A generalization of the optimal diagonal approximate inverse preconditioner[☆]

Luis González*, Antonio Suárez, Eduardo Rodríguez

Department of Mathematics, University of Las Palmas de Gran Canaria, 35017 Las Palmas de Gran Canaria, Spain

Abstract

The classical optimal (in the Frobenius sense) diagonal preconditioner for large sparse linear systems $Ax = b$ is generalized and improved. The new proposed approximate inverse preconditioner N is based on the minimization of the Frobenius norm of the residual matrix $AM - I$, where M runs over a certain linear subspace of $n \times n$ real matrices, defined by a prescribed sparsity pattern. The number of nonzero entries of the $n \times n$ preconditioning matrix N is less than or equal to $2n$, and n of them are selected as the optimal positions in each of the n columns of matrix N. All theoretical results are justified in detail. In particular, the comparison between the proposed preconditioner N and the optimal diagonal one is theoretically analyzed. Finally, numerical experiments reported confirm the theory and illustrate that our generalization of the optimal diagonal preconditioner improves (in general) its efficiency, when they do not coincide.

Keywords: Approximate inverse preconditioner, Frobenius norm minimization, Diagonal preconditioner

1. Introduction

The discretization of many different PDEs (modeling physical problems) by any adequate numerical method (finite differences, finite elements, finite volumes, meshless, etc.), generally leads to a large linear system

[☆]Short running title: A generalization of the optimal diagonal preconditioner

*Corresponding author.

Email address: luisglez@dma.ulpgc.es (Luis González)

$$Ax = b, \quad A \in \mathbb{R}^{n\times n}, \;\; x, b \in \mathbb{R}^{n\times 1} \tag{1.1}$$

in which the matrix A is nonsingular and sparse.

The solution of these linear systems is usually performed by iterative methods based on Krylov subspaces (see, e.g., [1, 2, 3, 4]). To improve the convergence of these Krylov methods, system (1.1) can be preconditioned with an adequate preconditioning matrix N, transforming it into any of the equivalent problems

$$NAx = Nb, \tag{1.2}$$

$$ANy = b, \quad x = Ny, \tag{1.3}$$

that is, the left and right preconditioned systems, respectively. In this paper, we address only the case of right-sided preconditioners (1.3), but analogous results can be obtained for the left-sided preconditioners (1.2). The study of preconditioning strategies for large linear systems is at present one of the most relevant research areas in Numerical Linear Algebra. In [5], we can find a very complete survey about this question. The preconditioning of system (1.1) is performed in order to obtain a preconditioned matrix AN as close as possible to the identity in some sense, and the preconditioner N is called an approximate inverse of A.

The different strategies to construct approximate inverse preconditioners can be grouped into three categories [6]: approximate inverse methods based on Frobenius norm minimization, factorized sparse approximate inverses (see, e.g., [7, 8, 9] and the references therein), and preconditioning methods consisting of an incomplete factorization followed by an approximate inversion of the incomplete factors.

The idea of using Frobenius norm minimization for preconditioning purposes was first described in [10], and other early works can be found in [11, 12, 13]. Some posterior approaches in this sense can be found, for instance, in [14, 15, 16, 17, 18, 19] and in the references therein.

In some cases, the Frobenius norm based preconditioners are parametrized by prescribed sparsity patterns. Otherwise, among the Frobenius norm minimization preconditioners not extracted from sparse matrix subspaces, let us mention here the preconditioners for structured matrices obtained by orthogonal projections onto unitary matrix algebras (like, for instance, circulant preconditioners for Toeplitz matrices); see, e.g., [20, 21, 22] and the references therein.

In [23, 24], the search of Frobenius norm based approximate inverses with a prescribed sparsity pattern is generalized by considering a more general case of linear parametrization where preconditioners belong to an arbitrary matrix subspace $\mathcal{S}$ of $\mathbb{R}^{n\times n}$. This procedure leads to a natural generalization of the classical Moore-Penrose inverse, the so-called $\mathcal{S}$-Moore-Penrose inverse introduced in [25].

The closeness of the preconditioned matrix AN to the identity may be measured by using a suitable matrix norm like, for instance, the Frobenius norm $\|\cdot\|_F$. In this way, the problem of obtaining the best preconditioner N (with respect to the Frobenius norm) of system (1.1) in the subspace $\mathcal{S}$ of $\mathbb{R}^{n\times n}$ is reduced to the minimization problem

$$\min_{M\in\mathcal{S}} \|AM - I\|_F = \|AN - I\|_F \tag{1.4}$$

and the solution N to problem (1.4) will be referred to as the "optimal" preconditioner of system (1.1) over the subspace $\mathcal{S}$.

It is important to highlight that, throughout this paper, the term "optimal" means that the approximate inverse N is the matrix that minimizes the Frobenius norm on $AN - I$ over a certain subspace $\mathcal{S}$ of $\mathbb{R}^{n\times n}$, but the preconditioner N is not necessarily optimal in any other sense of the word.

Let us briefly describe the basic idea of this work. Our starting point is the well-known optimal diagonal preconditioner; see, e.g., [4]. This is exactly the solution $\overline{D}$ to problem (1.4) for the subspace of all $n\times n$ diagonal matrices, and it is often used as a simple preconditioner for sparse linear systems.

Sometimes, the preconditioner $\overline{D}$ is efficient and it leads to fast convergence. For instance, this is usually the case when matrix A is symmetric positive definite [2]. However, in other cases, the diagonal preconditioner $\overline{D}$ is not effective enough for convergence.

Then, we improve $\overline{D}$ in the following natural way. Since, obviously, the diagonal matrix $\overline{D}$ has one and only one nonzero element per column, this suggests the idea of considering the best approximate inverse (in the Frobenius sense) of matrix A among all the $n \times n$ matrices that have exactly one nonzero element per column. Call each of such nonzero elements the optimal row position or entry for its corresponding column. Then, our proposed preconditioner N will contain the n diagonal entries and those optimal entries per column which do not coincide with the diagonal ones. Finally, N will be exactly the solution to problem (1.4) for the subspace $\mathcal{S} \subset \mathbb{R}^{n\times n}$, defined by the above described sparsity pattern.

Obviously, the so defined approximate inverse N of matrix A generalizes $\overline{D}$, and it has at least n nonzero entries (the diagonal ones), and at most $2n$ nonzero entries.

Moreover, the preconditioning matrix N has another advantage, compared with the classical diagonal approximate inverse $\overline{D}$. Namely, the reiteration of the preconditioning technique with the optimal diagonal approximate inverse makes no sense. On the contrary, when using our new preconditioner N, the well-known multistep preconditioning strategy (see, e.g., [15, 26]) not only makes sense, but as we shall prove, each step of this reiterated preconditioning strategy strictly reduces the Frobenius norm of the residual matrix, whenever the preconditioner N obtained in the previous step is not diagonal.

We propose a simple, natural generalization of the optimal diagonal preconditioner, which improves it (in the sense of Eq. (1.4)). Theoretical results will be justified in detail and illustrated with some numerical experiments. In addition, the proposed preconditioner is also compared with the AINV approximate inverse preconditioner [6].

This paper has been organized as follows. In Section 2, we recall explicit expressions for both the solution N to problem (1.4) and its corresponding minimum Frobenius norm $\|AN - I\|_F$, valid for any matrix subspace $\mathcal{S} \subset \mathbb{R}^{n\times n}$. Next, in Section 3, we derive explicit expressions for the proposed preconditioner N and for $\|AN - I\|_F$. Numerical experiments are presented in Section 4. Finally, Section 5 closes the paper with some concluding remarks.

2. A preliminary lemma

Now, we present a preliminary lemma required to make this paper self-contained.

Taking advantage of the prehilbertian character of the matrix Frobenius norm, the solution N to problem (1.4) can be directly obtained using the orthogonal projection theorem. Here and in the following, orthogonality is with respect to the Frobenius inner product $\langle \cdot, \cdot \rangle_F$. More precisely, the matrix product AN is the orthogonal projection of the identity onto the subspace $A\mathcal{S}$. Consequently, an explicit formula for matrix N can be obtained by expressing the orthogonal projection AN of the identity matrix onto the subspace $A\mathcal{S}$ by its expansion with respect to an orthonormal basis of $A\mathcal{S}$ [23]. This is the idea of the following lemma.

