_{h+1} \;\le\; \frac{\delta^{2h} det(|T_h|)\left[1 - \frac{a^2}{\delta^2} e_h^T |T_h|^{-1} e_h\right]}{\delta^{2(h-1)}(\mu_1)^{h-1}}. \tag{6.11}$$

From (6.11) (see also points (A) and (B) in Figure 6.1), in order to compute a lower [upper] *bound* $\tilde{\lambda}_1$ [$\tilde{\lambda}_{h+1}$]

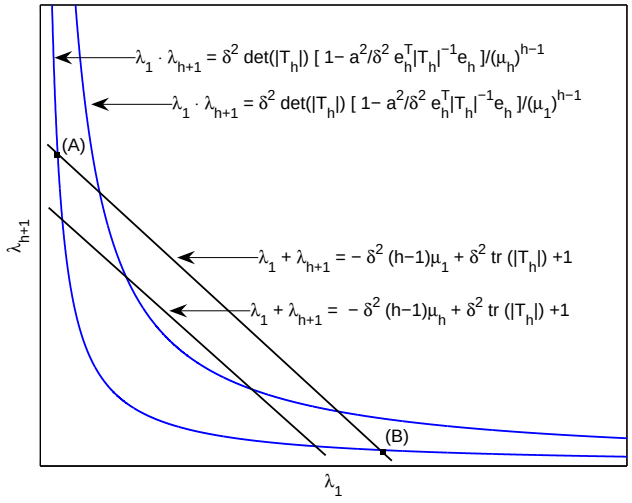

Figure 6.1: Relation between the eigenvalues λ_1 and λ_{h+1} of matrix (6.8).

for the smallest [largest] eigenvalue of matrix (6.8), we have to solve the linear system (σ_h and γ_h are defined in the statement of this proposition)

$$\begin{cases} \tilde{\lambda}_1 + \tilde{\lambda}_{h+1} = \gamma_h \\ \tilde{\lambda}_1 \cdot \tilde{\lambda}_{h+1} = \sigma_h, \end{cases}$$

which yields

$$\begin{aligned} \tilde{\lambda}_1 &= \frac{\gamma_h - (\gamma_h^2 - 4\sigma_h)^{1/2}}{2} \\ \tilde{\lambda}_{h+1} &= \frac{\gamma_h + (\gamma_h^2 - 4\sigma_h)^{1/2}}{2}, \end{aligned} \tag{6.12}$$

provided that $\gamma_h^2 - 4\sigma_h \ge 0$. However, the latter condition directly holds from Lemma 6.1. Now, observe that setting $N = [R_h \mid Du_{h+1} \mid DR_{n,h+1}]$ (where N is nonsingular by hypothesis), for $h \le n-1$ the preconditioners $M_h^\sharp(a,\delta,D)$ may be rewritten as

$$M_h^\sharp(a,\delta,D) = N \left[\begin{array}{cc|c} \delta^2|T_h| & ae_h & 0 \\ ae_h^T & 1 & \\ \hline 0 & & I_{n-(h+1)} \end{array}\right]^{-1} N^T, \qquad h \le n-1. \tag{6.13}$$

As a consequence, setting

$$W_h = \left[\begin{array}{cc|c} \delta^2|T_h| & ae_h & 0 \\ ae_h^T & 1 & \\ \hline 0 & & I_{n-(h+1)} \end{array}\right],$$

we have for the smallest [largest] eigenvalue of the symmetric matrices W_h and W_h^{-1} the expressions

$$\begin{cases} \lambda_m(W_h) = \min\{1, \lambda_1\} \\ \\ \lambda_M(W_h) = \max\{1, \lambda_{h+1}\} \end{cases}$$

$$\begin{cases} \lambda_m(W_h^{-1}) = \frac{1}{\max\{1,\lambda_{h+1}\}} \\ \\ \lambda_M(W_h^{-1}) = \frac{1}{\min\{1,\lambda_1\}}. \end{cases}$$

Thus, if $\lambda_m(A)$ $[\lambda_m(A^{-1})]$ and $\lambda_M(A)$ $[\lambda_M(A^{-1})]$ are the smallest [largest] eigenvalues of matrix A $[A^{-1}]$, from (6.13) we have

$$\|M_h^\sharp(a,\delta,D)A\| \;\le\; \lambda_M(A)\cdot\|N\|^2\cdot\lambda_M(W_h^{-1}) \;=\; \lambda_M(A)\cdot\|N\|^2\cdot\frac{1}{\min\{1,\lambda_1\}}$$

and

$$\begin{aligned} \|(M_h^\sharp(a,\delta,D)A)^{-1}\| &= \|A^{-1}(M_h^\sharp(a,\delta,D))^{-1}\| \;\le\; \lambda_M(A^{-1})\cdot\|N^{-1}\|^2\cdot\lambda_M(W_h) \\ &= \frac{1}{\lambda_m(A)}\cdot\|N^{-1}\|^2\cdot\max\{1,\lambda_{h+1}\}, \end{aligned}$$

so that from (6.12)

$$\kappa\left(M_h^\sharp(a,\delta,D)A\right) \;=\; \|M_h^\sharp(a,\delta,D)A\|\cdot\|(M_h^\sharp(a,\delta,D)A)^{-1}\| \;\le\; \frac{\max\left\{1,\tilde\lambda_{h+1}\right\}}{\min\left\{1,\tilde\lambda_1\right\}}\kappa(N)^2\kappa(A),$$

which is relation (6.6). Finally, when $D = I_n$ in (3.5) then $\kappa(N) = 1$. □

6.1 On the assessment of the bound (6.6)

In order to better specify the bound (6.6) we can now prove the next proposition.

Proposition 6.3 *Let us consider the hypotheses of Proposition 6.2 and the quantity ξ_h defined in (6.7). Then, for any choice of 'δ' and 'a' satisfying (6.5) we have*

$$\xi_h = \frac{\gamma_h + (\gamma_h^2 - 4\sigma_h)^{1/2}}{\gamma_h - (\gamma_h^2 - 4\sigma_h)^{1/2}}. \tag{6.14}$$

Proof: The proof consists to analyze the following three cases:

1) $\gamma_h < 2$ (i.e. $\delta^2 < 1/[tr(|T_h|) - (h-1)\mu_1]$)

2) $\gamma_h = 2$ (i.e. $\delta^2 = 1/[tr(|T_h|) - (h-1)\mu_1]$)

3) $\gamma_h > 2$ (i.e. $\delta^2 > 1/[tr(|T_h|) - (h-1)\mu_1]$).

In case 1) is satisfied, observe that the inequality

$$\frac{\gamma_h + (\gamma_h^2 - 4\sigma_h)^{1/2}}{2} < 1$$

cannot hold, since (consider that $\gamma_h - 2 < 0$ and see Lemma 6.1) it requires that

$$\gamma_h < 1 + \sigma_h \qquad \text{if and only if} \qquad a^2 < \left[1 - \frac{(\gamma_h - 1)\mu_h^{h-1}}{\delta^2 det(|T_h|)}\right]\frac{\delta^2}{e_h^T|T_h|^{-1}e_h}$$

which can hold only if

$$\frac{(\gamma_h - 1)\mu_h^{h-1}}{\delta^2 det(|T_h|)} \le 1$$

or equivalently

$$\delta^2 \geq \frac{(\gamma_h - 1)\mu_h^{h-1}}{det(|T_h|)}.$$

However, the last inequality cannot hold because it is equivalent to

$$1 \geq \frac{[tr(|T_h|) - (h-1)\mu_1]\mu_h^{h-1}}{det(|T_h|)},$$

which cannot be satisfied from Lemma 6.1. Moreover, in case 1), also

$$\frac{\gamma_h - (\gamma_h^2 - 4\sigma_h)^{1/2}}{2} > 1$$

cannot hold, since $\gamma_h - 2 < 0$. Therefore, when $\gamma_h < 2$ relation (6.14) holds.

The case 2) is pretty similar to the case 1), so that again (6.14) follows almost immediately.

In case 3), the inequality

$$\frac{\gamma_h + (\gamma_h^2 - 4\sigma_h)^{1/2}}{2} < 1$$

cannot hold since it is equivalent to $(\gamma_h^2 - 4\sigma_h)^{1/2} < 2 - \gamma_h < 0$. Moreover, from Lemma 6.1 and considering that $\gamma_h - 2 > 0$, the condition

$$\frac{\gamma_h - (\gamma_h^2 - 4\sigma_h)^{1/2}}{2} > 1$$

can be satisfied if

$$\gamma_h < 1 + \sigma_h \quad \text{if and only if} \quad a^2 < \left[1 - \frac{(\gamma_h - 1)\mu_h^{h-1}}{\delta^2 det(|T_h|)}\right] \frac{\delta^2}{e_h^T |T_h|^{-1} e_h},$$

which holds only if

$$\frac{(\gamma_h - 1)\mu_h^{h-1}}{\delta^2 det(|T_h|)} \leq 1$$

or equivalently

$$\delta^2 \geq \frac{(\gamma_h - 1)\mu_h^{h-1}}{det(|T_h|)}.$$

However, since $\gamma_h - 1 = tr(|T_h|) - (h-1)\mu_1$, the last inequality is again equivalent to

$$1 \geq \frac{[tr(|T_h|) - (h-1)\mu_1]\mu_h^{h-1}}{det(|T_h|)}$$

which cannot hold from Lemma 6.1. Thus relation (6.14) holds.

□

Lemma 6.4 *Consider the matrix $M_h^\sharp(a, \delta, D)$ in (3.5)-(3.6), with $h \leq n-1$. Let $\mu_1 \leq \cdots \leq \mu_h$ be the (ordered) eigenvalues of $|T_h|$, with $\mu_1, \cdots, \mu_h$ not all coincident, and let the parameters 'a' and 'δ' satisfy condition (6.5). Then, for any choice of the nonsingular matrix D in (3.5)*

- *the coefficient ξ_h in (6.14) increases when $|a| \to \rho$, with $\rho = |\delta|(e_h^T |T_h|^{-1} e_h)^{-1/2}$, and*

$$\lim_{|a| \uparrow \rho} \xi_h = +\infty$$

- *the coefficient ξ_h in (6.14) attains its minimum when $a = 0$, and for $a = 0$ we have*

$$\xi_h = \frac{\gamma_h + \left(\gamma_h^2 - 4\frac{\delta^2 det(|T_h|)}{(\mu_h)^{h-1}}\right)^{1/2}}{\gamma_h - \left(\gamma_h^2 - 4\frac{\delta^2 det(|T_h|)}{(\mu_h)^{h-1}}\right)^{1/2}}. \tag{6.15}$$

Proof: Observe that $\lim_{|a|\uparrow\rho} \xi_h = +\infty$. Indeed, when $|a| \to \rho$ we have $\sigma_h \to 0$, so that $\gamma_h - (\gamma_h^2 - 4\sigma_h)^{1/2} \to 0$ and $\gamma_h + (\gamma_h^2 - 4\sigma_h)^{1/2} \to 2\gamma_h$, with $\gamma_h > 1$. Thus, since from Lemma 6.1 $\gamma_h - 4\sigma_h \geq 0$, Proposition 6.3 ensures that ξ_h satisfies (6.14), so that ξ_h increases as $|a| \to \rho$, with $\lim_{|a|\uparrow\rho} \xi_h = +\infty$. Moreover, from (6.14) and since ξ_h is a continuous function of the parameter 'a' (see also (6.5)), we have

$$\frac{\partial \xi_h}{\partial a} = \frac{\partial \xi_h}{\partial \sigma_h} \cdot \frac{\partial \sigma_h}{\partial a} = \frac{-4\gamma_h}{[\gamma_h - (\gamma_h^2 - 4\sigma_h)^{1/2}]^2 (\gamma_h^2 - 4\sigma_h)^{1/2}} \cdot \frac{-2a \cdot det(|T_h|) e_h^T |T_h|^{-1} e_h}{(\mu_h)^{h-1}},$$

so that from (6.5) we have $sgn\{\partial \xi_h / \partial a\} = sgn\{a\}$, which implies that ξ_h attains its minimum for $a = 0$.

Finally, by Lemma 6.1 we obtain $\gamma_h^2 - 4\sigma_h \geq 0$ for any choice of a satisfying (6.5), and when $a = 0$ it results $\sigma_h = \delta^2 det(|T_h|)/(\mu_h)^{h-1}$. Thus, from Proposition 6.3 the value of ξ_h when $a = 0$ is given by (6.15). □

Remark 6.1 By (6.15) we observe that as expected, the parameter 'δ' affects both the distribution of the singular values of $M_h^\sharp(a, \delta, D)A$ (see item d) of Theorem 3.1), and also its condition number $\kappa(M_h^\sharp(a, \delta, D)A)$, when computed according with (6.1).

7 Numerical experiments

In order to preliminarily test our proposal in a general framework, where no information is known about the sparsity pattern of the matrix A, we used our class of preconditioners $M_h^\sharp(a, \delta, D)$, setting $a = 0$, $\delta = 1$ and $D = I_n$. As expected, in our numerical experience we obtain results which match the theory in Theorem 3.1 and Theorem 3.3.

In this section, in order to test the class of preconditioners (3.5)-(3.6), we used different sets of test problems. Moreover, the results are reported also for *LMP* preconditioner $\mathcal{M}_h$, which is built as in (5.3) and (5.5) (so that the conjugacy among the directions $s_1, \ldots, s_h$ is fully exploited), setting $\mathcal{M}_0 = I_n$. Of course, the latter choice implies that after h steps of the iterative method, we will use $h + 1$ vectors to construct $\mathcal{M}_h$ and only h vectors to build $M_h^\sharp(0, 1, I_n)$. Thus, $\mathcal{M}_h$ *collects* more information than $M_h^\sharp(0, 1, I_n)$. In addition, since we build the *LMP* preconditioner directly using conjugate directions in place of L-BFGS steps (as described in [19]), we expect the latter fact will play in favor of its accuracy. Nevertheless, we show that, to a large extent, the performance of our proposal is comparable with that of *LMP* and sometimes superior.

