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

Rounding error analysis of orthogonalization with a non-standard inner product ⋆

Jiří Kopal · Miroslav Rozložník · Miroslav Tůma · Alicja Smoktunowicz

Received: date / Accepted: date

Abstract In this paper we study the numerical properties of several orthogonalization schemes where the inner product is induced by a nontrivial symmetric and positive definite matrix. We analyze the effect of its conditioning on the factorization and the loss of orthogonality between vectors computed in finite precision arithmetic. We consider the implementation based on the backward stable eigendecomposition, modified and classical Gram-Schmidt algorithms, Gram-Schmidt process with reorthogonalization as well as the implementation motivated by the AINV approximate inverse preconditioner.

Keywords Orthogonalization schemes · QR factorization · Gram-Schmidt process · Preconditioning · Rounding error analysis

The work of J. Kopal was supported by the Ministry of Education of the Czech Republic under the project no. 7822/115. The work of M. Rozložník and M. Tůma was supported by Grant Agency of the Academy of Sciences of the Czech Republic under the project IAA100300802, by the international collaboration support M100300902 of AS CR and by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications".

Jiří Kopal
Technical University of Liberec, Institute of New Technologies and Applied Mathematics, Hálkova 6, CZ-461 17 Liberec, Czech Republic
E-mail: jiri.kopal@tul.cz.

Miroslav Rozložník
Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, CZ-182 07 Prague 8, Czech Republic
E-mail: miro@cs.cas.cz.

Miroslav Tůma
Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, CZ-182 07 Prague 8, Czech Republic
E-mail: tuma@cs.cas.cz.

Alicja Smoktunowicz
Faculty of Mathematics and Information Science, Warsaw University of Technology, Pl. Politechniki 1, 00-661, Warsaw, Poland
E-mail: A.Smoktunowicz@mini.pw.edu.pl.

1 Introduction

Let A be a given $m \times m$ symmetric positive definite matrix and $Z^{(0)}$ be an $m \times n$ matrix of full column rank n, $m \geq n$. We want to compute matrices Z and U such that $Z^{(0)} = ZU$, where Z is a $m \times n$ matrix satisfying $Z^T AZ = I$ and U is an upper triangular $n \times n$ matrix with positive diagonal entries. It is clear that the matrix U can be seen as the Cholesky factor of the matrix $(Z^{(0)})^T AZ^{(0)} = U^T U$ with the norm and minimum singular value bounded as

$$\sigma_m^{1/2}(A)\sigma_n(Z^{(0)}) \leq \sigma_n(A^{1/2}Z^{(0)}) = \sigma_n(U) \leq \|U\| = \|A^{1/2}Z^{(0)}\| \leq \|A\|^{1/2}\|Z^{(0)}\| \tag{1}$$

and the condition number $\kappa(U)$ satisfying $\kappa(U) = \kappa(A^{1/2}Z^{(0)}) \leq \kappa^{1/2}(A)\kappa(Z^{(0)})$. It is also easy to see that $U = Z^T AZU = Z^T AZ^{(0)}$. Due to the orthogonality relation $(A^{1/2}Z)^T(A^{1/2}Z) = I$ we have for the extremal singular values of Z

$$\|A\|^{-1/2} = \sigma_m(A^{-1/2}) \leq \sigma_n(Z) \leq \|Z\| \leq \|A^{-1/2}\| = \|A^{-1}\|^{1/2}. \tag{2}$$

Indeed, it follows from (2) that $\kappa(Z) \leq \kappa^{1/2}(A)$. Since $Z = Z^{(0)}U^{-1}$ the product ZZ^T can be written as $ZZ^T = Z^{(0)}\left[(Z^{(0)})^T AZ^{(0)}\right]^{-1}(Z^{(0)})^T$. Then AZZ^T represents the oblique projector onto $R(AZ^{(0)})$ and orthogonal to $R(Z^{(0)})$. Similarly, $ZZ^T A$ is the oblique projector onto $R(Z^{(0)})$ and orthogonal to $R(AZ^{(0)})$.