Lemma 2.1. *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular. Let $\mathcal{S}$ be a linear subspace of $\mathbb{R}^{n\times n}$ of dimension d, and $\{M_1, ..., M_d\}$ a basis of $\mathcal{S}$ such that $\{AM_1, ..., AM_d\}$ is an orthogonal basis of $A\mathcal{S}$. Then, the solution to problem* (1.4) *is*

$$N = \sum_{i=1}^{d} \frac{tr\,(AM_i)}{\|AM_i\|_F^2} M_i, \tag{2.1}$$

and the minimum Frobenius norm is

$$\|AN - I\|_F^2 = n - \sum_{i=1}^{d} \frac{[tr\,(AM_i)]^2}{\|AM_i\|_F^2}. \tag{2.2}$$

Remark 2.1. If we have a basis $\{M_i\}_{i=1}^d$ of subspace $\mathcal{S}$ such that the corresponding basis $\{AM_i\}_{i=1}^d$ of subspace $A\mathcal{S}$ is not orthogonal, then we only need to use the Gram-Schmidt orthogonalization procedure to obtain an orthogonal basis of $A\mathcal{S}$, in order to apply Lemma 2.1. This procedure has been formalized in [23], obtaining several explicit expressions for both the optimal preconditioner N defined by (1.4) and $\|AN - I\|_F$, that have been applied to the sparse preconditioning of large linear systems arising from real-world cases.

For different spectral properties of matrix AN, and for the theoretical effectiveness analysis of the optimal approximate inverse preconditioners N defined by Eq. (1.4), we refer the reader to [24, 25].

3. The proposed approximate inverse preconditioner

In this section, the proposed preconditioner N of system (1.1) is introduced. First, we need to give a definition and to set some notations.

Comparing two different approximate inverses for the same matrix A, as stated by the following definition, is an essential point for preconditioning purposes.

Definition 3.1. *Let $A, N, N' \in \mathbb{R}^{n\times n}$ and suppose that A is nonsingular. Then, we say that N is better approximate inverse of A than N', or that N improves N' as approximate inverse of A if and only if*

$$\|AN - I\|_F < \|AN' - I\|_F .$$

Throughout this paper, the subspace of all $n \times n$ diagonal matrices is denoted by $\mathcal{D}_n$. From now on, $M_{i,j}$ denotes the $n \times n$ matrix whose only nonzero term is $m_{ij} = 1$, e_i denotes the ith column of the identity matrix (i.e., Ae_i is the ith column of A), and the symbols $\|\cdot\|_2$ and $\langle \cdot, \cdot \rangle_2$ stand for the usual Euclidean vector norm and inner product, respectively.

Remark 3.1. Note that since $M_{i,j} = e_i e_j^T$, then the only non-null column of matrix $AM_{i,j}$ is its jth one, which coincides with the ith column Ae_i of matrix A. Consequently, we have

$$tr\left(AM_{i,j}\right) = tr\left(Ae_i e_j^T\right) = a_{ji}, \qquad \|AM_{i,j}\|_F^2 = \left\|Ae_i e_j^T\right\|_F^2 = \|Ae_i\|_2^2 . \quad (3.1)$$

Moreover,

$$\langle AM_{i,j}, AM_{i',j} \rangle_F = \left\langle Ae_i e_j^T, Ae_{i'} e_j^T \right\rangle_F = \langle Ae_i, Ae_{i'} \rangle_2 , \quad (3.2)$$

$$\langle AM_{i,j}, AM_{i',j'} \rangle_F = \left\langle Ae_i e_j^T, Ae_{i'} e_{j'}^T \right\rangle_F = 0 \quad \text{for all } j \neq j', \quad (3.3)$$

so that any system of matrices $\{AM_{i,j}\}_{j=1}^n$ is orthogonal with respect to the Frobenius inner product.

When all diagonal entries of matrix A are not null, the preconditioner

$$diag\left(a_{11}^{-1}, a_{22}^{-1}, \ldots, a_{nn}^{-1}\right)$$

is often used as an approximate inverse preconditioner of system (1.1); see, e.g., [2]. However, in general, this is not the optimal choice (in the sense of Eq. (1.4)) among the diagonal approximate inverses. Indeed, as it is well-known, the best diagonal preconditioner $\overline{D}$ of system (1.1), that is, the solution to problem (1.4) for the subspace $\mathcal{D}_n$ is given by (see, e.g., [27])

$$\overline{D} = \sum_{j=1}^{n} \frac{a_{jj}}{\|Ae_j\|_2^2} M_{jj} = diag\left(\frac{a_{11}}{\|Ae_1\|_2^2}, \cdots, \frac{a_{nn}}{\|Ae_n\|_2^2} \right), \quad (3.4)$$

while the corresponding minimum Frobenius norm is given by

$$\left\|A\overline{D} - I\right\|_F^2 = n - \sum_{j=1}^{n} \frac{a_{jj}^2}{\|Ae_j\|_2^2}. \quad (3.5)$$

Obviously, the optimal diagonal approximate inverse (3.4) of matrix A, has exactly one nonzero element per column. As mentioned in Section 1,

this suggests the idea of considering the best approximate inverse of matrix A among all the $n \times n$ matrices that have exactly one nonzero element per column, the so-called optimal row position or entry per column. Suppose that the n nonzero optimal entries are placed at positions

$$(i_1, 1), (i_2, 2), \ldots, (i_n, n),$$

i.e., for each column $j = 1, 2, \ldots, n$, the optimal entry for preconditioning the linear system (1.1), by using Eq. (1.4), is placed at the i_jth row.

Then, our new preconditioner N is defined as follows.
(i) If $i_j = j$, the best entry in the jth column is the diagonal one. We select this entry, and no other entries are added to (j, j) in column j.
(ii) If $i_j \neq j$, the best entry in the jth column is not the diagonal one. We select the diagonal entry (j, j) and, besides, the optimal entry (i_j, j) is added to column j.
(iii) Finally, our preconditioner N is defined as the solution to problem (1.4) for the subspace $\mathcal{S}$ of $\mathbb{R}^{n\times n}$ whose only nonzero entries are the ones defined by steps (i) and (ii), i.e.,

$$\mathcal{S} = \mathcal{S}_n := span\left(\{M_{j,j}\}_{j=1}^n \cup \left\{M_{i_j,j} \mid i_j \neq j\right\}_{j=1}^n\right).$$

In this way, the number of nonzero entries of each column $j = 1, 2, \ldots, n$ of matrix N is either 1 (if $i_j = j$) or 2 (if $i_j \neq j$). Hence, the total number of nonzero entries of the $n \times n$ preconditioning matrix N is at least n and at most $2n$.

For instance, let $n = 4$ and suppose that for a certain coefficient matrix $A \in \mathbb{R}^{4\times 4}$, the optimal positions per column are

$$(i_1, 1) = (1, 1), \quad (i_2, 2) = (4, 2), \quad (i_3, 3) = (3, 3), \quad (i_4, 4) = (1, 4).$$

Then, the sparsity patterns of the preconditioning matrices $\overline{D}$ and N will be

$$\overline{D} = \begin{pmatrix} n_{11} & 0 & 0 & 0 \\ 0 & n_{22} & 0 & 0 \\ 0 & 0 & n_{33} & 0 \\ 0 & 0 & 0 & n_{44} \end{pmatrix}, \qquad N = \begin{pmatrix} n_{11} & 0 & 0 & n_{14} \\ 0 & n_{22} & 0 & 0 \\ 0 & 0 & n_{33} & 0 \\ 0 & n_{42} & 0 & n_{44} \end{pmatrix},$$

where n_{42} and n_{14} are the new (optimal) entries in N, not appearing in $\overline{D}$. Hence

$$\mathcal{S}_4 = span\left\{M_{1,1}, M_{2,2}, M_{3,3}, M_{4,4}, M_{4,2}, M_{1,4}\right\}.$$

Remark 3.2. Note that the preconditioning matrix N generalizes the optimal diagonal approximate inverse $\overline{D}$. Indeed, in the special case that the optimal entry for each column $j = 1, 2, \ldots, n$ is the diagonal one, we have

$$i_j = j \quad \text{for all } j = 1, 2, \ldots, n \quad \Rightarrow \quad N = \overline{D}.$$

Moreover, the preconditioner N improves, in general, the optimal diagonal preconditioner $\overline{D}$. Indeed, since $\overline{D}$ and N are the solutions to problem (1.4) for the subspaces

$$\mathcal{D}_n = span\left\{M_{j,j}\right\}_{j=1}^{n} \quad \text{and} \quad \mathcal{S}_n = span\left(\left\{M_{j,j}\right\}_{j=1}^{n} \cup \left\{M_{i_j,j} \mid i_j \neq j\right\}_{j=1}^{n}\right),$$

respectively, then we have

$$\mathcal{S}_n \supseteq \mathcal{D}_n \;\Rightarrow\; \|AN - I\|_F \leq \left\|A\overline{D} - I\right\|_F.$$

The following is the main result of this paper. It provides us with explicit expressions for both matrix N and the minimum Frobenius norm $\|AN - I\|_F$.