Finally, in our numerical experience we solved all the linear systems without following the guidelines in Remark 3.1. The choice to step aside from Remark 3.1 is motivated by a specific fact. Indeed, for opportunistic values of the vector w in Remark 3.1 we risk to introduce an undesired bias in favor of one of the preconditioners in the comparisons.

7.1 Test set 1

First, we considered a set of indefinite linear systems as in (2.1), where the number of unknowns n is set as $n = 1000$, and the matrix A has also a moderate condition number. We simply wanted to experience how our class of preconditioners modifies the condition number of A. In particular (see also [14]), a possible choice for the latter class of matrices is given by

$$A = \{a_{i,j}\}, \qquad a_{i,j} \in U[-10, 10], \quad i, j = 1, \ldots, n, \tag{7.1}$$

where $a_{i,j} = a_{j,i}$ are random entries in the uniform distribution $U[-10, 10]$, between -10 and $+10$. Then, also the vector b in (2.1) is computed randomly with entries in $U[-10, 10]$. We computed the preconditioners $M_h^\sharp(0, 1, I_n)$ in (3.5) and $\mathcal{M}_h$ in (5.5) by using the CG method, which is one of the most popular Krylov-subspace methods to solve (2.1) [17]. We remark that the CG is often used also in case the matrix A is indefinite, though it can untimely stop (which did not happen on our experiment). As an alternative choice, in order to satisfy Assumption 2.1 with A indefinite, we can use the Lanczos process (see also SYMMLQ and MINRES in [31]) or Planar-CG methods [11]. In (3.5) and (5.5) we set the parameter h in the range (see also [19])

$$h \in \{\ 20\ ,\ 40\ ,\ 80\ ,\ 160\ ,\ 320\ ,\ 640\ \}.$$

We plotted in Figure 7.1 the condition number $Cond(A)$ of the matrix A, along with the condition number $Cond(AINVk \cdot A)$ of $M_h^\sharp(0, 1, I_n)A$ and $Cond(LMP \cdot A)$ of $\mathcal{M}_h A$ (we recall that the trick in Remark 3.1 was

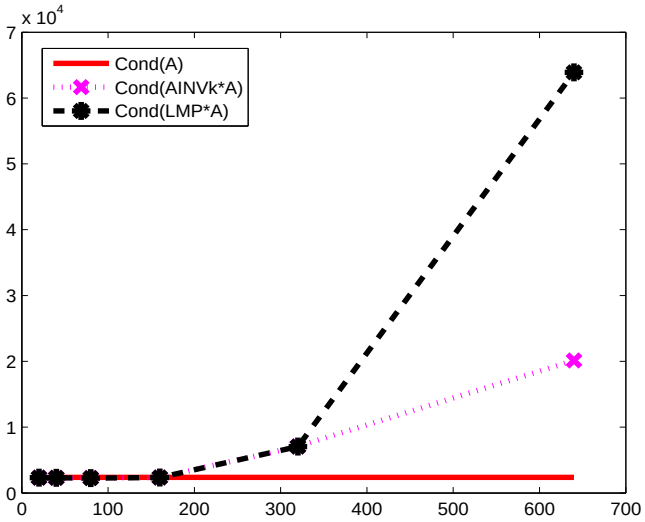

Figure 7.1: The condition number of matrix A ($Cond(A)$) along with the condition number of matrix $M_h^\sharp(0,1,I_n)A$ ($Cond(AINVk \cdot A)$) and matrix $\mathcal{M}_h A$ ($Cond(LMP \cdot A)$), when $h \in \{20, 40, 80, 160, 320, 640\}$, and $A \in \mathbb{R}^{1000\times 1000}$ is randomly chosen with entries in the uniform distribution $U[-10, 10]$. The condition number remains of the same order of magnitude (even though the trick in Remark 3.1 was not applied) and great similarities are observed between $M_h^\sharp(0,1,I_n)A$ and $\mathcal{M}_h A$.

not applied): an estimation of the condition number κ is calculated by preliminarily computing the eigenvalues $\lambda_1, \ldots, \lambda_n$ (using Matlab [24] routine `eigs()`), then determining the ratio

$$\kappa = \frac{\max_i \ |\lambda_i|}{\min_i \ |\lambda_i|}. \tag{7.2}$$

Thus, considering that $n = 1000$, and due to round-off and finite precision of computation, the quantity κ satisfies (3.7) sufficiently well. Evidently, numerical results confirm that the *order of magnitude* of the condition number of A is pretty similar to that of the condition number of $M_h^\sharp(0,1,I_n)A$ and $\mathcal{M}_h A$. This indicates that if the preconditioners (3.5) and (5.5) are used as a tool to solve (2.1), then most preconditioned iterative methods which are sensible to the condition number (e.g. the Krylov-subspace methods), on average are not expected to perform worse with respect to the unpreconditioned case. In addition, it is important to remark that the spectra $\Lambda[M_h^\sharp(0,1,I_n)A]$ and $\Lambda[\mathcal{M}_h A]$ tend to be shifted with respect to $\Lambda[A]$, inasmuch as the eigenvalues in $\Lambda[A]$ whose absolute value is larger than $+1$ tend to be scaled in $\Lambda[M_h^\sharp(0,1,I_n)A]$ and $\Lambda[\mathcal{M}_h A]$ (see Figure 7.2). The latter property is an appealing result, since the eigenvalues of $M_h^\sharp(0,1,I_n)A$ and $\mathcal{M}_h A$ will be 'more clustered' as in Theorem 3.3. Finally, the matrix A in (7.1) was generated several times yielding pictures similar to Figures 7.1–7.2.

7.2 Test set 2

In a second experiment we generated the indefinite matrix A in (2.1) such that

$$A = H\mathcal{D}H, \tag{7.3}$$

where $H \in \mathbb{R}^{n\times n}$, $n = 1000$, is an Householder transformation given by $H = I_n - 2vv^T$, with $v \in \mathbb{R}^n$ a unit vector, randomly chosen. The matrix $\mathcal{D} \in \mathbb{R}^{n\times n}$ is diagonal (so that its non-zero entries are also eigenvalues of A, while each column of H is also an eigenvector of A) and its entries are randomly chosen in the uniform distribution $U[-30, 30]$. The matrix $\mathcal{D}$ is such that its $\texttt{perc} \cdot n$ eigenvalues are larger (about one order of magnitude) than the remaining $(1 - \texttt{perc}) \cdot n$ eigenvalues (we set without loss of generality $\texttt{perc} = 0.3$). Finally, again we computed the preconditioners $M_h^\sharp(0,1,I_n)$ in (3.5) and $\mathcal{M}_h$ in (5.5) by using the CG, setting the starting point x_0 so that the initial residual $b - Ax_0$ was a linear combination (with coefficients -1 and $+1$ randomly chosen) of *all* the n eigenvectors of A. We strongly highlight that the latter choice of x_0 is expected

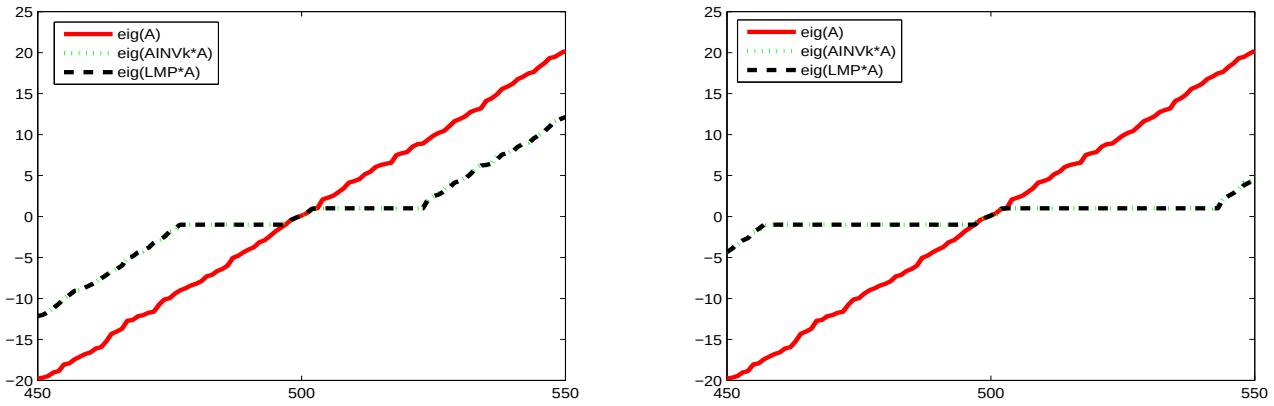

Figure 7.2: Comparison among the spectra (detail from the 450-th to the 550-th eigenvalue) $\Lambda[A]$ (eig(A)), $\Lambda[M_h^\sharp(0,1,I_n)A]$ (eig$(AINVk\cdot A)$) and $\Lambda[\mathcal{M}_h A]$ (eig$(LMP\cdot A)$), with A randomly chosen (eigenvalues of A are approximately in the range $[-300,300]$ and are sorted for simplicity. Each entry of A is a random variable with a uniform distribution in $[-10,10]$). Without loss of generality we show the results for the values $h=40$ (*left*) and $h=80$ (*right*). Apart from some exceptions on the extreme eigenvalues of the preconditioned matrix (which arise both for AINV$\mathcal{K}$ and *LMP*), the intermediate eigenvalues in the spectra $\Lambda[M_h^\sharp(0,1,I_n)A]$ and $\Lambda[\mathcal{M}_h A]$, whose absolute value is larger than $+1$, are in general smaller than the corresponding eigenvalues in $\Lambda[A]$. The eigenvalues in $\Lambda[M_h^\sharp(0,1,I_n)A]$ and $\Lambda[\mathcal{M}_h A]$ have similar values (almost overlapped) and are more clustered near ± 1 than those in $\Lambda[A]$. The results of Theorem 3.3 are confirmed.

to be not favorable when applying the CG, to build the preconditioners. In the latter case the CG method is indeed expected to perform exactly n iterations before stopping (see also [29, 32]), so that the matrix (7.3) may be significant to test the effectiveness of our preconditioners, in case of *small values* of h (broadly speaking, here h *small* implies that the preconditioners contain correspondingly a little information on the inverse matrix A^{-1}). We compared the spectra $\Lambda[A]$, $\Lambda[M_h^\sharp(a,1,I_n)A]$ and $\Lambda[\mathcal{M}_h A]$, in order to verify again how the preconditioners are able to *cluster the eigenvalues of* A. Following the choice in [25], in order to test our proposal also on a different range of values for the parameter h, we set

$$h \in \{\ 4\ ,\ 8\ ,\ 12\ ,\ 16\ ,\ 20\ ,\ 40\ \}.$$

The results are given in Figure 7.3 (condition numbers) and Figure 7.4 which includes all the 1000 eigenvalues (*left*) and a detail of the eigenvalues from the 840-th to the 875-th (*right*). Again, the Matlab routine `eigs()` is used. Observe that both the preconditioners are able to shift some eigenvalues of A towards -1 or $+1$, so that the clustering of the eigenvalues is enhanced when the parameter h increases. The spectra $\Lambda[M_h^\sharp(0,1,I_n)A]$ and $\Lambda[\mathcal{M}_h A]$ are almost overlapped.

7.3 Test set 3

We used another small test set, obtained by considering a couple of linear systems as (2.1), described in [25] and references therein, which come up from finite element problems. We addressed the latter linear systems as $A_0 x = b_0$ (*from one-dimensional model, consisting of a line of two-node elements with support conditions at both ends, and a linearly varying body force*) and $A_1 x = b_1$ (where A_1 is the *stiffness matrix from a two-dimensional finite element model of a cantilever beam*) respectively. The spectral properties of both the matrices A_0 and A_1 are extensively described in [25]. In particular $A_0 \in \mathbb{R}^{50\times 50}$ is positive definite with condition number $\kappa(A_0) = 0.2\cdot 10^{10}$ and with a suitable pattern of the eigenvalues; similarly, $A_1 \in \mathbb{R}^{170\times 170}$ is also positive

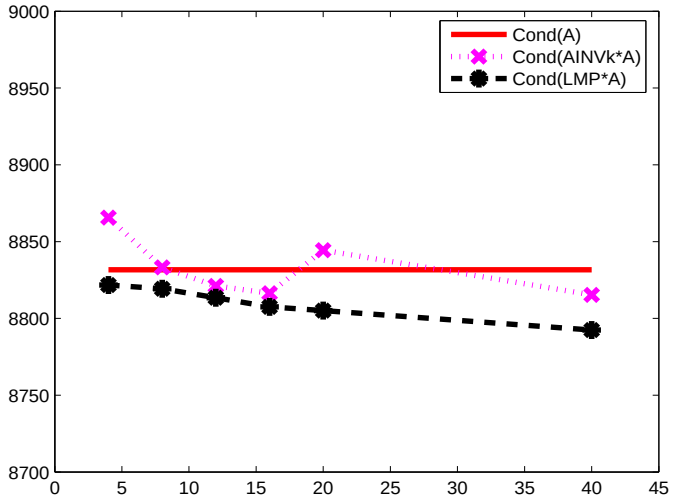

Figure 7.3: The condition number of matrix A ($Cond(A)$) along with the condition number of matrix $M_h^\sharp(0,1,I_n)A$ ($Cond(AINVk \cdot A)$) and matrix $\mathcal{M}_h A$ ($Cond(LMP \cdot A)$), when $h \in \{4, 8, 12, 16, 20, 40\}$. To a large extent, considering round-off and finite precision of the Matlab routine `eigs()` when computing the eigenvalues, and recalling that the trick in Remark 3.1 was not applied, the condition number (computed as in (7.2)) of $M_h^\sharp(0,1,I_n)A$ and $\mathcal{M}_h A$ remains of the same order of magnitude of A.

definite, with condition number $\kappa(A_1) = 0.13 \cdot 10^9$ and a different pattern of eigenvalues. In addition, we have

$$b_0 = \begin{pmatrix} 0 \\ 200/49 \\ 300/49 \\ \vdots \\ 4900/49 \\ 0 \end{pmatrix},$$

and

$$b_1 = 0, \quad \text{but} \quad b_1(34) = b_1(68) = b_1(102) = b_1(136) = b_1(170) = -8000,$$

and the CG is again used to compute both the preconditioners $M_h^\sharp(0,1,I_{50})$ and $\mathcal{M}_h$, adopting both the starting points $x_0 = 0$ and $x_0 = 100e$, $e = (1 \cdots 1)^T$, as indicated in [25].