For $m = n$ and $Z^{(0)}$ square nonsingular we have $ZZ^T = A^{-1}$ and $AZZ^T = ZZ^T A = I$. If the matrix $Z^{(0)}$ is in addition upper triangular then the matrix Z is also upper triangular and it represents an inverse factor in the triangular factorization $A^{-1} = ZZ^T$. In the particular case with $Z^{(0)} = I$ the matrix U is a Cholesky factor of A and $Z = U^{-1}$. This fact is heavily used in many applications. One of the important preconditioning classes involves computing such an approximate inverse factorization [4]. Another well-known examples are the symmetric definite generalized eigenvalue problem which can be using such factors transformed into the standard eigenproblem in well-conditioned case [32], [23], [13], nonsymmetric eigenvalue problem [26], and the generalized least squares problems [7] including the weighted least squares problem [22], [16] as a particular case. Sparse implementations of generalized orthogonalization schemes are efficiently used in linear scaling electronic theory [8], information retrieval [33] or solving complex systems from quantum chemistry or systems arising from Helmholtz's equation [24].

Given the matrices A and $Z^{(0)}$, there are numerous ways how to compute the factors Z and U. If we have the spectral decomposition $A = V\Lambda V^T$, the factor U can be obtained from the standard QR decomposition $\Lambda^{1/2}V^T Z^{(0)} = QU$ and the factor Z can be then recovered as $Z = V\Lambda^{-1/2}Q$. Similarly, if we have the Cholesky decomposition $A = LL^T$, then U is the upper triangular factor from $L^T Z^{(0)} = QU$ and Z can be then computed as $Z = L^{-T}Q$. Significantly less attention has been paid to the QR decomposition using the A-invariant reflections - only the case of weighted QR factorization has been thoroughly analyzed in [14]. One of the most frequently used and probably the most straightforward approach is the Gram-Schmidt orthogonalization, which consecutively A-orthogonalizes the columns of $Z^{(0)}$ against previously computed vectors from factor Z using the orthogonalization coefficients that form then the triangular factor U. In the classical Gram-Schmidt algorithm (CGS), the A-orthogonal vectors are computed via matrix-vector updates which are relatively easy to parallelize. The rearrangement of this scheme has led to the modified Gram-Schmidt

algorithm (MGS) with better numerical properties. However, introducing sequential orthogonalization of the current vector destroys desirable parallel properties of the algorithm. We will discuss also yet another variant of sequential orthogonalization, which is motivated originally by the AINV preconditioner [4] and which uses oblique projections. We will refer to this scheme as the AINV orthogonalization (AINV) here. For the particular case $Z^{(0)} = I$ the situation is even more developed due to progress of recent preconditioning techniques. The early papers on inverse factorization have various motivations and do not study numerical properties of algorithms [25,10,19,20, 18,29,30]. Although the main motivation for the development of approximate inverse techniques came from parallel processing, concerns on their robustness and accuracy immediately became an important aspect. While the initial schemes like the basic AINV algorithm [4] were based on oblique projections or the CGS orthogonalization, recent development has lead to their stabilization both in terms of the orthogonalization scheme (MGS in the SAINV algorithm [3]) and in terms of appropriate computation of diagonal entries in U (one-sided versus stabilized versions of AINV [3,5,21]).

While for the case of standard inner product there exist complete rounding error analysis for all main schemes [17], [6], [11], [12], [31] numerical properties of orthogonalization schemes with non-standard inner product are much less understood. The main motivation of this paper is to review several orthogonalization approaches and to give bounds for corresponding quantities computed in finite precision arithmetic. Given some approximations $\bar{Z}$ and $\bar{U}$ to Z and U, respectively, we will be especially interested at the magnitude of quantities as the factorization error $Z^{(0)} - \bar{Z}\bar{U}$, the error in computing the Cholesky factor $(Z^{(0)})^T AZ^{(0)} - \bar{U}^T\bar{U}$ and the most important loss of orthogonality between computed vectors measured by $\bar{Z}^T A\bar{Z} - I$. Eventually we will look at the error in the approximation of the inverse $A^{-1} - \bar{Z}\bar{Z}^T$ and/or the right or left residual $A\bar{Z}\bar{Z}^T - I$ or $\bar{Z}\bar{Z}^T A - I$. We will formulate them mainly in terms of quantities proportional to the roundoff unit u, in terms of the condition number $\kappa(A)$ which represents an upper bound for the relative error in computing the A-inner product as well as the condition number of the matrix $A^{1/2}Z^{(0)}$ which plays an important role in the factorization $(Z^{(0)})^T AZ^{(0)} \approx \bar{U}^T\bar{U}$. We believe that these results are an initial step towards understanding the behavior of practical strategies in approximate inverse preconditioning which are based on sparse approximation to the factors Z and U using some inexact orthogonalization scheme. For a survey of such preconditioning techniques we refer to [2]. The organization of the paper is as follows. Section 2 is devoted to the ideal implementation based on the eigenvalue decomposition of A. Section 3 recalls the modified Gram-Schmidt algorithm with the inner product induced by the matrix A. In Section 4 we consider two orthogonalization schemes with this inner product, namely the classical Gram-Schmidt and AINV orthogonalizations and show that they behave in a similar way. Finally, in Section 5 we focus on the roundoff analysis of the Gram-Schmidt algorithm with reorthogonalization and show that it is numerically similar to the ideal implementation discussed in Section 2.