Theorem 3.1. *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular. Let N be the solution to problem* (1.4) *for the subspace*

$$\mathcal{S}_n = span\left(\left\{M_{j,j}\right\}_{j=1}^{n} \cup \left\{M_{i_j,j} \mid i_j \neq j\right\}_{j=1}^{n}\right). \tag{3.6}$$

Then, for each $j = 1, 2, \ldots, n$, its corresponding index i_j is defined by the condition

$$\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2} = \max\left\{\frac{|a_{j1}|}{\|Ae_1\|_2}, \frac{|a_{j2}|}{\|Ae_2\|_2}, \cdots, \frac{|a_{jn}|}{\|Ae_n\|_2}\right\}. \tag{3.7}$$

Moreover,

$$\begin{aligned} N = & \sum_{\substack{j=1\\ i_j = j}}^{n} \frac{a_{jj}}{\|Ae_j\|_2^2} M_{j,j} + \sum_{\substack{j=1\\ i_j \neq j}}^{n} \frac{a_{jj}\left\|Ae_{i_j}\right\|_2^2 - a_{ji_j}\left\langle Ae_j, Ae_{i_j}\right\rangle_2}{\|Ae_j\|_2^2\left\|Ae_{i_j}\right\|_2^2 - \left\langle Ae_j, Ae_{i_j}\right\rangle_2^2} M_{j,j} \\ & + \sum_{\substack{j=1\\ i_j \neq j}}^{n} \frac{a_{ji_j}\|Ae_j\|_2^2 - a_{jj}\left\langle Ae_j, Ae_{i_j}\right\rangle_2}{\|Ae_j\|_2^2\left\|Ae_{i_j}\right\|_2^2 - \left\langle Ae_j, Ae_{i_j}\right\rangle_2^2} M_{i_j,j} \end{aligned} \tag{3.8}$$

and the corresponding minimum Frobenius norm is given by

$$\|AN - I\|_F^2 = n - \sum_{j=1}^{n} \frac{a_{jj}^2}{\|Ae_j\|_2^2} - \sum_{\substack{j=1 \\ i_j \neq j}}^{n} \frac{\left[a_{ji_j} \|Ae_j\|_2^2 - a_{jj} \left\langle Ae_j, Ae_{i_j} \right\rangle_2\right]^2}{\|Ae_j\|_2^2 \left(\|Ae_j\|_2^2 \left\|Ae_{i_j}\right\|_2^2 - \left\langle Ae_j, Ae_{i_j} \right\rangle_2^2\right)}. \quad (3.9)$$

Proof. First, we determine the (optimal) positions $\{(i_j, j)\}_{j=1}^n$ of the nonzero entries in the best approximate inverse of matrix A among all $n \times n$ matrices that have exactly one nonzero element per column.

Let $j \in \{1, 2, \ldots, n\}$ be arbitrary, but fixed. The optimal approximate inverse $N_{i,j}$, among all the $n \times n$ matrices whose only nonzero term is placed at the ith row, jth column, can be obtained as the solution to problem (1.4) for the one-dimensional subspace $\mathcal{S} = span\{M_{i,j}\}$. That is, using Eqs. (2.1) and (3.1), we obtain

$$N_{i,j} = \frac{tr(AM_{i,j})}{\|AM_{i,j}\|_F^2} M_{i,j} = \frac{a_{ji}}{\|Ae_i\|_2^2} M_{i,j},$$

for which, using Eqs. (2.2) and (3.1), we have

$$\|AN_{i,j} - I\|_F^2 = n - \frac{[tr(AM_{i,j})]^2}{\|AM_{i,j}\|_F^2} = n - \frac{a_{ji}^2}{\|Ae_i\|_2^2}.$$

Consequently, the index $i \in \{1, 2, \ldots, n\}$ that minimizes $\|AN_{i,j} - I\|_F^2$ for each fixed column j, is the one that maximizes the quotient $\frac{a_{ji}^2}{\|Ae_i\|_2^2}$, that is, the index i_j, defined by Eq. (3.7).

Now, consider the set

$$T = \{j \in \{1, 2, \ldots, n\} \mid i_j \neq j\}.$$

There are two possible cases.

Case 1. If $i_j = j$ for all $j = 1, 2, \ldots, n$ then $T = \emptyset$. In this case,

$$\mathcal{S}_n = span\left(\{M_{j,j}\}_{j=1}^n \cup \left\{M_{i_j,j} \mid i_j \neq j\right\}_{j=1}^n\right) = span\{M_{j,j}\}_{j=1}^n = \mathcal{D}_n.$$

Hence, $\dim(\mathcal{S}_n) = \dim(\mathcal{D}_n) = n$ and, according to Remark 3.1, the basis $\{AM_{j,j}\}_{j=1}^n$ of subspace $A\mathcal{S}_n$ is orthogonal. Lemma 2.1 can be applied. Using Eqs. (2.1), (2.2) and (3.1), we obtain

$$N = \sum_{j=1}^{n} \frac{tr(AM_{j,j})}{\|AM_{j,j}\|_F^2} M_{j,j} = \sum_{j=1}^{n} \frac{a_{jj}}{\|Ae_j\|_2^2} M_{j,j}, \tag{3.10}$$

$$\|AN - I\|_F^2 = n - \sum_{j=1}^{n} \frac{[tr(AM_{j,j})]^2}{\|AM_{j,j}\|_F^2} = n - \sum_{j=1}^{n} \frac{a_{jj}^2}{\|Ae_j\|_2^2}. \tag{3.11}$$

But, since $T = \emptyset$, the two right-most sums in Eq. (3.8) and the right-most sum in Eq. (3.9) vanish, so that these two expressions exactly coincide with those given by Eqs. (3.10) and (3.11), respectively. This proves the theorem in case 1.

Case 2. If $i_j \neq j$ for some $j = 1, 2, \ldots, n$ then $T \neq \emptyset$. In this case, $1 \leq |T| \leq n$ and

$$\mathcal{S}_n = span\left(\{M_{j,j}\}_{j=1}^n \cup \left\{M_{i_j,j} \,|i_j \neq j\right\}_{j=1}^n\right) = span\left(\{M_{j,j}\}_{j=1}^n \cup \left\{M_{i_j,j}\right\}_{j\in T}\right).$$

Hence, $\dim(\mathcal{S}_n) = n + |T|$ and

$$B = \{AM_{j,j}\}_{j=1}^n \cup \left\{AM_{i_j,j}\right\}_{j\in T}$$

is a basis of subspace $A\mathcal{S}_n$.

Now, note that, according to Eq. (3.2), for every $j \in T$ we have

$$\left\langle AM_{j,j}, AM_{i_j,j}\right\rangle_F = \left\langle Ae_j, Ae_{i_j}\right\rangle_2$$

so that the basis B of $A\mathcal{S}_n$ is not necessarily orthogonal. Then, to apply Lemma 2.1, it suffices to use the Gram-Schmidt procedure (see Remark 2.1). In this way, after applying Gram-Schmidt to the basis B of $A\mathcal{S}_n$, we obtain the following orthogonal basis $\widetilde{B}$ of subspace $A\mathcal{S}_n$ from B

$$\widetilde{B} = \{AU_j\}_{j=1}^n \cup \{AV_j\}_{j\in T} = \{AM_{j,j}\}_{j=1}^n \cup \left\{AM_{i_j,j} - \beta_j AM_{j,j}\right\}_{j\in T},$$

where, for simplicity, we denote

$$U_j = M_{j,j} \quad \text{for all} \quad j = 1, 2, \ldots, n, \quad V_j = M_{i_j,j} - \beta_j M_{j,j} \quad \text{for all} \quad j \in T$$

and

$$\beta_j = \frac{\left\langle Ae_j, Ae_{i_j} \right\rangle_2}{\|Ae_j\|_2^2} \quad \text{for all} \;\; j \in T. \tag{3.12}$$

Hence, in order to apply Lemma 2.1 for the orthogonal basis $\widetilde{B}$ of $A\mathcal{S}_n$, we only need to compute the traces and Frobenius norms of matrices AU_j and AV_j; see Eq. (2.1). On one hand, for all $j = 1, 2, \ldots, n$ using Eq. (3.1), we immediately obtain

$$\begin{aligned} tr\left(AU_j\right) &= tr\left(AM_{j,j}\right) = a_{jj}, \\ \|AU_j\|_F^2 &= \|AM_{j,j}\|_F^2 = \|Ae_j\|_2^2 \end{aligned} \tag{3.13}$$

and, on the other hand, for all $j \in T$ using Eqs. (3.1), (3.2) and (3.12), we immediately obtain

$$\begin{aligned} tr\left(AV_j\right) &= tr\left(AM_{i_j,j} - \beta_j AM_{j,j}\right) = a_{ji_j} - \beta_j a_{jj}, \\ \|AV_j\|_F^2 &= \left\langle AM_{i_j,j} - \beta_j AM_{j,j}, AM_{i_j,j} - \beta_j AM_{j,j} \right\rangle_F \\ &= \left\|Ae_{i_j}\right\|_2^2 - \beta_j^2 \left\|Ae_j\right\|_2^2. \end{aligned} \tag{3.14}$$