As regards the comparison between $M_h^\sharp(0,1,I_{50})$ and $\mathcal{M}_h$ on the linear system $A_0 x = b_0$ (the relative picture is not reported for the sake of brevity), we found that $\mathcal{M}_h$ shows a slightly better performance with respect to the condition number of the preconditioned matrix (both using $x_0 = 0$ and $x_0 = 100e$). The latter result is likely due to the fact that since n is relatively small, using one additional vector to build $\mathcal{M}_h$ (with respect to $M_h^\sharp(0,1,I_{50})$) may yield a better performance. Then, in Figure 7.5 we show that the spectra $\Lambda[M_h^\sharp(0,1,I_{50})A_0]$ and $\Lambda[\mathcal{M}_h A_0]$, setting $x_0 = 0$, are definitely very similar and reveal the clustering around the value $+1$. A similar conclusion also holds considering Figure 7.6, where we set $x_0 = 100e$.

As regards the comparison between $M_h^\sharp(0,1,I_{170})$ and $\mathcal{M}_h$ on the linear system $A_1 x = b_1$, we first recall that here $n = 170$. The results of the comparison are much similar to those for the linear system $A_0 x = b_0$, and are summarized in Figure 7.7 (setting $x_0 = 0$ for the CG) and Figure 7.8 (setting $x_0 = 100e$ for the CG), respectively.

7.4 Test set 4

Finally, we tested both $M_h^\sharp(0,1,I_n)$ and $\mathcal{M}_h$ on the solution of a *sparse* linear system from real applications, which was suggested in [19]. We considered the symmetric positive definite matrix BCSSTK27, denoted hereafter by $\mathcal{A} \in \mathbb{R}^{1224\times 1224}$, from dynamic analyses in structural engineering, provided by the Harwell-Boeing Sparse

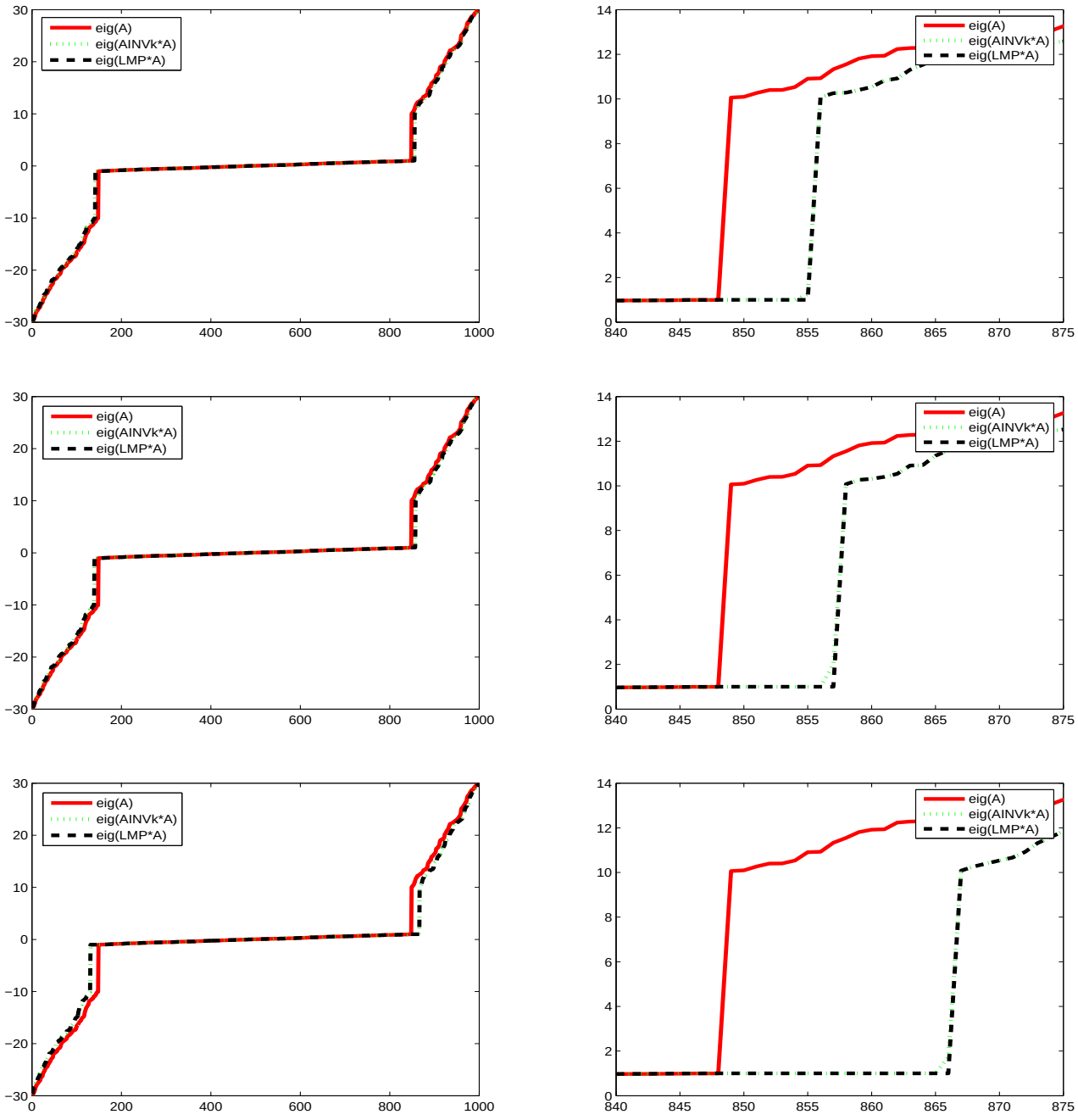

Figure 7.4: Comparison between the full (*left*) and detailed (*right*) spectra $\Lambda[A]$ (eig(A)), $\Lambda[M_h^\sharp(0,1,I_n)A]$ (eig$(AINVk \cdot A)$) and $\Lambda[\mathcal{M}_h A]$ (eig$(LMP \cdot A)$), with A nonsingular and given by (7.3) (eigenvalues are sorted for simplicity); we used $h = 16$ (*top*), $h = 20$ (*middle*) and $h = 40$ (*bottom*). A 'flatter' piecewise-line of the eigenvalues in $\Lambda[M_h^\sharp(0,1,I_n)A]$ and $\Lambda[\mathcal{M}_h A]$ indicates that the eigenvalues tend to cluster around -1 and $+1$, according to Theorem 3.3. The sequences of eigenvalues of $M_h^\sharp(0,1,I_n)A$ and $\mathcal{M}_h A$ are almost overlapped.

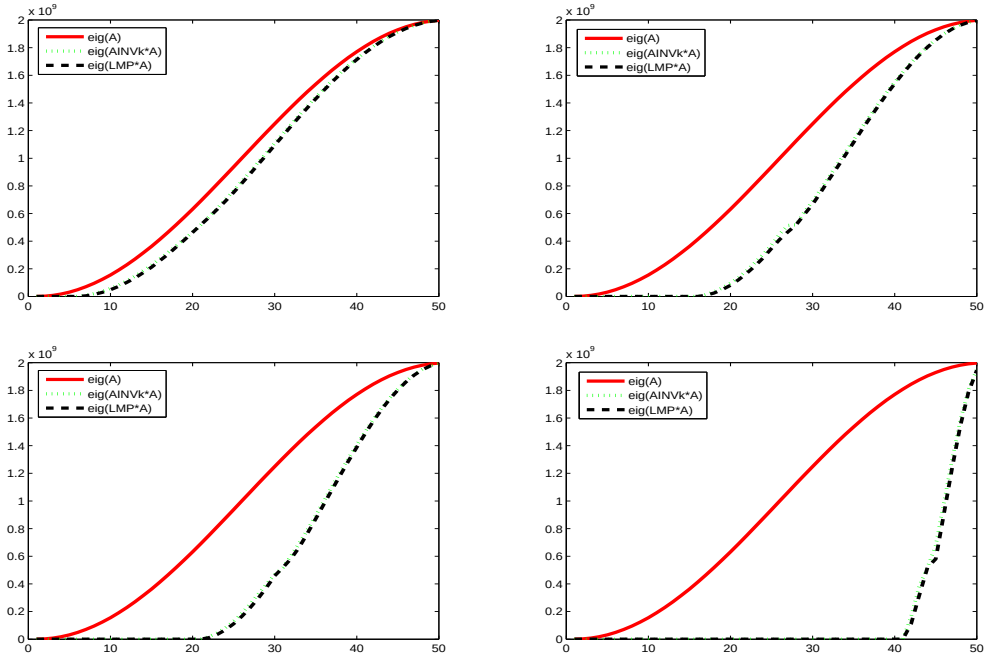

Figure 7.5: Comparison among the spectra $\Lambda[A_0]$ (eig(A)), $\Lambda[M_h^\sharp(0,1,I_{50})A_0]$ (eig$(AINVk \cdot A)$) and $\Lambda[\mathcal{M}_h A_0]$ (eig$(LMP \cdot A)$) — eigenvalues are sorted for simplicity; we set $x_0 = 0$ for the CG and we used $h = 5$ (*top left*), $h = 15$ (*top right*), $h = 20$ (*bottom left*) and $h = 40$ (*bottom right*). Several eigenvalues of $M_h^\sharp(0,1,I_{50})A_0$ and $\mathcal{M}_h A_0$ tend to cluster around $+1$, according to Theorem 3.3. The sequences of eigenvalues of $M_h^\sharp(0,1,I_{50})A_0$ and $\mathcal{M}_h A_0$ are almost overlapped.

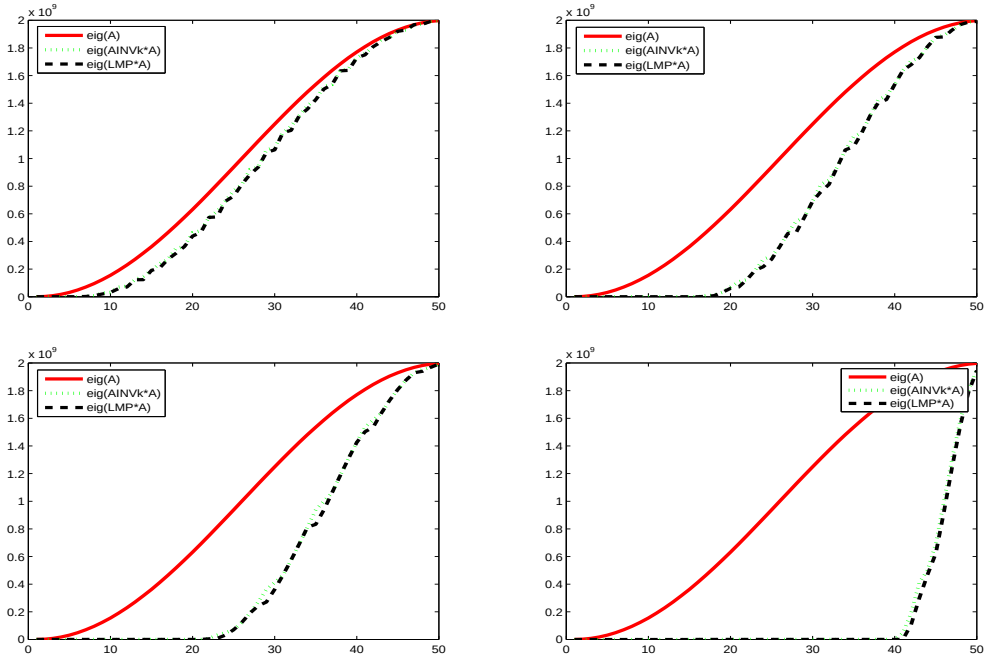

Figure 7.6: Comparison among the spectra $\Lambda[A_0]$ (eig(A)), $\Lambda[M_h^\sharp(0,1,I_{50})A_0]$ (eig$(AINVk \cdot A)$) and $\Lambda[\mathcal{M}_h A_0]$ (eig$(LMP \cdot A)$) — eigenvalues are sorted for simplicity; we set $x_0 = 100e$ for the CG and we used $h = 5$ (*top left*), $h = 15$ (*top right*), $h = 20$ (*bottom left*) and $h = 40$ (*bottom right*). We obtain results pretty similar to Figure 7.5.