Throughout the paper $X = [x_1, \ldots, x_n]$ denotes the $m \times n$ matrix X with columns $x_1, \ldots, x_n$. The quantity $\sigma_k(X)$ denotes its kth largest singular value and if X has a full column rank then $\kappa(X) = \sigma_1(X)/\sigma_n(X)$ refers to the condition number of the matrix X. The term $|X|$ denotes the absolute value of the matrix X; $\|X\| = \sigma_1(X)$ denotes its 2-norm ; $|x|$ is the absolute value of the vector x and $\|x\|$ denotes its Euclidean norm. By $\langle \cdot, \cdot \rangle$ we mean the Euclidean inner product of two vectors and $\langle \cdot, \cdot \rangle_A$ denotes the inner product defined by the positive definite matrix A.

For distinction with their exact arithmetic counterparts, we denote quantities computed in finite precision arithmetic using an extra upper-bar. We assume the standard model for floating-point computations, and use the notation $fl(\cdot)$ for the computed result of some expression (see e.g. [17]). The unit roundoff is denoted by u. The terms $O(u)$, $k = 1, 2, \ldots$ are low-degree polynomials in the problem dimensions m and n multiplied by the unit roundoff u; they are independent of the condition number $\kappa(A)$ but they do depend on details of the computer arithmetic. For simplicity we do not evaluate the terms the terms proportional to higher powers of u and also occasionally skip the technical details that would negatively affect the presentation of our results.

2 The implementation based on eigendecomposition

The eigendecomposition of the (symmetric positive definite) matrix $A = V\Lambda V^T$ can find its use also in our orthogonalization problem. Indeed, the factor Z can be computed as a product of two orthogonal and one diagonal matrix in the form $Z = V\Lambda^{-1/2}Q$, where Q is the orthogonal factor from the standard QR factorization $\Lambda^{1/2}V^TZ^{(0)} = QU$. The factor U is thus the triangular factor from the classical orthogonalization of $\Lambda^{1/2}V^TZ^{(0)}$ with respect to the Euclidean inner product. Assuming that these two main ingredients are implemented in a backward stable way this approach represents probably the most accurate algorithm one can get for the general case of a symmetric positive definite matrix A (if we look at the loss of orthogonality between computed vectors). The backward stable eigendecomposition delivers the computed eigendecomposition $\bar{V}\bar{\Lambda}\bar{V}^T$ which is nearly the exact eigendecomposition of a nearby matrix

$$A + \Delta A = (\bar{V} + \Delta V)\bar{\Lambda}(\bar{V} + \Delta V)^T, \quad \|\Delta A\| \leq O(u)\|A\|, \tag{3}$$

where $\hat{V} = \bar{V} + \Delta V$ is orthogonal and $\|\Delta V\| \leq O(u)$ (see [28]). Multiplying the matrix $Z^{(0)}$ with $\bar{\Lambda}^{1/2}\bar{V}^T$ from the left and applying a backward stable QR decomposition (such as Householder QR [17]) to the product $fl(\bar{\Lambda}^{1/2}\bar{V}^TZ^{(0)})$ we can write for the computed factors $\bar{Q}$ and $\bar{U}$ the identity

$$\bar{\Lambda}^{1/2}\bar{V}^TZ^{(0)} = (\bar{Q} + \Delta Q)\bar{U} + \Delta E_1, \quad \|\Delta E_1\| \leq O(u)\|A\|^{1/2}\|Z^{(0)}\|, \tag{4}$$

where $\hat{Q} = \bar{Q} + \Delta Q$ is orthogonal and $\|\Delta Q\| \leq O(u)$. The matrix $\bar{Z}$ is then computed as the product of two nearly orthogonal and one diagonal matrix $\bar{Z} = fl(\bar{V}\bar{\Lambda}^{-1/2}\bar{Q})$ satisfying

$$\bar{Z} = \hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2, \quad \|\Delta E_2\| \leq O(u)\|\bar{\Lambda}^{-1}\|^{1/2} \leq O(u)\|\bar{Z}\|. \tag{5}$$