Now, using Eq. (2.1) in Lemma 2.1 and Eqs. (3.13) and (3.14), we obtain the following first expression for matrix N (based on Gram-Schmidt)

$$\begin{aligned} N &= \sum_{j=1}^{n} \frac{tr\left(AU_j\right)}{\|AU_j\|_F^2} U_j + \sum_{j \in T} \frac{tr\left(AV_j\right)}{\|AV_j\|_F^2} V_j \\ &= \sum_{j=1}^{n} \frac{a_{jj}}{\|Ae_j\|_2^2} M_{j,j} + \sum_{j \in T} \frac{a_{ji_j} - \beta_j a_{jj}}{\left\|Ae_{i_j}\right\|_2^2 - \beta_j^2 \left\|Ae_j\right\|_2^2} \left(M_{i_j,j} - \beta_j M_{j,j}\right). \end{aligned}$$

Then, we split the above expression for N into three sums corresponding to linear combinations of the matrices: $M_{j,j}$ with $j \notin T$ (i.e., $i_j = j$); $M_{j,j}$ with $j \in T$ (i.e., $i_j \neq j$); and $M_{i_j,j}$ with $j \in T$ (i.e., $i_j \neq j$). Then, we replace β_j with its value given in Eq. (3.12). In this way, we obtain the following final expression for matrix N

$$\begin{aligned} N = &\sum_{j \notin T} \frac{a_{jj}}{\|Ae_j\|_2^2} M_{j,j} + \sum_{j \in T} \left(\frac{a_{jj}}{\|Ae_j\|_2^2} - \frac{\beta_j \left(a_{ji_j} - \beta_j a_{jj}\right)}{\left\|Ae_{i_j}\right\|_2^2 - \beta_j^2 \left\|Ae_j\right\|_2^2} \right) M_{j,j} \\ &+ \sum_{j \in T} \frac{a_{ji_j} - \beta_j a_{jj}}{\left\|Ae_{i_j}\right\|_2^2 - \beta_j^2 \left\|Ae_j\right\|_2^2} M_{i_j,j} \end{aligned}$$

$$= \sum_{\substack{j=1 \\ i_j = j}}^{n} \frac{a_{jj}}{\|Ae_j\|_2^2} M_{j,j} + \sum_{\substack{j=1 \\ i_j \neq j}}^{n} \frac{a_{jj} \left\|Ae_{i_j}\right\|_2^2 - a_{ji_j} \left\langle Ae_j, Ae_{i_j} \right\rangle_2}{\|Ae_j\|_2^2 \left\|Ae_{i_j}\right\|_2^2 - \left\langle Ae_j, Ae_{i_j} \right\rangle_2^2} M_{j,j}$$

$$+ \sum_{\substack{j=1 \\ i_j \neq j}}^{n} \frac{a_{ji_j} \|Ae_j\|_2^2 - a_{jj} \left\langle Ae_j, Ae_{i_j} \right\rangle_2}{\|Ae_j\|_2^2 \left\|Ae_{i_j}\right\|_2^2 - \left\langle Ae_j, Ae_{i_j} \right\rangle_2^2} M_{i_j,j}.$$

where the two right-most sums in the above expression contain the entries of the preconditioner N where it differs from the optimal diagonal preconditioner $\overline{D}$, while its left-most sum contains the diagonal entries (j,j) of N (such that $i_j = j$) whose values coincide with the ones placed at the same positions in $\overline{D}$; see Eq. (3.4).

Finally, using Eq. (2.2) in Lemma 2.1 and Eqs. (3.13), (3.14) and (3.12), we obtain

$$\begin{aligned}
\|AN - I\|_F^2 &= n - \sum_{j=1}^{n} \frac{\left[tr\left(AU_j\right)\right]^2}{\|AU_j\|_F^2} - \sum_{j \in T} \frac{\left[tr\left(AV_j\right)\right]^2}{\|AV_j\|_F^2} \\
&= n - \sum_{j=1}^{n} \frac{a_{jj}^2}{\|Ae_j\|_2^2} - \sum_{j \in T} \frac{\left(a_{ji_j} - \beta_j a_{jj}\right)^2}{\left\|Ae_{i_j}\right\|_2^2 - \beta_j^2 \|Ae_j\|_2^2} \\
&= n - \sum_{j=1}^{n} \frac{a_{jj}^2}{\|Ae_j\|_2^2} \\
&\quad - \sum_{\substack{j=1 \\ i_j \neq j}}^{n} \frac{\left[a_{ji_j} \|Ae_j\|_2^2 - a_{jj} \left\langle Ae_j, Ae_{i_j} \right\rangle_2\right]^2}{\|Ae_j\|_2^2 \left(\|Ae_j\|_2^2 \left\|Ae_{i_j}\right\|_2^2 - \left\langle Ae_j, Ae_{i_j} \right\rangle_2^2\right)}.
\end{aligned}$$

This proves the theorem in case 2. □

Remark 3.3. Formula (3.7) has the following geometric meaning. Since the cosine of the angle between the ith column of matrix A and the jth column of the identity is given by

$$\cos \angle \left(Ae_i, e_j\right) = \frac{\left\langle Ae_i, e_j \right\rangle_2}{\|Ae_i\|_2 \|e_j\|_2} = \frac{a_{ji}}{\|Ae_i\|_2}$$

then

$$\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2} = \max_{1\le i\le n} \frac{|a_{ji}|}{\|Ae_i\|_2} = \max_{1\le i\le n} |\cos\angle\left(Ae_i, e_j\right)| = \left|\cos\angle\left(Ae_{i_j}, e_j\right)\right|$$

and thus, Eq. (3.7) simply selects the column i_j of matrix A that maximizes $|\cos\angle\left(Ae_i, e_j\right)|$, i.e., it picks the column of A that is closest in angle to the jth column e_j of the identity ($a_{ji_j} > 0$), or to $-e_j$ ($a_{ji_j} < 0$).

Remark 3.4. Note that formulas (3.10) and (3.11) coincide with expressions (3.4) and (3.5), respectively, for the optimal diagonal preconditioner. Of course, this is due to the fact that, in case 1, we have

$$\mathcal{S}_n = span\left\{M_{j,j}\right\}_{j=1}^n = \mathcal{D}_n,$$

and thus $N = \overline{D}$.

Remark 3.5. The assumption that A is nonsingular in Theorem 3.1 is essential. Indeed, note that the factor (appearing in the denominators of Eqs. (3.8) and (3.9))

$$\|Ae_j\|_2^2 \left\|Ae_{i_j}\right\|_2^2 - \left\langle Ae_j, Ae_{i_j}\right\rangle_2^2 > 0$$

because of the Cauchy-Schwartz inequality and the assumption that A is nonsingular. In fact, the above positive factor is the Gram determinant of the jth and the i_jth columns of matrix A.

The following auxiliary lemma provides a lower bound on each summand of the right-most sum in Eq. (3.9), which will be used to compare the preconditioned matrix AN with both matrices $A\overline{D}$ (Corollary 3.1) and A itself (Corollary 3.2).

Lemma 3.1. *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular. For each $j = 1, 2, \ldots, n$, let i_j be its corresponding index defined by Eq. (3.7), and let $\theta_j = \angle\left(Ae_j, Ae_{i_j}\right)$. Then, for each j such that $i_j \neq j$, we have*

$$\frac{\left[a_{ji_j}\|Ae_j\|_2^2 - a_{jj}\left\langle Ae_j, Ae_{i_j}\right\rangle_2\right]^2}{\|Ae_j\|_2^2\left(\|Ae_j\|_2^2\left\|Ae_{i_j}\right\|_2^2 - \left\langle Ae_j, Ae_{i_j}\right\rangle_2^2\right)} \ge \left(\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2} - \frac{|a_{jj}|}{\|Ae_j\|_2}\right)^2 \frac{1}{\sin^2\theta_j}. \tag{3.15}$$