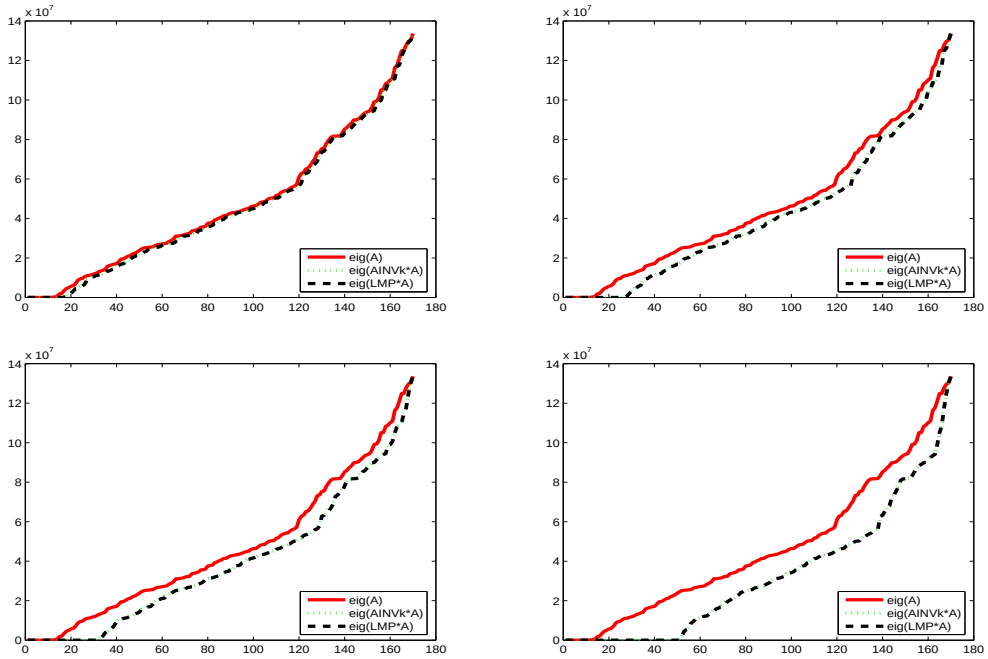

Figure 7.7: Comparison among the spectra $\Lambda[A_1]$ (eig(A)), $\Lambda[M_h^\sharp(0,1,I_{170})A_1]$ (eig$(AINVk \cdot A)$) and $\Lambda[\mathcal{M}_h A_1]$ (eig$(LMP \cdot A)$) — eigenvalues are sorted for simplicity; we set $x_0 = 0$ for the CG and we used $h = 5$ (*top left*), $h = 15$ (*top right*), $h = 20$ (*bottom left*) and $h = 40$ (*bottom right*). Several eigenvalues of $M_h^\sharp(0,1,I_{170})A_1$ and $\mathcal{M}_h A_1$ tend to cluster around $+1$, according to Theorem 3.3. The sequences of eigenvalues of $M_h^\sharp(0,1,I_{170})A_1$ and $\mathcal{M}_h A_1$ are almost overlapped.

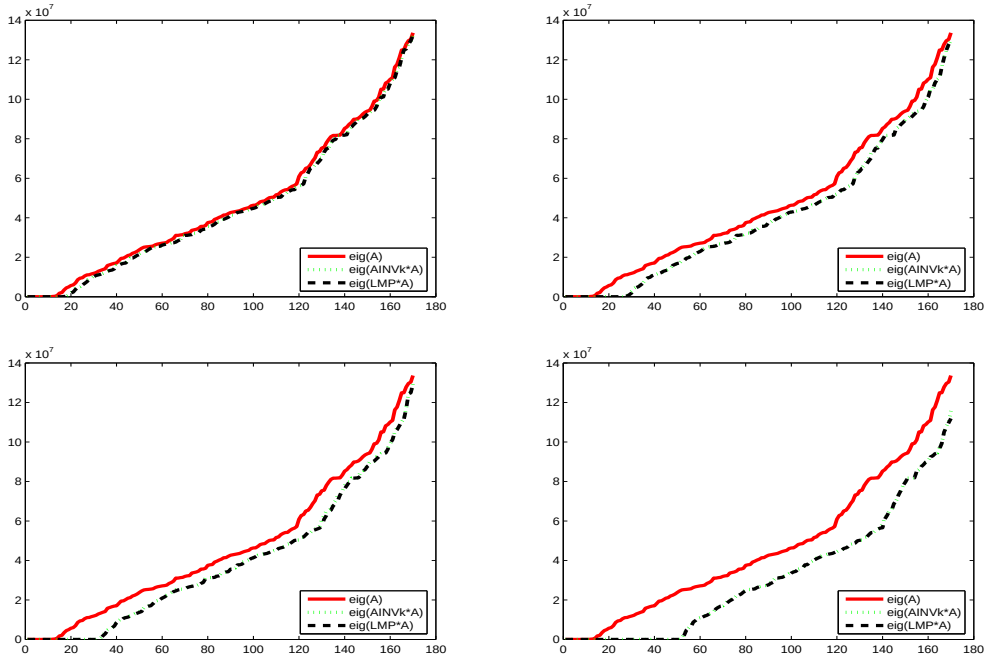

Figure 7.8: Comparison among the spectra $\Lambda[A_1]$ (eig(A)), $\Lambda[M_h^\sharp(0,1,I_{170})A_1]$ (eig$(AINVk \cdot A)$) and $\Lambda[\mathcal{M}_h A_1]$ (eig$(LMP \cdot A)$) — eigenvalues are sorted for simplicity; we set $x_0 = 100e$ for the CG and we used $h = 5$ (*top left*), $h = 15$ (*top right*), $h = 20$ (*bottom left*) and $h = 40$ (*bottom right*). We obtain results pretty similar to Figure 7.7.

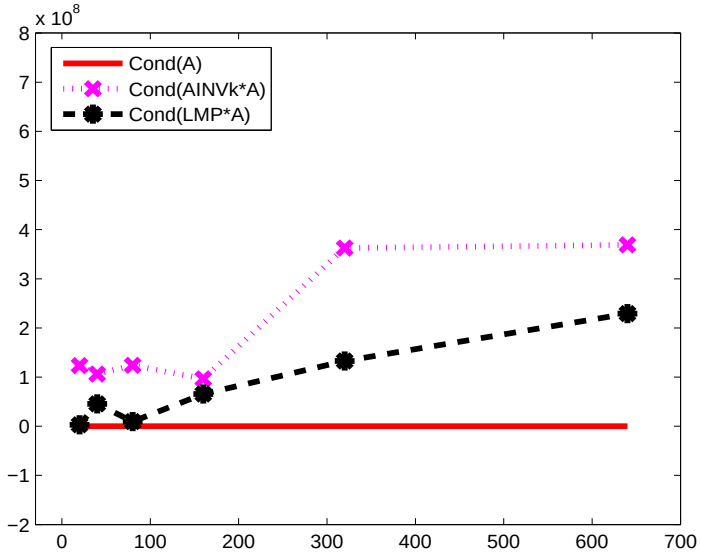

Figure 7.9: The condition number of matrix $\mathcal{A}$ ($Cond(A)$), along with the condition number of matrix $M_h^\sharp(0,1,I_n)\mathcal{A}$ ($Cond(AINVk \cdot A)$) and $\mathcal{M}_h\mathcal{A}$ ($Cond(LMP \cdot A)$), when $h \in \{20, 40, 80, 160, 320, 640\}$, using the matrix BCSSTK27 in the Harwell-Boeing Sparse Matrix collection (again Remark 3.1 does not apply). The two preconditioners show similar results with a slight preference for $\mathcal{M}_h$. The condition number of $M_h^\sharp(0,1,I_n)\mathcal{A}$ and $\mathcal{M}_h\mathcal{A}$ substantially satisfies the statement of Theorem 3.3.

Matrix collection (see `http://math.nist.gov/MatrixMarket/data/Harwell-Boeing`). This matrix has 56126 nonzero entries, with (on average) 46 nonzeros per row (column) and a bandwidth of 104. In addition, it is not diagonally dominant and has a condition number given by $7.7 \cdot 10^4$.

Then, we generated the vector $b \in \mathbb{R}^{1224}$, whose entries are in the uniform distribution $U[-10, 10]$. Again, we want to compare the performance of the two preconditioners $M_h^\sharp(0,1,I_n)$ and $\mathcal{M}_h$, considering the condition number (see Figure 7.9, where again the trick in Remark 3.1 does not apply) and the spectra (see Figure 7.10) of $M_h^\sharp(0,1,I_n)\mathcal{A}$ and $\mathcal{M}_h\mathcal{A}$, for different values of the parameter h. As for the Section 7.1, we set h in the range (see also [19])

$$h \in \{\ 20\ ,\ 40\ ,\ 80\ ,\ 160\ ,\ 320\ ,\ 640\ \}.$$

Observe that as suggested also in [19], a contraction for the condition number of the preconditioned matrix in principle can be induced by suitably scaling the linear system. However, the scaling parameter is in general unknown. In Figure 7.9 we find the comparison in terms of the condition number, which is again computed as in (7.2).

7.5 Test set 5

After the preliminary numerical tests in Sections 7.1-7.4 we can now apply our proposal on the sequence of linear systems arising in a well known optimization framework, namely truncated Newton methods. Again we tested our class of preconditioners vs. *LMP* preconditioners. In particular, we considered unconstrained optimization problems, which were solved using the standard linesearch-based truncated Newton method in Table 7.1, where the solution of the symmetric linear system (Newton's equation) $\nabla^2 f(z_k)d = -\nabla f(z_k)$ is required, at each outer iteration k. The fruitful use of preconditioning techniques within truncated Newton methods is clearly pointed out in several papers (see e.g. [23] and [28] for a survey).

As test problems we considered all the large scale unconstrained optimization problems from `CUTEr` [18] collection (112 test problems), with $n \in [1000, 10000]$. At the outset of the outer iteration k we computed the preconditioners $M_h^\sharp(0,1,I_n)$ and $\mathcal{M}_h$, using the information collected by the CG, after $h = 7$ (inner) iterations, when solving the equation $\nabla^2 f(z_k)d = -\nabla f(z_k)$ (for the choice $h = 7$ see [12, 25]). Then, from the 8-th (inner) iteration we adopted $M_h^\sharp(0,1,I_n)$ (and $\mathcal{M}_h$ respectively) as a preconditioner, for the solution of the linear system $\nabla^2 f(z_k)d = -\nabla f(z_k)$. All the parameters used within the preconditioning strategy, the truncated

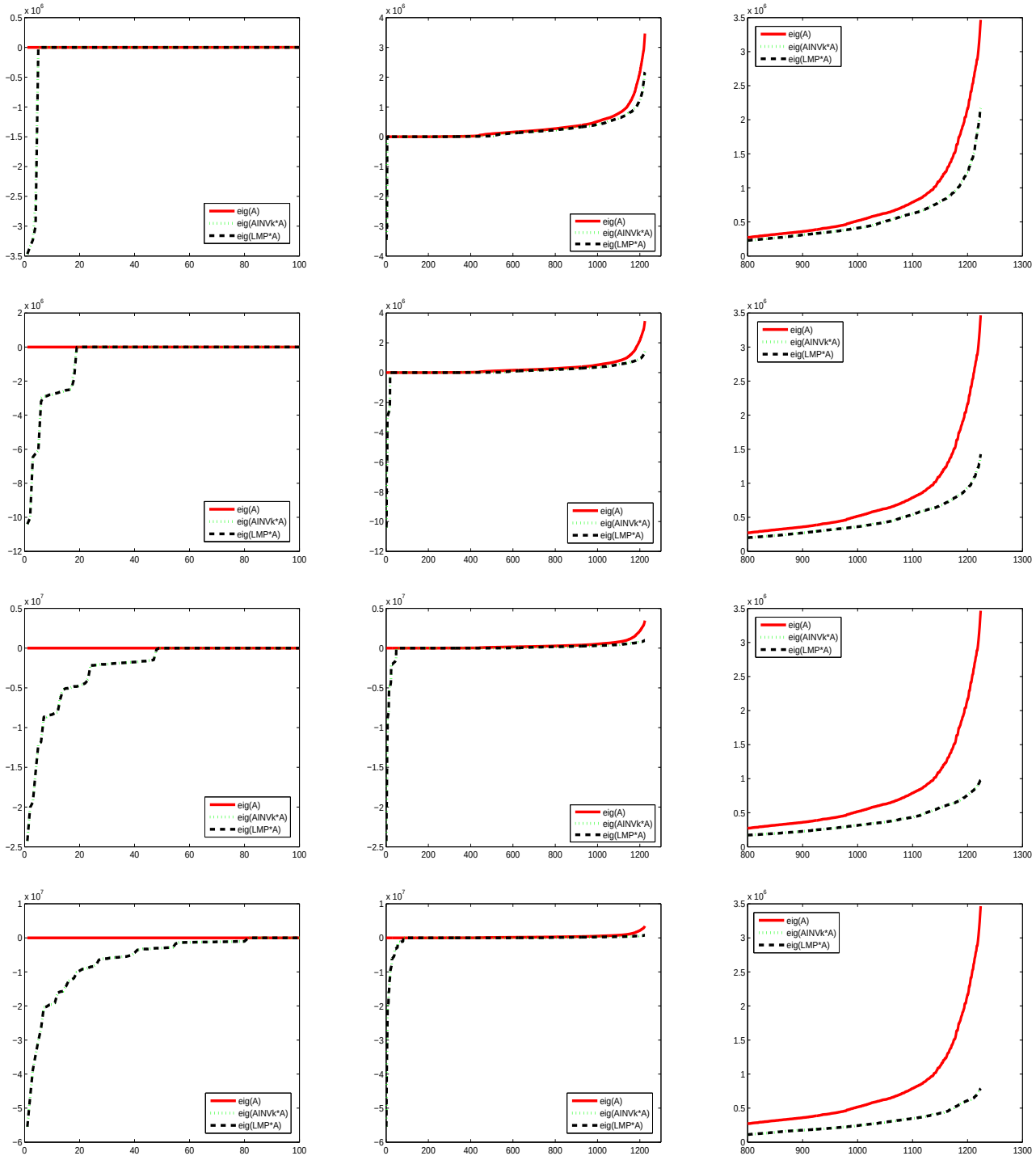

Figure 7.10: Comparison among the spectra of $\mathcal{A}$ (eig(A)), $M_h^\sharp(0,1,I_n)\mathcal{A}$ (eig($AINVk\cdot A$)) and $\mathcal{M}_h\mathcal{A}$ (eig($LMP\cdot A$)) — full spectra in the *center* and details on the *left-right*, for the matrix BCSSTK27 in the Harwell-Boeing Sparse Matrix collection, with $h \in \{80, 160, 320, 640\}$. The preconditioned matrices $M_h^\sharp(0,1,I_n)\mathcal{A}$ and $\mathcal{M}_h\mathcal{A}$ produce sequences of eigenvalues which are almost overlapped.