Proof. For each $j = 1, 2, \ldots, n$ such that $i_j \neq j$, we have

$$
\begin{aligned}
&\frac{\left[a_{ji_j}\left\|Ae_j\right\|_2^2 - a_{jj}\left\langle Ae_j, Ae_{i_j}\right\rangle_2\right]^2}{\left\|Ae_j\right\|_2^2\left(\left\|Ae_j\right\|_2^2\left\|Ae_{i_j}\right\|_2^2 - \left\langle Ae_j, Ae_{i_j}\right\rangle_2^2\right)} \\
=&\frac{a_{ji_j}^2\left\|Ae_j\right\|_2^4 - 2a_{jj}a_{ji_j}\left\|Ae_j\right\|_2^2\left\langle Ae_j, Ae_{i_j}\right\rangle_2 + a_{jj}^2\left\langle Ae_j, Ae_{i_j}\right\rangle_2^2}{\left\|Ae_j\right\|_2^2\left(\left\|Ae_j\right\|_2^2\left\|Ae_{i_j}\right\|_2^2 - \left\langle Ae_j, Ae_{i_j}\right\rangle_2^2\right)} \\
=&\frac{a_{ji_j}^2\left\|Ae_j\right\|_2^2 - 2a_{jj}a_{ji_j}\left\langle Ae_j, Ae_{i_j}\right\rangle_2 + \frac{a_{jj}^2}{\left\|Ae_j\right\|_2^2}\left\langle Ae_j, Ae_{i_j}\right\rangle_2^2}{\left\|Ae_j\right\|_2^2\left\|Ae_{i_j}\right\|_2^2 - \left\langle Ae_j, Ae_{i_j}\right\rangle_2^2} \\
=&\frac{\frac{a_{ji_j}^2}{\left\|Ae_{i_j}\right\|_2^2} - 2\frac{a_{jj}}{\left\|Ae_j\right\|_2}\frac{a_{ji_j}}{\left\|Ae_{i_j}\right\|_2}\cos\theta_j + \frac{a_{jj}^2}{\left\|Ae_j\right\|_2^2}\cos^2\theta_j}{1 - \cos^2\theta_j} \\
\geq&\frac{\frac{a_{ji_j}^2}{\left\|Ae_{i_j}\right\|_2^2} - 2\frac{|a_{jj}|}{\left\|Ae_j\right\|_2}\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2}\left|\cos\theta_j\right| + \frac{a_{jj}^2}{\left\|Ae_j\right\|_2^2}\cos^2\theta_j}{\sin^2\theta_j} \\
=&\frac{\frac{\left|a_{ji_j}\right|^2}{\left\|Ae_{i_j}\right\|_2^2} - 2\frac{|a_{jj}|}{\left\|Ae_j\right\|_2}\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2}\left|\cos\theta_j\right| + \frac{|a_{jj}|^2}{\left\|Ae_j\right\|_2^2}\cos^2\theta_j}{\sin^2\theta_j} \\
=&\frac{\left(\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2} - \frac{|a_{jj}|}{\left\|Ae_j\right\|_2}\left|\cos\theta_j\right|\right)^2}{\sin^2\theta_j} \geq \frac{\left(\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2} - \frac{|a_{jj}|}{\left\|Ae_j\right\|_2}\right)^2}{\sin^2\theta_j},
\end{aligned}
$$

since

$$
\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2} - \frac{|a_{jj}|}{\left\|Ae_j\right\|_2}\left|\cos\theta_j\right| \geq \frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2} - \frac{|a_{jj}|}{\left\|Ae_j\right\|_2} > 0
$$

because of Eq. (3.7). □

Next corollary, a key result in this paper, compares the minimum Frobenius norms for the optimal diagonal preconditioner $\overline{D}$ and for the optimal preconditioner N given by Theorem 3.1.

Corollary 3.1. *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular. For each $j = 1, 2, \ldots, n$, let i_j be its corresponding index defined by Eq. (3.7), and let $\theta_j = \angle\left(Ae_j, Ae_{i_j}\right)$. Let $\overline{D}$ be the optimal diagonal preconditioner for matrix A. Let N be the*

solution to problem (1.4) *for the subspace* $\mathcal{S}_n$ *defined by Eq.* (3.6). *Then*

$$\left\|A\overline{D} - I\right\|_F^2 - \|AN - I\|_F^2 \geq \sum_{\substack{j=1 \\ i_j \neq j}}^{n} \left(\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2} - \frac{|a_{jj}|}{\|Ae_j\|_2} \right)^2 \frac{1}{\sin^2\theta_j}. \quad (3.16)$$

Proof. Using Eqs. (3.5), (3.9) and (3.15) we see that N improves $\overline{D}$ by the quantity

$$\begin{aligned} \left\|A\overline{D} - I\right\|_F^2 - \|AN - I\|_F^2 &= \sum_{\substack{j=1 \\ i_j \neq j}}^{n} \frac{\left[a_{ji_j}\|Ae_j\|_2^2 - a_{jj}\left\langle Ae_j, Ae_{i_j}\right\rangle_2\right]^2}{\|Ae_j\|_2^2\left(\|Ae_j\|_2^2\left\|Ae_{i_j}\right\|_2^2 - \left\langle Ae_j, Ae_{i_j}\right\rangle_2^2\right)} \\ &\geq \sum_{\substack{j=1 \\ i_j \neq j}}^{n} \left(\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2} - \frac{|a_{jj}|}{\|Ae_j\|_2} \right)^2 \frac{1}{\sin^2\theta_j}. \quad \square \end{aligned}$$

Next corollary compares the Frobenius norms of the residual matrices $A - I$ and $AN - I$.

Corollary 3.2. *Let* $A \in \mathbb{R}^{n\times n}$ *be nonsingular. For each* $j = 1, 2, \ldots, n$, *let* i_j *be its corresponding index defined by Eq.* (3.7), *and let* $\theta_j = \angle\left(Ae_j, Ae_{i_j}\right)$. *Then, the solution* N *to problem* (1.4) *for the subspace* $\mathcal{S}_n$ *defined by Eq.* (3.6) *satisfies*

$$\begin{aligned} \|A - I\|_F^2 - \|AN - I\|_F^2 \geq &\sum_{j=1}^{n} \frac{\left(\|Ae_j\|_2^2 - a_{jj}\right)^2}{\|Ae_j\|_2^2} \\ &+ \sum_{\substack{j=1 \\ i_j \neq j}}^{n} \left(\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2} - \frac{|a_{jj}|}{\|Ae_j\|_2} \right)^2 \frac{1}{\sin^2\theta_j}. \quad (3.17) \end{aligned}$$

In particular, if $\|Ae_j\|_2^2 \neq a_{jj}$ *for at least one index* $j = 1, 2, \ldots, n$ *or if matrix* N *is not diagonal then we have*

$$\|A - I\|_F > \|AN - I\|_F. \quad (3.18)$$

Proof. Using the obvious fact that

$$\|A - I\|_F^2 = n - \left(2tr\,(A) - \|A\|_F^2\right)$$

and Eqs. (3.9) and (3.15), we obtain

$$\begin{aligned}
\|A - I\|_F^2 - \|AN - I\|_F^2 &= \|A\|_F^2 - 2tr\,(A) + \sum_{j=1}^{n} \frac{a_{jj}^2}{\|Ae_j\|_2^2} \\
&+ \sum_{\substack{j=1 \\ i_j \neq j}}^{n} \frac{\left[a_{ji_j}\|Ae_j\|_2^2 - a_{jj}\left\langle Ae_j, Ae_{i_j}\right\rangle_2\right]^2}{\|Ae_j\|_2^2\left(\|Ae_j\|_2^2\left\|Ae_{i_j}\right\|_2^2 - \left\langle Ae_j, Ae_{i_j}\right\rangle_2^2\right)} \\
&\geq \sum_{j=1}^{n}\|Ae_j\|_2^2 - \sum_{j=1}^{n} 2a_{jj} + \sum_{j=1}^{n}\frac{a_{jj}^2}{\|Ae_j\|_2^2} \\
&+ \sum_{\substack{j=1 \\ i_j \neq j}}^{n} \left(\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2} - \frac{|a_{jj}|}{\|Ae_j\|_2}\right)^2 \frac{1}{\sin^2\theta_j} \\
&= \sum_{j=1}^{n}\frac{\left(\|Ae_j\|_2^2 - a_{jj}\right)^2}{\|Ae_j\|_2^2} \\
&+ \sum_{\substack{j=1 \\ i_j \neq j}}^{n} \left(\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2} - \frac{|a_{jj}|}{\|Ae_j\|_2}\right)^2 \frac{1}{\sin^2\theta_j}.
\end{aligned}$$

Finally, if $\|Ae_j\|_2^2 \neq a_{jj}$ for at least one index $j = 1, 2, \ldots, n$ then the left sum in Eq. (3.17) contains at least one positive summand, and thus we conclude that $\|A - I\|_F > \|AN - I\|_F$. Moreover, if matrix N is not diagonal then $i_j \neq j$ for at least one column $j \in \{1, 2, \ldots, n\}$ (case 2 in the proof of Theorem 3.1). Hence, the right sum in Eq. (3.17) contains at least one summand, which is necessarily positive due to Eq. (3.7), and thus $\|A - I\|_F > \|AN - I\|_F$. $\square$

Remark 3.6. Corollary 3.1 has established the comparison between the optimal diagonal preconditioner $\overline{D}$ and the proposed preconditioner N, in the

following terms. First, note that from Eqs. (3.7) and (3.16), we conclude that N improves $\overline{D}$ in the sense of Definition 3.1, i.e.,

$$\text{If } N \neq \overline{D} \Rightarrow \ \exists\, j \in \{1, 2, \ldots, n\} \text{ s.t. } i_j \neq j \Rightarrow \left\|A\overline{D} - I\right\|_F > \|AN - I\|_F .$$

Second, Eq. (3.16) provides us with the following analysis of the improvement achieved when using preconditioner N instead of preconditioner $\overline{D}$. For the second item, we use the obvious fact that the function $f(\theta) = \frac{1}{\sin^2\theta}$ is strictly decreasing in the interval $\left(0, \frac{\pi}{2}\right)$, and strictly increasing in the interval $\left(\frac{\pi}{2}, \pi\right)$.
(i) For each $j = 1, 2, \ldots, n$ such that $i_j \neq j$, the more the maximum quotient $\frac{\left|a_{ji_j}\right|}{\left\|Ae_{i_j}\right\|_2}$ (given by Eq. (3.7)) exceeds the quotient $\frac{|a_{jj}|}{\|Ae_j\|_2}$, the larger the difference $\left\|A\overline{D} - I\right\|_F^2 - \|AN - I\|_F^2$ will be, and thus the more the preconditioner N improves the diagonal preconditioner $\overline{D}$ (in the sense of Definition 3.1).
(ii) For each $j = 1, 2, \ldots, n$ such that $i_j \neq j$, the closer the angle θ_j between the jth and the i_jth columns of the coefficient matrix A is either to 0 or to π (i.e., the larger the difference between θ_j and $\frac{\pi}{2}$ is), the larger $f(\theta_j)$, and then the larger the difference $\left\|A\overline{D} - I\right\|_F^2 - \|AN - I\|_F^2$ will be, and thus the more the preconditioner N improves the diagonal preconditioner $\overline{D}$ (in the sense of Definition 3.1).

Remark 3.7. Note that the right-most sum in Eq. (3.17) coincides with the sum in Eq. (3.16). Thus, the above two comments (i) and (ii) in Remark 3.6, concerning the comparison between $\left\|A\overline{D} - I\right\|_F$ and $\|AN - I\|_F$ (analyzed in Corollary 3.1), remain true for the comparison between $\|A - I\|_F$ and $\|AN - I\|_F$ (analyzed in Corollary 3.2).

Remark 3.8. Call $N_1 = N$ the approximate inverse of matrix A, constructed in Theorem 3.1. According to the well-known multistep preconditioning strategy (see, e.g., [15, 26]), we can obtain a sequence N_1, N_1N_2, $N_1N_2N_3, \ldots$ of approximate inverses of A where, for every $k \geq 2$, matrix N_k is the best sparse approximate inverse of matrix $AN_1N_2 \cdots N_{k-1}$, among all matrices defined by the sparsity pattern (3.6).

Note that since subspace $\mathcal{D}_n$ of all $n \times n$ diagonal matrices is closed for the matrix product, then we have

$$\begin{aligned}
&\min_{M \in \mathcal{D}_n} \|(AN_1)\, M - I\|_F = \|(AN_1)\, N_2 - I\|_F = \|A\,(N_1N_2) - I\|_F \\
&\min_{M \in \mathcal{D}_n} \|A\,(N_1M) - I\|_F = \min_{M \in \mathcal{D}_n} \|AM - I\|_F = \|AN_1 - I\|_F ,
\end{aligned}$$

and then, due to the uniqueness of solution of problem (1.4), we conclude that $N_1 N_2 = N_1$. This means that the multistep strategy does not make sense for the optimal diagonal preconditioner. However, this does not happen with the optimal preconditioners N belonging to the subspaces $\mathcal{S}_n$ defined by the prescribed sparsity patterns (3.6) and thus, in our case the multistep preconditioning strategy makes sense. In particular, Eq. (3.18) implies that each step k of our multistep preconditioning strategy strictly reduces the Frobenius norm, whenever the preconditioner N_{k-1}, obtained in the previous step, is not diagonal.

4. Numerical experiments

We present some numerical experiments to illustrate the behavior of the proposed preconditioner N. We compare the preconditioned linear system using our preconditioner N with both the unpreconditioned linear system and the preconditioned system using the optimal diagonal preconditioner $\overline{D}$. At the end of this section, the preconditioner N is also compared with the approximate inverse preconditioner AINV. We have studied a number of linear systems $Ax = b$, where the test coefficient matrices are taken from the University of Florida Sparse Matrix Collection [28]. We carried out our numerical problems with the Krylov solvers GMRES [29] and BiCGStab [30]. Both solvers led to similar results for most test matrices, but a small advantage was observed when using the latter for solving the systems preconditioned with matrix N. For this reason, we only present here the results obtained with the (right-preconditioned) BiCGStab. In any case, our purpose in this paper is to analyze the effectiveness of the proposed preconditioner (especially in comparison with the optimal diagonal approximate inverse), rather than to compare different Krylov subspace methods. The initial guess was always $x_0 = 0$, and the right-hand side vector was $b = [1, \ldots, 1]^T$. The stopping criterion was either

$$\frac{\|b - Ax_k\|_2}{\|b\|_2} < 10^{-8},$$

or when this condition about the relative residual was not satisfied, within $2n$ iterations (n being the order of the coefficient matrix A). We run all numerical experiments in double precision arithmetic, on *Intel*(*R*) *Xeon*(*R*) E5620 with 2.40 GHz clock frequency and 24GB of main memory using GNU Octave 3.2.4.

In Table 1, n and $nnz\,(A)$ stand for the order and the number of nonzero entries of matrix A, respectively. This table also provides the Frobenius norms of matrices $A - I$, $A\overline{D} - I$ and $AN - I$.

In Table 2, $nnz\,(N)$ denotes the number of nonzero entries of our preconditioner N, which is compared, in its second column, with the number n of entries of the optimal diagonal preconditioner $\overline{D}$ (obviously, $n \leq nnz\,(N) \leq 2n$). In the third and fourth columns, $\overline{D}$-time and N-time denote the CPU time (in seconds) for constructing the preconditioners $\overline{D}$ and N, respectively. In the three right-most columns, Unprec-iter, $\overline{D}$-iter, and N-iter stand for the number of iterations of the BiCGStab method for the unpreconditioned system, and the preconditioned systems with $\overline{D}$ and N, respectively. When convergence is not attained, within the maximum number $2n$ of allowed iterations, we indicate it by writing "†", in any of the corresponding columns Unprec-iter, $\overline{D}$-iter and N-iter.

Numerical tests reported confirm the theoretical results and illustrate the effectiveness of the proposed preconditioner in comparison with the optimal diagonal one.

Test problems have been grouped together into five classes, according to the behavior of the preconditioner N in comparison with $\overline{D}$. Looking at the two right-most columns ($\overline{D}$-iter and N-iter) in Table 2, one can easily identify each of these five groups of test matrices. For each of these classes, problems are arranged in increasing order of the size n of the test coefficient matrices.

The first five test problems correspond to matrices for which the optimal diagonal preconditioner and our preconditioner coincide, i.e., $N = \overline{D}$. Of course, this is in accordance with the theory, since N has been defined as a generalization of $\overline{D}$, and they do coincide when $i_j = j$ for all $j = 1, 2, \ldots, n$. Obviously, in such cases, Table 2 shows that $nnz\,(N) = n$ (each column of N consists only of its diagonal entry), and the number of iterations needed for convergence coincide for both preconditioners.

The rest of test matrices corresponds to the case $N \neq \overline{D}$, so that the number $nnz\,(N)$ of nonzero entries of N will be greater than n. When $nnz\,(N) = 2n$, this means that $i_j \neq j$ for all $j = 1, 2, \ldots, n$, and each column of N consists of two nonzero entries.

Table 1
The test matrices and the Frobenius norms of $A - I$, $A\overline{D} - I$ and $AN - I$.

Matrix	n	$nnz\,(A)$	$\|A - I\|_F$	$\left\|A\overline{D} - I\right\|_F$	$\|AN - I\|_F$
orsirr_2	886	5970	1.55×10^6	18	18
sherman1	1000	3750	1.00×10^3	15.6	15.6
sherman4	1104	3786	1.21×10^3	10.4	10.4
bcsstk09	1083	18437	8.57×10^8	21.2	21.2
sherman3	5005	20033	1.36×10^7	27.2	27.2
hor_131	434	4182	4.34×10^2	14.7	14.7
rdb450l	450	2580	5.01×10^2	16.2	13.7
pores_3	532	3474	6.63×10^5	15.1	14.9
steam2	600	5660	5.27×10^{10}	16.8	12.3
young3c	841	3988	6.40×10^3	14.8	14.7
bcsstk10	1086	22070	2.97×10^8	22.6	22.6
olm500	500	1996	2.24×10^5	18.3	15.6
olm1000	1000	3996	1.26×10^6	25.8	22.1
tols1090	1090	3546	1.23×10^7	16.8	14.6
fpga_trans_01	1220	7382	1.22×10^3	24.6	22.4
adder_dcop_14	1813	11246	1.81×10^3	24.3	23.9
adder_dcop_15	1813	11246	1.81×10^3	24.4	24
adder_dcop_16	1813	11246	1.81×10^3	24.1	23.7
adder_dcop_17	1813	11246	1.81×10^3	24	23.6
adder_dcop_20	1813	11246	1.81×10^3	24	23.5
adder_trans_02	1814	14579	1.81×10^3	14.6	14.1
tols2000	2000	5184	5.40×10^7	21.6	18.7
psmigr_1	3140	543160	3.54×10^6	15.4	15.4
tols4000	4000	8784	2.98×10^8	29.5	25.5
meg4	5860	25258	9.87×10^5	17.4	15.8
steam1	240	2248	8.41×10^7	8.95	8.92
sherman5	3312	20793	1.44×10^4	32.4	32.3
sherman2	1080	23094	7.00×10^9	30.7	28.9
adder_dcop_10	1813	11232	1.81×10^3	25.2	24.7

Table 2
Convergence results for $\overline{D}$ and N.