Set $z_0 \in \mathbb{R}^n$
Set $\eta_k \in [0,1)$ for any k, with $\{\eta_k\} \to 0$
OUTER ITERATIONS
for $k = 0, 1, \ldots$
Compute $\nabla f(z_k)$; if $\|\nabla f(z_k)\|$ is small then STOP
INNER ITERATIONS
Compute d_k which approximately solves $\nabla^2 f(z_k)d = -\nabla f(z_k)$
and satisfies the *truncation rule*
$$\|\nabla^2 f(z_k)d_k + \nabla f(z_k)\| \le \eta_k \|\nabla f(z_k)\|$$
Compute the steplength a_k by an Armijo-type linesearch procedure
Update $z_{k+1} = z_k + a_k d_k$
endfor

Table 7.1: The standard linesearch-based truncated Newton method we adopted.

scheme and the linesearch adopted were exactly those chosen in [12]. We show in Figure 7.11 the performance profiles (see [10]) where the comparison is summarized in terms of inner iterations. The profiles are drawn including three cases: $M_h^\sharp(0,1,I_n)$ is adopted, $\mathcal{M}_h$ is adopted and no preconditioning is considered. Moreover, for a fair comparison, only those test problems where the latter three schemes converge to the same stationary point were considered. As we can see, although our proposal yields two more failures, $M_h^\sharp(0,1,I_n)$ seems more efficient than $\mathcal{M}_h$, even though it is built using one less vector, i.e. it contains in principle less information on the Hessian matrix $\nabla^2 f(z_k)$.

As another test in optimization frameworks, we wanted to verify the impact of our preconditioner on the solution of a sequence of slightly changing linear systems. In particular, considering the truncated Newton scheme in Table 7.1, we first computed $M_h^\sharp(0,1,I_n)$ and $\mathcal{M}_h$ at the k-th outer iteration, by applying the CG. Then, at the $(k+1)$-th outer iteration we solved the linear system $\nabla^2 f(z_{k+1})d = -\nabla f(z_{k+1})$ using the latter preconditioners. Thus, we used at the $(k+1)$-th outer iteration the preconditioners computed at the k-th outer iteration. Note that this strategy was adopted in [25] using the preconditioner *PREQN*. The iteration index k was the first value such that

$$\frac{\|z_{k+1} - z_k\|}{a_k} \le 10^{-3}\|z_k\| \qquad \text{and} \qquad a_k \ge 0.95, \tag{7.4}$$

i.e. $\{z_k\}$ is likely converging to a local minimizer, so that the entries of the Hessian matrices $\nabla^2 f(z_k)$ and $\nabla^2 f(z_{k+1})$ are not expected to differ significantly. For the sake of brevity we report the results on just two test problems, with $n = 1000$, in the set of all the 112 optimization problems experienced. Very similar results were obtained for almost all the other test problems. The Hessian matrices in the two test problems reported were nearly singular, so that they represented instances where the CG has had difficulties to detect the solution (without stopping untimely), and may have yielded poor information on the Hessian matrix $\nabla^2 f(z_k)$. Nonetheless, the application of $M_h^\sharp(0,1,I_n)$ and $\mathcal{M}_h$ gave results on the guidelines of Theorem 3.3.

In Figures 7.12 and 7.14 we consider the problem `NONCVXUN`; we show the numerical results for values of h proposed in [25]. Observe that according to (7.4) we have here $k = 175$, and since z_{k+1} is close to z_k (i.e. we are eventually converging to a local minimizer) the Hessian matrix $\nabla^2 f(z_{k+1})$ is positive semidefinite. Again the eigenvalues larger than $+1$ in $\Lambda[\nabla^2 f(z_{k+1})]$ are scaled in $\Lambda[M_h^\sharp(0,1,I_n)\nabla^2 f(x_{k+1})]$ and $\Lambda[\mathcal{M}_h A]$.

Similarly we show in Figures 7.13 and 7.15 the results for the test function `NONDQUAR`; now we have $k = 40$ which satisfies (7.4). The test problems in this optimization framework, where the preconditioners $M_h^\sharp(0,1,I_n)$ and $\mathcal{M}_h$ are computed at the outer iteration k and used at the outer iteration $k+1$, confirm that the properties of Theorem 3.3 approximately hold also when $M_h^\sharp(0,1,I_n)$ and $\mathcal{M}_h$ are used on a short sequence of linear systems $A_k d = b_k$, when A_k changes *slightly* with k.

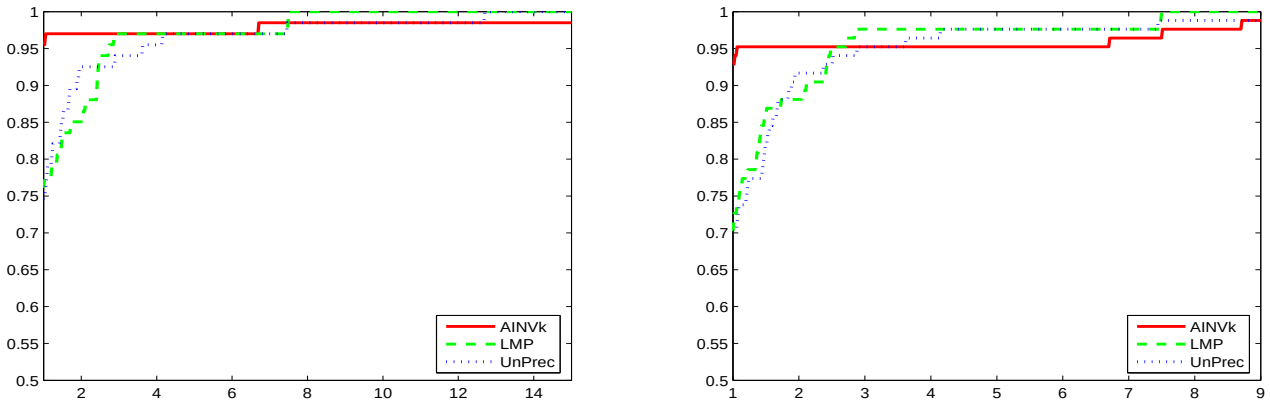

Figure 7.11: Performance profiles for a comparison in terms of number of inner iterations, on 112 large scale unconstrained `CUTEr` problems, among three truncated Newton's schemes, where $M_h^\sharp(0,1,I_n)$ (AINVk), $\mathcal{M}_h$ (LMP) and no preconditioner (UnPrec) is respectively adopted to solve Newton's equation $\nabla^2 f(z_k)d = -\nabla f(z_k)$. The two preconditioners show different performance (on the *left* we include only problems where no negative curvatures were encountered by the CG, on the *right* all test problems are included).

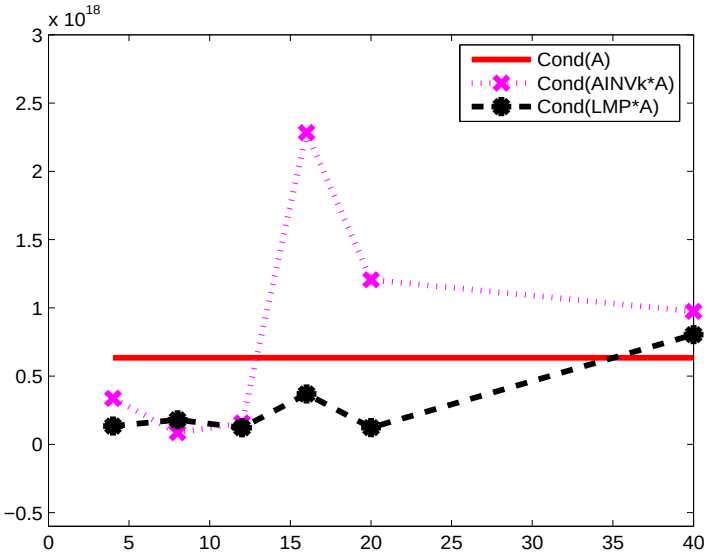

Figure 7.12: The condition number of matrix $\nabla^2 f(z_{k+1})$ $(Cond(A))$, along with the condition number of matrix $M_h^\sharp(0,1,I_n)\nabla^2 f(z_{k+1})$ $(Cond(AINVk \cdot A))$ and matrix $\mathcal{M}_h\nabla^2 f(z_{k+1})$ $(Cond(LMP \cdot A))$, when $h \in \{4,8,12,16,20,40\}$, for the optimization problem NONCVXUN (again Remark 3.1 does not apply). The CG did not stop prematurely and round-off yielded inaccurate results due to the high condition number of $\nabla^2 f(z_k)$. For some values of h the preconditioner $\mathcal{M}_h$ seems to perform slightly better than $M_h^\sharp(0,1,I_n)$. Nonetheless the condition number of $M_h^\sharp(0,1,I_n)\nabla^2 f(z_{k+1})$ and $\mathcal{M}_h\nabla^2 f(z_{k+1})$ remains of the same order of magnitude of $\nabla^2 f(z_{k+1})$. Considering that $n = 1000$ and at the value $k = 175$ we have $\|z_{175} - z_{176}\| \approx 0.083$, then z_{175} and z_{176} are pretty close.

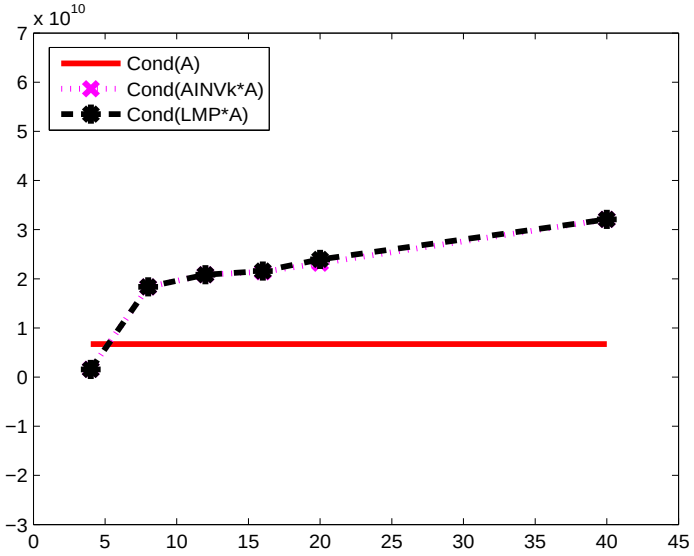

Figure 7.13: The condition number of matrix $\nabla^2 f(z_{k+1})$ ($Cond(A)$), along with the condition number of matrix $M_h^\sharp(0,1,I_n)\nabla^2 f(z_{k+1})$ ($Cond(AINVk \cdot A)$) and matrix $\mathcal{M}_h\nabla^2 f(z_{k+1})$ ($Cond(LMP \cdot A)$), when $h \in \{4,8,12,16,20,40\}$, for the optimization problem NONDQUAR (again Remark 3.1 does not apply). The two preconditioners show similar results. The CG did not stop prematurely and round-off yielded inaccurate results due to the high condition number of $\nabla^2 f(z_k)$. Nonetheless the condition number of $M_h^\sharp(0,1,I_n)\nabla^2 f(z_{k+1})$ and $\mathcal{M}_h\nabla^2 f(z_{k+1})$ remains of the same order of magnitude of $\nabla^2 f(z_{k+1})$. Considering that $n = 1000$ and at the value $k = 40$ we have $\|z_{40} - z_{41}\| \approx 0.203$, then z_{40} and z_{41} are pretty close.

8 Conclusions

We have given theoretical and numerical results for a new class of preconditioners, which are parameter dependent. The preconditioners can be built by using any Krylov-subspace method for the indefinite linear system (2.1), as well as L-BFGS updates, provided that the general conditions (2.2)-(2.3) in Assumption 2.1 are satisfied. We gave evidence that on several test problems and real applications, a few iterations of the Krylov-subspace method adopted may suffice to compute effective preconditioners. In particular, in many problems using a relatively small value of the index h, a significant information on the system matrix A can be captured.

On this guideline our proposal seems tailored also for those cases where a sequence of linear systems of the form

$$A_k x = b_k, \qquad k = 1, 2, \ldots \tag{8.1}$$

requires a solution (e.g., see also [8, 25] for details), where A_k slightly changes with the index k. In the latter case, the preconditioners $M_h^\sharp(a,\delta,D)$ in (3.5)-(3.6) can be computed applying the Krylov-subspace method to the first linear system $A_1 x = b_1$. Then, the resulting preconditioners can be used to efficiently solve (8.1) for $k = 2, 3, \ldots$

A full investigation was also included, where our proposal was compared with *LMP* preconditioners [19], showing that the two proposals have some similarities. In particular, though our preconditioners are also suited for indefinite linear systems, and are in general cheaper than *LMP* (see Section 4.1), when $A \succ 0$ the two proposals tend to generate a similar spectrum of the preconditioned matrix, endowed with similar theoretical properties (with *LMP* being apparently slightly superior).