Matrix	$nnz(N)/n$	$\overline{D}$-time	N-time	Unprec-iter	$\overline{D}$-iter	N-iter
orsirr_2	886/886	0.072	0.360	1232	448	448
sherman1	1000/1000	0.084	0.368	407	192	192
sherman4	1104/1104	0.096	0.408	92	64	64
bcsstk09	1083/1083	0.124	0.832	188	154	154
sherman3	5005/5005	1.56	6	†	453	453
hor_131	435/434	0.028	0.140	†	369	255
rdb450l	900/450	0.024	0.380	†	254	51
pores_3	560/532	0.032	0.188	†	717	637
steam2	750/600	0.044	0.324	448	10	7
young3c	849/841	0.072	0.304	1080	974	804
bcsstk10	1088/1086	0.136	0.936	†	736	657
olm500	1000/500	0.032	0.428	†	†	307
olm1000	2000/1000	0.084	0.996	†	†	657
tols1090	1568/1090	0.088	0.688	†	†	1560
fpga_trans_01	1352/1220	0.128	0.684	†	†	205
adder_dcop_14	1834/1813	0.248	1.224	†	†	437
adder_dcop_15	1836/1813	0.248	1.220	†	†	310
adder_dcop_16	1837/1813	0.248	1.224	†	†	418
adder_dcop_17	1843/1813	0.248	1.228	†	†	329
adder_dcop_20	1857/1813	0.248	1.248	†	†	518
adder_trans_02	1832/1814	0.260	1.384	229	†	9
tols2000	2842/2000	0.256	1.560	†	†	1534
psmigr_1	3143/3140	4.348	46.951	1812	†	58
tols4000	5642/4000	0.896	4.492	†	†	1328
meg4	5968/5860	2.104	8.584	†	†	5
steam1	320/240	0.012	0.108	†	236	388
sherman5	3327/3312	0.756	3.53	2251	849	892
sherman2	1628/1080	0.144	1.38	†	†	†
adder_dcop_10	1872/1813	0.248	1.256	†	†	†

In some of these cases (e.g., for the test matrices displayed in rows 6-11), the number of iterations needed to reach convergence is reduced when using our preconditioner N instead of $\overline{D}$.

Moreover, in many cases, like for instance for the test matrices displayed in rows 12-25 in Table 2, convergence is not attained with the diagonal pre-

conditioner $\overline{D}$, but however it is reached when we use our preconditioner N. We think that this is the main advantage of our preconditioner pointed out by the numerical tests.

In particular, we highlight the case of the test matrix *psmigr_1*. In this case, although the preconditioner N has only 3 nonzero entries more than the diagonal preconditioner $\overline{D}$ ($nnz(N)/n = 3143/3140$), the system transits from non-convergence to convergence if we use N (with a small number of iterations) instead of $\overline{D}$ as preconditioning matrix. Moreover, for this test problem, both preconditioners only differ in the numerical values of 6 entries: the $nnz(N) - n = 3$ diagonal entries (j, j) for which $i_j \neq j$, and the corresponding 3 nondiagonal entries (i_j, j) that do not appear in the diagonal matrix $\overline{D}$. The numerical values of the remaining $n - 3 = 3137$ nonzero diagonal entries of $\overline{D}$ and N coincide (see Eqs. (3.4) and (3.8)). So, $\overline{D}$ and N are very similar but, as mentioned, we obtain convergence with the latter, but not with the former. This happens not only for the test problem *psmigr_1*, but also for other test matrices reported.

Next, we present two problems, namely the third and fourth ones from the bottom, for which the number of iterations required for reaching convergence is greater when we use the preconditioner N instead of $\overline{D}$. Finally, for the last two test matrices reported, convergence is reached neither with N nor with $\overline{D}$. This can be explained by the small number of nonzero entries in both preconditioners, which makes them inefficient for some very ill-conditioned systems.

For all test problems reported, the last three columns in Table 1 confirm our theoretical results, in the sense that

$$\|A - I\|_F \geq \left\|A\overline{D} - I\right\|_F \geq \|AN - I\|_F .$$

On one hand, the first inequality is due to the fact that the optimal diagonal preconditioner $\overline{D}$ (by definition) minimizes $\|AM - I\|_F$ over the subspace of all $n \times n$ diagonal matrices (and the identity is a diagonal matrix). On the other hand, the second inequality is an obvious consequence of the set inclusion $\mathcal{S}_n \supseteq \mathcal{D}_n$ (as commented in Remark 3.2).

The large difference (observed for all test matrices reported) between $\|A - I\|_F$ and any of the values $\left\|A\overline{D} - I\right\|_F$ and $\|AN - I\|_F$ is mainly due to the following fact. While, obviously, $\|A - I\|_F$ can be arbitrarily large, both norms $\left\|A\overline{D} - I\right\|_F$ and $\|AN - I\|_F$ are never greater than $\sqrt{n}$. The reason is that, from Eq. (2.2) we immediately derive that the optimal approximate

inverse M (in the Frobenius sense) over any matrix subspace $\mathcal{S} \subset \mathbb{R}^{n\times n}$ (and, in particular, $M = \overline{D}$ and $M = N$) always satisfies $\|AM - I\|_F \leq \sqrt{n}$. Also, we can observe that the test matrices for which the difference between $\left\|A\overline{D} - I\right\|_F$ and $\|AN - I\|_F$ is larger usually correspond to those cases for which the ratio between the numbers of nonzero entries in matrices N and $\overline{D}$ (denoted by $nnz\,(N)\,/n$ in Table 2) is larger. In addition to this ratio, other parameters that also determine the difference between (the squares of) the Frobenius norms of $A\overline{D} - I$ and $AN - I$ have been analyzed in Remark 3.6.

Summing up, in almost all the cases, the iterations required by the BiCGStab method when using preconditioner N are fewer than those needed by this solver when using the diagonal preconditioner $\overline{D}$. The differences between the CPU times for constructing the preconditioners $\overline{D}$ and N, as well as the differences between the execution times of the BiCGStab method for both preconditioners, are not significant, and in most cases the total CPU times for solving the system with $\overline{D}$ and N are of the same order of magnitude. In any case, the small increment in computational cost when using the preconditioner N instead of $\overline{D}$, is compensated by the fact that, in many cases, convergence is reached with the preconditioner N but not with the optimal diagonal preconditioner $\overline{D}$.

For both preconditioners $\overline{D}$ and N, their respective sparsity patterns consist of small numbers of nonzero entries (n entries for $\overline{D}$, and a number of entries between n and $2n$ for N). On one hand, this implies low computational costs for constructing them. On the other hand, for both of them, the number of required iterations is large in comparison with other more dense and expensive optimal approximate inverse preconditioners based on the same idea (Frobenius norm minimization). In any case, as the numerical experiments have shown, the proposed preconditioner improves the classical diagonal one, with a small increment in the computation cost required for constructing the former instead of the latter.

To finish this section, we compare the proposed preconditioner N with a more expensive approximate inverse preconditioner, namely the well-known AINV preconditioner [6]. We have implemented the AINV preconditioner, with a drop tolerance $Tol = 0.25$, for the same set of test matrices used for comparing our preconditioner N with the optimal diagonal one $\overline{D}$. For most of these test problems, AINV was found to be more efficient (in terms of the overall solution time and using the Krylov solver BiCGStab) than N. However, for some test matrices, namely *sherman4*, *steam2*, *adder_trans_02*,

psmigr_1 and *meg4*, the proposed preconditioner was found to be more efficient than AINV. In conclusion, our proposed preconditioner was more effective than the optimal diagonal $\overline{D}$ for most of the numerical problems considered in this paper, and it was more effective than the AINV preconditioner in a few cases.