Acknowledgments. The authors wish to thank both Jorge Nocedal, for his comments when the contents in this paper were at their early beginning, and Serge Gratton for his more recent suggestions.

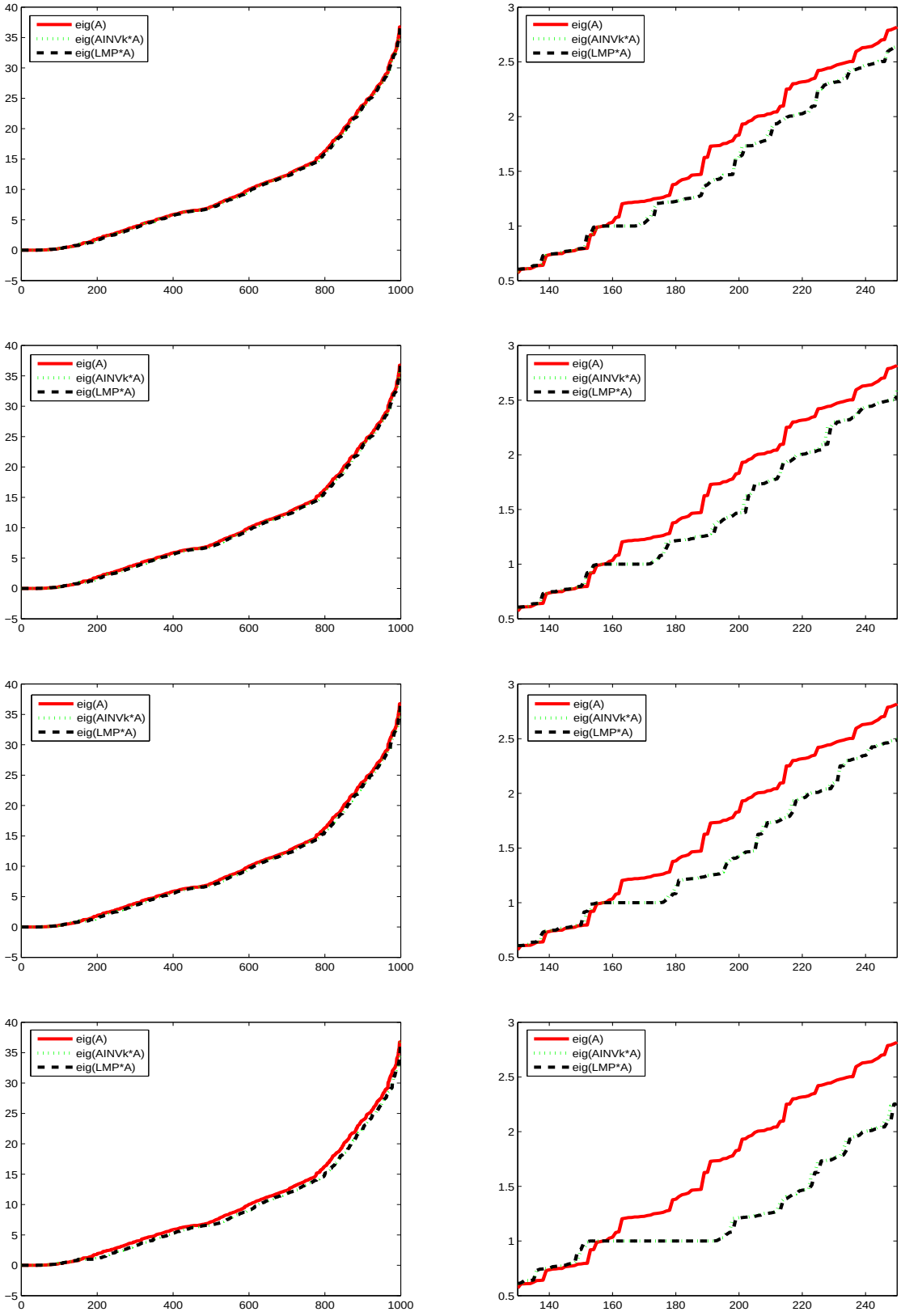

Figure 7.14: Comparison between the full spectra (*left*) and detailed spectra (*right*) of $\nabla^2 f(z_{k+1})$ (eig(A)), $M_h^\sharp(0,1,I_n)\nabla^2 f(z_{k+1})$ (eig$(AINVk \cdot A)$) and $\mathcal{M}_h \nabla^2 f(z_{k+1})$ (eig$(LMP \cdot A)$), for the optimization problem NONCVXUN, with $h \in \{12, 16, 20, 40\}$ and $k = 175$. The clustering near $+1$ of the eigenvalues of $M_h^\sharp(0,1,I_n)\nabla^2 f(x_{k+1})$ and $\mathcal{M}_h \nabla^2 f(z_{k+1})$, when h increases is evident.

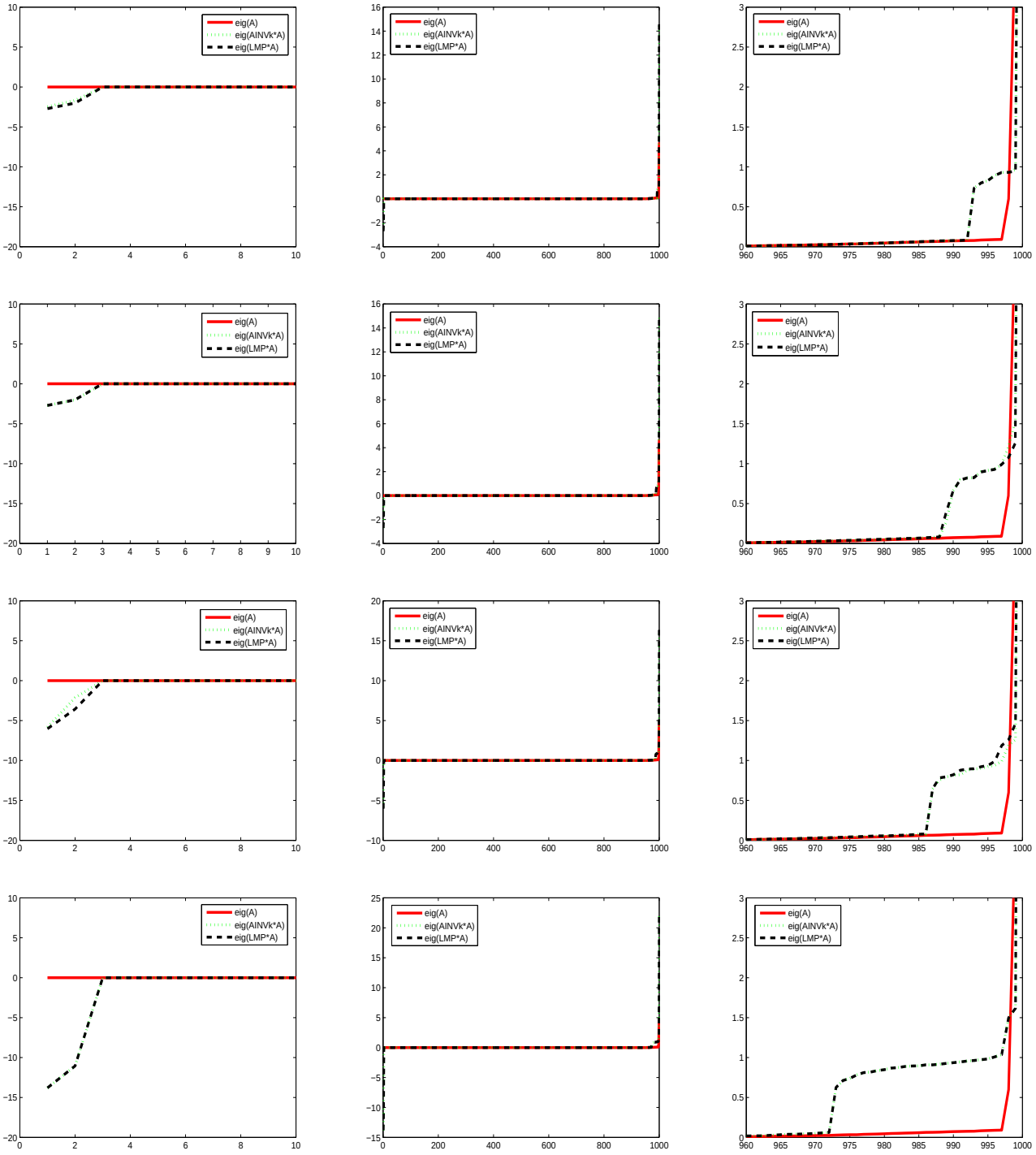

Figure 7.15: Comparison among the spectra of $\nabla^2 f(z_{k+1})$ (eig(A)), $M_h^\sharp(0,1,I_n)\nabla^2 f(z_{k+1})$ (eig$(AINVk \cdot A)$) and $\mathcal{M}_h \nabla^2 f(z_{k+1})$ (eig$(LMP \cdot A)$) — full spectra in the *center* and details on the *left-right*, for the optimization problem NONDQUAR, with $h \in \{12, 16, 20, 40\}$ and $k = 40$. Some nearly 0 eigenvalues in the spectrum $\Lambda[\nabla^2 f(z_{k+1})]$ are shifted to non-zero values in $\Lambda[M_h^\sharp(0,1,I_n)\nabla^2 f(z_{k+1})]$ and $\Lambda[\mathcal{M}_h \nabla^2 f(z_{k+1})]$, both on the left and on the right (nearly 1 right-shifted eigenvalues). Since many eigenvalues in $\Lambda[\nabla^2 f(z_{k+1})]$ are nearly 0, the preconditioners might be of scarce effect, unless large values of the parameter h are considered. $M_h^\sharp(0,1,I_n)$ and $\mathcal{M}_h$ show a similar behavior.

A Appendix

Lemma A.1 *Given the symmetric matrices $H \in \mathbb{R}^{h\times h}$, $P \in \mathbb{R}^{(n-h)\times(n-h)}$ and the matrix $\Phi \in \mathbb{R}^{h\times(n-h)}$, suppose*

$$\Phi^T H = \begin{bmatrix} z_1^T \\ \vdots \\ z_m^T \\ 0_{[n-(h+m)],h} \end{bmatrix}, \qquad z_1, \ldots, z_m \in \mathbb{R}^h, \tag{A.1}$$

with $H = \lambda[I_h + u_1 w_1^T + \cdots + u_p w_p^T]$, $p \le h$, $0 \le m \le h-p$, $\lambda \in \mathbb{R}$, $u_i, w_i \in \mathbb{R}^h$, $i = 1, \ldots, p$. Then, the symmetric matrix

$$\left(\begin{array}{c|c} H & H\Phi \\ \hline \Phi^T H & P \end{array}\right) \tag{A.2}$$

has the eigenvalue λ with multiplicity at least equal to $h - rk[w_1\ w_2\ \cdots\ w_p\ z_1\ z_2\ \cdots\ z_m]$.

Proof: Observe that H has the eigenvalue λ with a multiplicity at least $h-p$, since $Hs = \lambda s$ for any $s \perp span\{w_1, \ldots, w_p\}$. Moreover, imposing the condition (with $x_1,\ x_2$ not simultaneously zero vectors)

$$\left(\begin{array}{c|c} H & H\Phi \\ \hline \Phi^T H & P \end{array}\right) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

is equivalent to impose the conditions

$$\begin{cases} H(x_1 + \Phi x_2) = \lambda x_1 \\ \\ \Phi^T H x_1 + P x_2 = \lambda x_2. \end{cases}$$

By (A.1), choosing $x_2 = 0$ and x_1 any h-real vector such that $x_1 \perp span\{w_1, \ldots, w_p, z_1, \ldots, z_m\}$, then λ is eigenvalue of (A.2) with multiplicity given by h minus the largest number of linearly independent vectors in the set $\{w_1, \ldots, w_p, z_1, \ldots, z_m\}$. □

Proof of Theorem 3.1.

Let $N = [R_h \mid Du_{h+1} \mid DR_{n,h+1}]$, where N is nonsingular by hypothesis. Observe that for $h \le n-1$ the preconditioners $M_h^\sharp(a, \delta, D)$ may be rewritten as

$$M_h^\sharp(a,\delta,D) = N \left[\begin{array}{c|c} \left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right)^{-1} & 0 \\ \hline 0 & I_{n-(h+1)} \end{array}\right] N^T, \qquad h \le n-1. \tag{A.3}$$

The property a) follows from the symmetry of T_h. In addition, observe that $R_{n,h+1}^T R_{n,h+1} = I_{n-(h+1)}$. Thus, from (A.3) the matrix $M_h^\sharp(a,\delta,D)$ is nonsingular if and only if the matrix

$$\left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right) \tag{A.4}$$

is invertible. Furthermore, by a direct computation we observe that for $h \le n-1$ the following identity holds

$$\left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right) = \left(\begin{array}{c|c} I_h & 0 \\ \hline \frac{a}{\delta^2} e_h^T |T_h|^{-1} & 1 \end{array}\right) \left(\begin{array}{c|c} \delta^2|T_h| & 0 \\ \hline 0 & 1 - \frac{a^2}{\delta^2} e_h^T |T_h|^{-1} e_h \end{array}\right) \left(\begin{array}{c|c} I_h & \frac{a}{\delta^2}|T_h|^{-1}e_h \\ \hline 0 & 1 \end{array}\right). \tag{A.5}$$

Thus, since T_h is nonsingular and $\delta \neq 0$, for $h \le n-1$ the determinant of matrix (A.4) is nonzero if and only if $a \neq \pm\delta(e_h^T|T_h|^{-1}e_h)^{-1/2}$. Finally, for $h = n$ the matrix $M_h^\sharp(a,\delta,D)$ is nonsingular, since R_n and T_n are nonsingular in (3.6).

As regards b), recalling that $R_h^T R_h = I_h$ and $|T_h|$ is nonsingular from Assumption 2.1, when $h \le n-1$ relations (3.3) and (A.3) trivially yield the result, as well as (3.4) and (3.6) for the case $h = n$.

As regards c), observe that from (A.3) the matrix $M_h^\sharp(a,\delta,D)$ is positive definite, as long as the matrix (A.4) is positive definite. Thus, from (A.5) and relation $|T_h| \succ 0$ we immediately infer that $M_h^\sharp(a,\delta,D)$ is positive definite as long as $|a| < |\delta|(e_h^T|T_h|^{-1}e_h)^{-1/2}$. Moreover, we recall that when $D = I_n$ then N is orthogonal.

Item d) may be proved considering that $D = I_n$ and computing the eigenvalues of the matrix

$$\left[M_h^\sharp(a,\delta,I_n)A\right]\left[M_h^\sharp(a,\delta,I_n)A\right]^T = M_h^\sharp(a,\delta,I_n)A^2M_h^\sharp(a,\delta,I_n).$$

On this purpose, for $h \le n-1$ we have for $M_h^\sharp(a,\delta,I_n)A^2M_h^\sharp(a,\delta,I_n)$ the expression (see (A.3))

$$\begin{array}{l} M_h^\sharp(a,\delta,I_n)A^2M_h^\sharp(a,\delta,I_n) \;=\; \\ N\left[\begin{array}{c|c} \left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right)^{-1} & 0 \\ \hline 0 & I_{n-(h+1)} \end{array}\right] C \left[\begin{array}{c|c} \left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right)^{-1} & 0 \\ \hline 0 & I_{n-(h+1)} \end{array}\right] N^T \end{array} \tag{A.6}$$

where $C \in \mathbb{R}^{n\times n}$, with

$$C = N^TA^2N = \left[\begin{array}{c|c|c} R_h^TA^2R_h & R_h^TA^2u_{h+1} & R_h^TA^2R_{n,h+1} \\ \hline u_{h+1}^TA^2R_h & u_{h+1}^TA^2u_{h+1} & u_{h+1}^TA^2R_{n,h+1} \\ \hline R_{n,h+1}^TA^2R_h & R_{n,h+1}^TA^2u_{h+1} & R_{n,h+1}^TA^2R_{n,h+1} \end{array}\right].$$

From (2.2) and the symmetry of T_h we obtain

$$R_h^TA^2R_h \;=\; (AR_h)^T(AR_h) \;=\; (R_hT_h + \rho_{h+1}u_{h+1}e_h^T)^T(R_hT_h + \rho_{h+1}u_{h+1}e_h^T) = T_h^2 + \rho_{h+1}^2e_he_h^T \tag{A.7}$$

$$R_h^TA^2u_{h+1} \;=\; (AR_h)^TAu_{h+1} \;=\; v_1 \in \mathbb{R}^h, \tag{A.8}$$

and considering relation (2.2) we obtain

$$\begin{array}{rcl} AR_{h+1} &=& A(R_h \mid u_{h+1}) \;=\; R_{h+1}T_{h+1} + \rho_{h+2}u_{h+2}e_{h+1}^T \\ &=& (R_h \mid u_{h+1})\left(\begin{array}{c|c} T_h & \rho_{h+1}e_h \\ \hline \rho_{h+1}e_h^T & t_{h+1,h+1} \end{array}\right) + \rho_{h+2}u_{h+2}e_{h+1}^T \end{array}$$

i.e.

$$\begin{array}{rcl} AR_h &=& R_hT_h + \rho_{h+1}u_{h+1}e_h^T \\ Au_{h+1} &=& \rho_{h+1}u_h + t_{h+1,h+1}u_{h+1} + \rho_{h+2}u_{h+2}, \end{array} \tag{A.9}$$

so that

$$A^2R_h = (AR_h)T_h + \rho_{h+1}Au_{h+1}e_h^T = (R_hT_h + \rho_{h+1}u_{h+1}e_h^T)T_h + \rho_{h+1}(\rho_{h+1}u_h + t_{h+1,h+1}u_{h+1} + \rho_{h+2}u_{h+2})e_h^T.$$

As a consequence, from (A.9) we also have that $Au_{h+2} = span\{u_{h+1}, u_{h+2}, u_{h+3}\}$ and

$$\begin{array}{rcl} R_h^TA^2R_{n,h+1} &=& (A^2R_h)^TR_{n,h+1} = \rho_{h+1}(\rho_{h+2}u_{h+2}e_h^T)^TR_{n,h+1} = \\ &=& \rho_{h+1}\rho_{h+2}\begin{pmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{h\times[n-(h+1)]} = \rho_{h+1}\rho_{h+2}E_{h,1}, \\ u_{h+1}^TA^2u_{h+1} &=& c > 0 \\ u_{h+1}^TA^2R_{n,h+1} &=& \left[A(\rho_{h+1}u_h + t_{h+1,h+1}u_{h+1} + \rho_{h+2}u_{h+2})\right]^TR_{n,h+1} = \\ &=& \left[A(t_{h+1,h+1}u_{h+1} + \rho_{h+2}u_{h+2})\right]^TR_{n,h+1} = (\alpha \;\; \beta \;\; 0\cdots0) \in \mathbb{R}^{n-(h+1)} \end{array}$$

with $\alpha, \beta \in \mathbb{R}$ and

$$R_{n,h+1}^TA^2R_{n,h+1} = V_2 \in \mathbb{R}^{[n-(h+1)]\times[n-(h+1)]},$$

where $E_{i,j}$ has all zero entries but $+1$ at position (i,j). Thus,

$$C \;=\; \left[\begin{array}{c|c|c} T_h^2 + \rho_{h+1}^2e_he_h^T & v_1 & \rho_{h+1}\rho_{h+2}E_{h,1} \\ \hline v_1^T & c & \alpha \;\; \beta \;\; 0 \;\cdots\; 0 \\ \hline & \alpha & \\ & \beta & \\ \rho_{h+1}\rho_{h+2}E_{1,h} & 0 & V_2 \\ & \vdots & \\ & 0 & \end{array}\right].$$

Moreover, from (A.5) we can readily infer that

$$\left[\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right]^{-1} = \left(\begin{array}{c|c} I_h & -\frac{a}{\delta^2}|T_h|^{-1}e_h \\ \hline 0 & 1 \end{array}\right)\left(\begin{array}{c|c} \frac{1}{\delta^2}|T_h|^{-1} & 0 \\ \hline 0 & \frac{1}{1-\frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h} \end{array}\right)\left(\begin{array}{c|c} I_h & 0 \\ \hline -\frac{a}{\delta^2}e_h^T|T_h|^{-1} & 1 \end{array}\right)$$

$$= \left(\begin{array}{c|c} \frac{1}{\delta^2}|T_h|^{-1} - \frac{a}{\delta^4}\omega|T_h|^{-1}e_he_h^T|T_h|^{-1} & \frac{\omega}{\delta^2}|T_h|^{-1}e_h \\ \hline \frac{\omega}{\delta^2}e_h^T|T_h|^{-1} & -\frac{\omega}{a} \end{array}\right), \tag{A.10}$$

with

$$\omega = -\frac{a}{1-\frac{a^2}{\delta^2}e_h^T|T_h|^{-1}e_h}. \tag{A.11}$$

Now, recalling that since $D = I_n$ then $N = [R_h \mid u_{h+1} \mid R_{n,h+1}]$, for any $h \le n-1$ we obtain from (A.6)

$$M_h^\sharp(a,\delta,I_n)A^2M_h^\sharp(a,\delta,I_n) =$$

$$N\left[\begin{array}{c|c} \left[\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right]^{-1}\left[\begin{array}{c|c} T_h^2+\rho_{h+1}^2e_he_h^T & v_1 \\ \hline v_1^T & c \end{array}\right]\left[\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right]^{-1} & \left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right)^{-1}\left(\begin{array}{c} \rho_{h+1}\rho_{h+2}E_{h,1} \\ \hline \alpha\ \beta\ 0\ \cdots\ 0 \end{array}\right) \\ \hline \left(\begin{array}{c} \rho_{h+1}\rho_{h+2}E_{h,1} \\ \hline \alpha\ \beta\ 0\ \cdots\ 0 \end{array}\right)^T\left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right)^{-1} & V_2 \end{array}\right]N^T,$$

with

$$\left(\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right)^{-1}\left(\begin{array}{c} \rho_{h+1}\rho_{h+2}E_{h,1} \\ \hline \alpha\ \beta\ 0\ \cdots\ 0 \end{array}\right) = \left(\begin{array}{cc} * & * \\ \vdots & \vdots \\ * & * \end{array}\ \begin{array}{c} \\ 0_{h+1,[n-(h+3)]} \\ \\ \end{array}\right) \in \mathbb{R}^{(h+1)\times[n-(h+1)]},$$

where the '$*$' indicates entries whose computation is not relevant to our purposes.

Now, considering the second last relation, we focus on computing the submatrix $H_{h\times h}$ corresponding to the first h rows and h columns of the matrix

$$\left[\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right]^{-1}\left[\begin{array}{c|c} T_h^2+\rho_{h+1}^2e_he_h^T & v_1 \\ \hline v_1^T & c \end{array}\right]\left[\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right]^{-1}. \tag{A.12}$$

After a brief computation, from (A.10) and (A.12) we obtain for the submatrix $H_{h\times h}$

$$\begin{aligned} H_{h\times h} = {} & \left[\left(\frac{1}{\delta^2}|T_h|^{-1} - \frac{a}{\delta^4}\omega|T_h|^{-1}e_he_h^T|T_h|^{-1}\right)\left(T_h^2+\rho_{h+1}^2e_he_h^T\right) + \right. \\ & \left.\frac{\omega}{\delta^2}|T_h|^{-1}e_hv_1^T\right]\cdot\left[\frac{1}{\delta^2}|T_h|^{-1} - \frac{a}{\delta^4}\omega|T_h|^{-1}e_he_h^T|T_h|^{-1}\right] + \\ & \left[\left(\frac{1}{\delta^2}|T_h|^{-1} - \frac{a}{\delta^4}\omega|T_h|^{-1}e_he_h^T|T_h|^{-1}\right)v_1 + \frac{\omega}{\delta^2}c|T_h|^{-1}e_h\right]\cdot\frac{\omega}{\delta^2}e_h^T|T_h|^{-1}, \end{aligned}$$

and for the case of T_h not factorizable, from (2.3) we obtain (a similar analysis holds for the case of T_h positive definite, too)

$$\begin{aligned} H_{h\times h} = {} & \left[\frac{1}{\delta^2}V_h\hat{I}_hV_h^TT_h + \frac{\rho_{h+1}^2}{\delta^2}|T_h|^{-1}e_he_h^T - \frac{a}{\delta^4}\omega|T_h|^{-1}e_he_h^TV_h\hat{I}_hV_h^TT_h\right. \\ & \left. - \frac{a}{\delta^4}\omega\rho_{h+1}^2(e_h^T|T_h|^{-1}e_h)|T_h|^{-1}e_he_h^T + \frac{\omega}{\delta^2}|T_h|^{-1}e_hv_1^T\right]\cdot\left[\frac{1}{\delta^2}|T_h|^{-1} - \frac{a}{\delta^4}\omega|T_h|^{-1}e_he_h^T|T_h|^{-1}\right] \\ & + \frac{\omega}{\delta^2}\left[\frac{1}{\delta^2}|T_h|^{-1}v_1 - \frac{a}{\delta^4}\omega|T_h|^{-1}e_he_h^T|T_h|^{-1}v_1 + \frac{\omega}{\delta^2}c|T_h|^{-1}e_h\right]e_h^T|T_h|^{-1}. \end{aligned}$$

Recalling that $(V_h\hat{I}_hV_h^T)(V_h\hat{I}_hV_h^T) = I_h$ (so that $e_h^T(V_h\hat{I}_hV_h^T)(V_h\hat{I}_hV_h^T)e_h = 1$), from the last relation we finally have for $H_{h\times h}$ the expression

$$H_{h\times h} = \frac{1}{\delta^4}\left\{I_h + \left[\eta|T_h|^{-1}e_h - \frac{a\omega}{\delta^2}e_h + \omega|T_h|^{-1}v_1\right]e_h^T|T_h|^{-1} + \omega|T_h|^{-1}e_h\left[v_1^T|T_h|^{-1} - \frac{a}{\delta^2}e_h^T\right]\right\}, \tag{A.13}$$

where

$$\eta = \rho_{h+1}^2 - 2\frac{a}{\delta^2}\omega\rho_{h+1}^2(e_h^T|T_h|^{-1}e_h) + \frac{a^2\omega^2}{\delta^4} + \frac{a^2}{\delta^4}\omega^2\rho_{h+1}^2(e_h^T|T_h|^{-1}e_h)^2 - 2\frac{a}{\delta^2}\omega^2(e_h^T|T_h|^{-1}v_1) + \omega^2c; \tag{A.14}$$

moreover, since $M_h^\sharp(a,\delta,I_n)A^2M_h^\sharp(a,\delta,I_n) \succ 0$ then also $H_{h\times h}$ is positive definite.
Let us now define the subspace (see the vectors which define the dyads in relation (A.13))

$$\mathcal{T}_2 = span\left\{|T_h|^{-1}e_h\ ,\ \omega\left[|T_h|^{-1}v_1 - \frac{a}{\delta^2}e_h\right]\right\}. \tag{A.15}$$

Observe that since $D = I_n$ then after some computation $v_1 = \rho_{h+1}\left[T_h + t_{h+1,h+1}I_h\right]e_h$. Thus, from (A.15) the subspace $\mathcal{T}_2$ has dimension 2, unless

(i) T_h is proportional to I_h,

(ii) $a = 0$ (which from (A.11) also implies $\omega = 0$).