5. Summary and conclusions

In this paper, a new approximate inverse preconditioner N for large sparse linear systems has been constructed and theoretically analyzed. N has been defined as the optimal preconditioner (in the Frobenius sense) among all the $n \times n$ matrices whose only nonzero entries, for each column $j = 1, 2, \ldots, n$, are the diagonal one (j, j) and, in addition, the optimal entry (i_j, j) in column j, whenever it does not coincide with the diagonal one. In this way, our preconditioning matrix N generalizes the optimal diagonal preconditioner $\overline{D}$. Explicit expressions for both matrix N and the minimum Frobenius norm $\|AN - I\|_F$ have been presented. We have proved that, whenever $N \neq \overline{D}$, the preconditioner N (which has at least n and at most $2n$ nonzero entries) improves $\overline{D}$, in the sense of the Frobenius norm. We have also analyzed the difference between the Frobenius norms of $A - I$ and $AN - I$.

Numerical experiments have confirmed the theoretical results, presenting a number of test matrices for which the proposed preconditioner improves the convergence of the optimal diagonal one, when they do not coincide. In particular, Table 1 shows that the Frobenius norm $\|AN - I\|_F$ is always smaller (as we have theoretically shown) and, in fact, much smaller in most cases than the Frobenius norm $\|A - I\|_F$. Table 2 illustrates the reduction, in most cases, of the number of iterations when we use the preconditioning matrix N instead of $\overline{D}$.

The main advantage of our preconditioner pointed out by the numerical tests is the following. For many test matrices, the small additional CPU time required for constructing N instead of $\overline{D}$ is compensated by the fact that system transits from non-convergence to convergence when using N instead of $\overline{D}$ as preconditioning matrix. Moreover, this also occurs for some test problems for which the number of nonzero entries of the preconditioner N exceeds only by a very small quantity the number of nonzero entries of the diagonal preconditioner $\overline{D}$, and the numerical values of most of the entries placed at the same diagonal positions coincide for both preconditioners ($\overline{D}$ and N are very similar).

For future researches, it would be interesting to analyze in more detail the practical value of the preconditioner proposed in this paper. For this purpose, our preconditioner N could be compared, using new test matrices, with the AINV approximate inverse preconditioner and with other preconditioners not considered in this paper. In addition, it is worth trying to study some common characteristics/features of those test matrices for which the generalized preconditioner N has a better behavior for convergence purposes.

Finally, regarding an additional line for future work, our method can be improved by considering the optimal preconditioner N (in the sense of the Frobenius norm) among all the $n \times n$ matrices having exactly a (small) fixed number $m_j \geq 2$ of nonzero entries for each column $j = 1, 2, \ldots, n$.

The determination of the sparsity pattern of such preconditioner N is not a simple problem because of the following fact. Assume that the jth column Ne_j of the preconditioner N consists of only one nonzero entry, i.e., $m_j = 1$. Then, as shown in Section 3, the optimal position $\left(i_j^1, j\right)$ in the jth column Ne_j of N for minimizing $\|ANe_j - e_j\|_2$ can be easily determined simply by using Eq. (3.7). Similarly, we can easily determine the second, third,...,m_jth best positions $\left(i_j^2, j\right)$, $\left(i_j^3, j\right)$,...,$\left(i_j^{m_j}, j\right)$ in the jth column of N for minimizing $\|ANe_j - e_j\|_2$. Unfortunately, the set of these best m_j positions (obtained separately, i.e., when Ne_j consists of only one nonzero entry) for approximating the jth column e_j of I, does not necessarily coincide with the optimal set of cardinality m_j for approximating e_j (when the sparsity pattern of N is defined by the condition that Ne_j consists of $m_j \geq 2$ nonzero entries). Consequently, the determination of this optimal m_j-set (in order to find the optimal sparsity pattern for an optimal preconditioner defined by Eq. (1.4)) is a very difficult problem (when $m_j \geq 2$). Alternatively, one can consider the possibility of using an algorithm based on the LU factorization with partial pivoting to determine the optimal sparsity pattern for each column.

Acknowledgments

The authors thank the anonymous referees for their detailed revisions and valuable comments and suggestions, which have substantially improved the earlier version of this paper. This work was partially supported by the "Ministerio de Economía y Competitividad" of the Spanish Government, and FEDER, through Grant contract: CGL2011-29396-C03-01.

References

[1] O. Axelsson, Iterative Solution Methods, Cambridge University Press, Cambridge, MA, 1994.

[2] A. Greenbaum, Iterative Methods for Solving Linear Systems, Frontiers Appl. Math., vol. 17, SIAM, Philadelphia, PA, 1997.

[3] C.T. Kelley, Iterative Methods for Linear and Nonlinear Equations, Frontiers Appl. Math., vol. 16, SIAM, Philadelphia, PA, 1995.

[4] Y. Saad, Iterative Methods for Sparse Linear Systems, PWS Publishing Co., Boston, MA, 1996.

[5] M. Benzi, Preconditioning techniques for large linear systems: a survey, J. Comput. Phys. 182 (2002) 418-477.

[6] M. Benzi, M. Tůma, A comparative study of sparse approximate inverse preconditioners, Appl. Numer. Math. 30 (1999) 305-340.

[7] M. Benzi, C.D. Meyer, M. Tůma, A sparse approximate inverse preconditioner for the conjugate gradient method, SIAM J. Sci. Comput. 17 (1996) 1135-1149.

[8] M. Benzi, J.K. Cullum, M. Tůma, Robust approximate inverse preconditioning for the conjugate gradient method, SIAM J. Sci. Comput. 22 (2000) 1318-1332.

[9] E.-J. Lee, J. Zhang, Factored approximate inverse preconditioners with dynamic sparsity patterns, Comput. Math. Appl. 62 (2011) 235-242.

[10] M.W. Benson, Iterative solution of large scale linear systems, Masters thesis, Lakehead University, Thunder Bay, Canada, 1973.

[11] P.O. Frederickson, Fast approximate inversion of large sparse linear systems, Math. Report 7, Lakehead University, Thunder Bay, Canada, 1975.

[12] M.W. Benson, P.O. Frederickson, Iterative solution of large sparse linear systems arising in certain multidimensional approximation problems, Util. Math. 22 (1982) 127-140.

[13] M.W. Benson, J. Krettmann, M. Wright, Parallel algorithms for the solution of certain large sparse linear systems, Int. J. Comput. Math. 16 (1984) 245-260.

[14] B. Carpentieri, I.S. Duff, L. Giraud, Sparse pattern selection strategies for robust Frobenius-norm minimization preconditioners in electromagnetism, Numer. Linear Algebra Appl. 7 (2000) 667-685.

[15] E. Chow, Y. Saad, Approximate inverse preconditioners via sparse-sparse iterations, SIAM J. Sci. Comput. 19 (1998) 995-1023.

[16] N.I.M. Gould, J.A. Scott, Sparse approximate-inverse preconditioners using norm-minimization techniques, SIAM J. Sci. Comput. 19 (1998) 605-625.

[17] M. Grote, T. Huckle, Parallel preconditioning with sparse approximate inverses, SIAM J. Sci. Comput. 18 (1997) 838-853.

[18] H.L. Ong, Fast approximate solution of large-scale sparse linear systems, J. Comput. Appl. Math. 10 (1984) 45-54.

[19] V. del Olmo, R. Fuster, Some iterative methods related to Frobenius norm minimization, Comput. Math. Appl. 22 (1991) 121-126.

[20] D.A. Bini, P. Favati, O. Menchi, A family of modified regularizing circulant preconditioners for two-levels Toeplitz systems, Comput. Math. Appl. 48 (2004) 755-768.

[21] R.H. Chan, M.K. Ng, Conjugate gradient methods for Toeplitz systems, SIAM Rev. 38 (1996) 427-482.

[22] E. Tyrtyshnikov, Optimal and superoptimal circulant preconditioners, SIAM J. Matrix Anal. Appl. 13 (1992) 459-473.

[23] G. Montero, L. González, E. Flórez, M.D. García, A. Suárez, Approximate inverse computation using Frobenius inner product, Numer. Linear Algebra Appl. 9 (2002) 239-247.

[24] L. González, Orthogonal projections of the identity: spectral analysis and applications to approximate inverse preconditioning, SIAM Rev. 48 (2006) 66-75.

[25] A. Suárez, L. González, A generalization of the Moore-Penrose inverse related to matrix subspaces of $\mathbb{C}^{n\times m}$, Appl. Math. Comput. 216 (2010) 514-522.

[26] K. Wang, J. Zhang, MSP: A class of parallel multistep successive sparse approximate inverse preconditioning strategies, SIAM J. Sci. Comput. 24 (2003) 1141-1156.

[27] P. Tarazaga, D. Cuellar, Preconditioners generated by minimizing norms, Comput. Math. Appl. 57 (2009) 1305-1312.

[28] T.A. Davis, Y. Hu, The University of Florida Sparse Matrix Collection. http://www.cise.ufl.edu/research/sparse/matrices (accessed June 2013).

[29] Y. Saad, M.H. Schultz, GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems, SIAM J. Sci. Statist. Comput. 7 (1986) 856-869.

[30] H.A. van der Vorst, BICGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems, SIAM J. Sci. Statist. Comput. 13 (1992) 631-644.