We analyze separately the two cases. The condition *(i)* cannot hold since (2.2) would imply that the vector Au_i is proportional to u_i, $i = 1,\ldots,h-1$, i.e. the Krylov-subspace method had to stop at the very first iteration, since the Krylov-subspace generated at the first iteration did not change. As a consequence, considering any subspace $\mathcal{S}_{h-2} \subseteq \mathbb{R}^n$, such that $\mathcal{S}_{h-2} \bigoplus \mathcal{T}_2 = \mathbb{R}^h$, we can select any orthonormal basis $\{s_1,\ldots,s_{h-2}\}$ of the subspace $\mathcal{S}_{h-2}$ so that (see (A.13)) the $h-2$ vectors $\{s_1,\ldots,s_{h-2}\}$ can be thought as (the first) $h-2$ eigenvectors of the matrix $H_{h\times h}$, corresponding to the eigenvalue $+1/\delta^4$.
Now, from the formula after (A.11) the eigenvalues of $M_h^\sharp(a,\delta,I_n)A^2M_h^\sharp(a,\delta,I_n)$ coincide with the eigenvalues of (we recall that since $M_h^\sharp(a,\delta,I_n)A^2M_h^\sharp(a,\delta,I_n) \succ 0$ then $H_{h\times h} \succ 0$)

$$\left(\begin{array}{c|cc} H_{h\times h} & H_{h\times h}\Phi & \\ \hline & * & *\cdots\cdots*\\ \hline \Phi^T H_{h\times h} & * & \\ & \vdots & V_2 \\ & * & \end{array}\right), \qquad \Phi = H_{h\times h}^{-1}\left(z_1\ z_2\ z_3\ 0_{h\times[n-(h+3)]}\right), \quad z_i \in \mathbb{R}^h, \tag{A.16}$$

which becomes, after setting

$$P = \left(\begin{array}{c|c} * & *\cdots\cdots*\\ \hline * & \\ \vdots & V_2 \\ * & \end{array}\right),$$

of the form

$$\left[\begin{array}{c|c} H_{h\times h} & H_{h\times h}\Phi \\ \hline \Phi^T H_{h\times h} & P \end{array}\right].$$

Thus, using Lemma A.1 with $w_1 = |T_h|^{-1}e_h$, $w_2 = \omega\left[|T_h|^{-1}v_1 - a/\delta^2 e_h\right]$ and $m = 3$, recalling that $T_h \succ 0$ or T_h is not factorizable, and observing that we have by (A.7)

$$\begin{aligned}\left[\begin{array}{c} z_1 \\ \hline * \end{array}\right] &= \left[\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right]^{-1}\left[\begin{array}{c|c} T_h^2 + \rho_{h+1}^2 e_h e_h^T & v_1 \\ \hline v_1^T & c \end{array}\right]\left[\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right]^{-1} e_{h+1} \\ &= \left[\begin{array}{c|c} \frac{1}{\delta^2}|T_h|^{-1} - \frac{a}{\delta^4}\omega|T_h|^{-1}e_h e_h^T|T_h|^{-1} & \frac{\omega}{\delta^2}|T_h|^{-1}e_h \\ \hline \frac{\omega}{\delta^2}e_h^T|T_h|^{-1} & -\frac{\omega}{a} \end{array}\right]\left(\begin{array}{c} \frac{\omega}{\delta^2}T_h^2|T_h|^{-1}e_h + \rho_{h+1}^2\frac{\omega}{\delta^2}(e_h^T|T_h|^{-1}e_h)e_h - \frac{\omega}{a}v_1 \\ \hline \frac{\omega}{\delta^2}v_1^T|T_h|^{-1}e_h - \frac{c\omega}{a} \end{array}\right)\end{aligned}$$

so that $z_1 \in span\left\{\omega e_h\ ,\ \omega|T_h|^{-1}e_h\ ,\ |T_h|^{-1}v_1\right\}$,

$$\begin{aligned}\left[\begin{array}{c} z_2 \\ \hline * \end{array}\right] &= \left[\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right]^{-1}\left(\begin{array}{c} 0 \\ \vdots \\ 0 \\ \rho_{h+1}\rho_{h+2} \\ \alpha \end{array}\right) \\ &= \left[\begin{array}{c} \frac{\rho_{h+1}\rho_{h+2}}{\delta^2}|T_h|^{-1}e_h - \rho_{h+1}\rho_{h+2}\frac{a\omega}{\delta^4}|T_h|^{-1}e_h(e_h^T|T_h|^{-1}e_h) + \frac{\alpha\omega}{\delta^2}|T_h|^{-1}e_h \\ \hline * \end{array}\right]\end{aligned}$$

so that $z_2 \in span\{|T_h|^{-1}e_h\}$, and

$$\left[\begin{array}{c} z_3 \\ \hline * \end{array}\right] = \left[\begin{array}{c|c} \delta^2|T_h| & ae_h \\ \hline ae_h^T & 1 \end{array}\right]^{-1} \left(\begin{array}{c} 0 \\ \vdots \\ 0 \\ 0 \\ \beta \end{array}\right) = \left[\begin{array}{c} \frac{\beta\omega}{\delta^2}|T_h|^{-1}e_h \\ \hline * \end{array}\right]$$

so that $z_3 \in span\{\omega|T_h|^{-1}e_h\}$, we conclude that considering the expression of $H_{h\times h}$, at least $h-3$ eigenvalues of $M_h^\sharp(a,\delta,I_n)A^2M_h^\sharp(a,\delta,I_n)$ coincide with $+1/\delta^4$. As a consequence, the matrix $M_h^\sharp(a,\delta,I_n)A$ has at least $h-3$ singular values equal to $+1/\delta^2$, which proves the first statement of d).
As regards the case (ii) with $a=0$, observe that by the definition (A.11) of ω, $a=0$ implies $\omega=0$, and from relations (A.13)-(A.14), we have $H_{h\times h} = 1/\delta^4[I_h + \rho_{h+1}^2|T_h|^{-1}e_he_h^T|T_h|^{-1}]$. Thus, the subspace $\mathcal{T}_2$ in (A.15) reduces to $\mathcal{T}_1 = span\{|T_h|^{-1}e_h\}$. Now, reasoning as in the case (i), we conclude that the matrix $M_h^\sharp(a,\delta,I_n)A$ has at least $(h-2)$ singular values equal to $+1/\delta^2$.

As regards item e), observe that for $h=n$ the matrix R_n is orthogonal, so that by (3.4) and (3.6) $\Lambda[M_h^\sharp(a,\delta,D)] = \Lambda[M_h^{-1}] = \Lambda[|T_h|^{-1}]$. Furthermore, by (2.2) and (3.6) we have for the case of T_n not factorizable (a similar analysis holds for the case T_h factorizable, too)

$$M_n^\sharp(a,\delta,D)A = M_n^{-1}A = R_n|T_n|^{-1}R_n^TR_nT_nR_n^T = R_nV_n\hat{I}_nV_n^TR_n^T = (R_nV_n)\hat{I}_n(R_nV_n)^T. \tag{A.17}$$

Since both R_n and V_n are orthogonal so is the matrix R_nV_n; thus, relation (A.17) proves that $M_n^\sharp(a,\delta,D)A$ has all the n eigenvalues in the set $\{-1,+1\}$. □

References

[1] J. Baglama, D. Calvetti, G. Golub, and L. Reichel, *Adaptively preconditioned GMRES algorithms*, SIAM Journal on Scientific Computing, 20 (1998), pp. 243–269.

[2] M. Benzi, *Preconditioning techniques for large linear systems: a survey*, Journal of Computational Physics, 182 (2002), pp. 418–477.

[3] M. Benzi, J. Cullum, and M. Tuma, *Robust approximate inverse preconditioner for the conjugate gradient method*, SIAM Journal on Scientific Computing, 22 (2000), pp. 1318–1332.

[4] M. Benzi and M. Tuma, *A comparative study of sparse approximate inverse preconditioners*, Applied Numerical Mathematics, 30 (1999), pp. 305–340.

[5] L. Bergamaschi, J. Gondzio, and G. Zilli, *Preconditioning indefinite linear systems in interior point methods for optimization*, Computational Optimization and Application, 28 (2004), pp. 149–171.

[6] D. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas – Second Edition*, Princeton University Press, Princeton, 2009.

[7] J. Bunch and L. Kaufman, *Some stable methods for calculating inertia and solving symmetric linear equations*, Mathematics of Computations, 31 (1977), pp. 163–179.

[8] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust–region methods*, MPS–SIAM Series on Optimization, Philadelphia, PA, 2000.

[9] J. Dennis and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear equations*, Prentice-Hall, Englewood Cliffs, 1983.

[10] E. D. Dolan and J. Moré, *Benchmarking optimization software with performance profiles*, Mathematical Programming, 91 (2002), pp. 201–213.

[11] G. Fasano, *Planar–conjugate gradient algorithm for large–scale unconstrained optimization, Part 1: Theory*, Journal of Optimization Theory and Applications, 125 (2005), pp. 523–541.

[12] G. Fasano and M. Roma, *Preconditioning Newton-Krylov methods in nonconvex large scale optimization*, submitted to Computational Optimization and Applications.

[13] ——, *Iterative computation of negative curvature directions in large scale optimization*, Computational Optimization and Applications, 38 (2007), pp. 81–104.

[14] S. Geman, *A limit theorem for the norm of random matrices*, The Annals of Probability, 8 (1980), pp. 252–261.

[15] P. E. Gill, W. Murray, D. B. Ponceleon, and M. A. Saunders, *Preconditioners for indefinite systems arising in optimization*, SIAM Journal on Matrix Analysis and Applications, 13 (1992), pp. 292–311.

[16] P. E. Gill, M. A. Saunders, and J. R. Shinnerl, *On the stability of Cholesky factorization for symmetric quasidefinite systems*, SIAM Journal on Matrix Analysis and Applications, 17 (1996), pp. 35–46.

[17] G. Golub and C. Van Loan, *Matrix Computations*, The John Hopkins Press, Baltimore, 1996. Third Edition.

[18] N. I. M. Gould, D. Orban, and P. L. Toint, CUTEr *(and* sifdec*), a constrained and unconstrained testing environment, revised*, ACM Transaction on Mathematical Software, 29 (2003), pp. 373–394.

[19] S. Gratton, A. Sartenaer, and J. Tshimanga, *On a class of limited memory preconditioners for large scale linear systems with multiple right-hand sides*, SIAM Journal on Optimization, 21 (2011), pp. 912–935.

[20] M. Hestenes, *Conjugate Direction Methods in Optimization*, Springer Verlag, New York, 1980.

[21] N. Higham, *Accuracy and stability of numerical algorithms*, SIAM, Philadelphia, PA, 2002. Second Edition.

[22] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1999.

[23] L. Lukšan, C. Matonoha, and J. Vlček, *Band preconditioners for the matrix-free truncated Newton method*, Techical Report V-1079, Institute of Computer Science AS CR, Pod Vodarenskou Vezi 2, 18207 Prague 8, 2010.

[24] MATLAB, *MATLAB Release 2011a*, The MathWorks Inc., 2011.

[25] J. Morales and J. Nocedal, *Automatic preconditioning by limited memory quasi–Newton updating*, SIAM Journal on Optimization, 10 (2000), pp. 1079–1096.

[26] J. L. Morales and J. Nocedal, *Algorithm* PREQN*: Fortran 77 subroutine for preconditioning the conjugate gradient method*, ACM Transaction on Mathematical Software, 27 (2001), pp. 83–91.

[27] R. Nabben and C. Vuik, *A comparison of deflation and the balancing preconditioner*, SIAM Journal on Scientific Computing, 27 (2006), pp. 1742–1759.

[28] S. Nash, *A survey of truncated-Newton methods*, Journal of Computational and Applied Mathematics, 124 (2000), pp. 45–59.

[29] J. Nocedal and S. Wright, *Numerical Optimization (Springer Series in Operations Research and Financial Engineering) - Second Edition*, Springer, New York, 2000.

[30] D. O'Leary and A. Yeremin, *The linear algebra of block quasi-Newton algorithms*, Linear Algebra and its Applications, 212/213 (1994), pp. 153–168.

[31] C. Paige and M. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM Journal on Numerical Analysis, 12 (1975), pp. 617–629.

[32] Y. Saad, *Iterative Methods for Sparse Linear Systems – Second Edition*, SIAM, Philadelphia, PA, 2003.

[33] V. Simoncini and D. Szyld, *Recent computational developments in Krylov subspace methods for linear systems*, Numerical Linear Algebra with Applications, 14 (2007), pp. 1–59.

[34] J. Stoer, *Solution of large linear systems of equations by conjugate gradient type methods*, in Mathematical Programming. The State of the Art, A. Bachem, M.Grötschel, and B. Korte, eds., Berlin Heidelberg, 1983, Springer-Verlag, pp. 540–565.

[35] R. J. Vanderbei, *Symmetric quasi-definite matrices*, SIAM Journal of Optimization, 5 (1995), pp. 100–113.