Considering (4) we have $Z^{(0)} = \hat{V}\bar{\Lambda}^{-1/2}\hat{Q}\bar{U} + \hat{V}\bar{\Lambda}^{-1/2}\Delta E_1 + \hat{V}\Delta V^TZ^{(0)}$. Using (5) we get the factorization for computed quantities

$$Z^{(0)} + \Delta E_3 = \bar{Z}\bar{U}, \quad \|\Delta E_3\| \leq O(u)\kappa^{1/2}(A)\|Z^{(0)}\|. \tag{6}$$

Note that the factor $\kappa^{1/2}(A)\|Z^{(0)}\|$ appears due to the fact that there exists only a normwise bound for the matrix ΔE_1 for the QR factorization in (4). As we will see later the bound (6) that holds for general symmetric and positive definite matrix A can be further improved for the Gram-Schmidt implementations. If $A = \Lambda$ is in addition diagonal, there is no need for the decomposition (3) and one can compute $\bar{Z}$ directly from Λ and $\bar{Q}$ as $\bar{Z} = \Lambda^{-1/2}\bar{Q} + \Delta E_2$ with $|\Delta E_2| \leq O(u)|\Lambda^{-1/2}||\bar{Q}|$. Due

to (4) rewritten as $\Lambda^{1/2}Z^{(0)} = \bar{Q}\bar{U} + \Delta E_1$ we get $Z^{(0)} = \bar{Z}\bar{U} - \Delta E_2\bar{U} + \Lambda^{-1/2}\Delta E_1$ leading to the better bound $\|\Delta E_3\| \leq O(u)\|\Lambda^{-1}\|^{1/2}\|\Lambda^{1/2}Z^{(0)}\|$. For a general A the loss of orthogonality between the columns in the computed factor $\bar{Z}$ can be expressed as $\bar{Z}^T A\bar{Z} - I = (\hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2)^T A(\hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2) - I$ which gives then the following bound for its norm

$$\|\bar{Z}^T A\bar{Z} - I\| \leq O(u)\|A\|\|\bar{Z}\|^2 \leq O(u)\kappa(A). \tag{7}$$

Again, for a diagonal A we can write $\bar{Z}^T A\bar{Z} - I = (\Lambda^{-1/2}\bar{Q} + \Delta E_2)^T A(\Lambda^{-1/2}\bar{Q} + \Delta E_2) - I = \bar{Q}^T\bar{Q} - I + \bar{Q}\Lambda^{1/2}\Delta E_2 + (\Delta E_2)^T\Lambda^{1/2}\bar{Q} + (\Delta E_2)^T\Lambda\Delta E_2$. Since $\|\Lambda^{1/2}\Delta E_2\| \leq O(u)$ this identity then gives rise to the bound $\|\bar{Z}^T A\bar{Z} - I\| \leq O(u)$. Indeed for a diagonal A the orthogonality of computed vectors remains on the roundoff unit level and is not dependent on the matrix $A^{1/2}Z^{(0)}$.

The error in the inverse factorization can be written as $A^{-1} - \bar{Z}\bar{Z}^T = \hat{V}\bar{\Lambda}^{-1/2}(I - \bar{\Lambda}^{-1/2}\hat{V}^T\Delta A\hat{V}\bar{\Lambda}^{-1/2})^{-1}\bar{\Lambda}^{-1/2}\hat{V}^T - (\hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2)(\hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2)^T$ which leads to the factorization for computed factors

$$A^{-1} + \Delta E_4 = \bar{Z}\bar{Z}^T, \quad \|\Delta E_4\| \leq O(u)\|A\|\|\bar{\Lambda}^{-1}\|^2 \leq O(u)\kappa(A)\|A^{-1}\|. \tag{8}$$

Similarly we can express $A\bar{Z}\bar{Z}^T - I = (\hat{V}\bar{\Lambda}\hat{V}^T - \Delta A)(\hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2)(\hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2)^T - I$ and $\bar{Z}\bar{Z}^T A - I = (\hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2)(\hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2)^T(\hat{V}\bar{\Lambda}\hat{V}^T - \Delta A) - I$ and get the estimates for the right and left residuals

$$\|\bar{Z}\bar{Z}^T A - I\| = \|A\bar{Z}\bar{Z}^T - I\| \leq \|\bar{\Lambda}^{1/2}\|\|\Delta E_2\| + \|\bar{\Lambda}\|\|\bar{Z}\|\|\Delta E_2\| + \|\Delta A\|\|\bar{Z}\|^2, \tag{9}$$

ending up with $\|A\bar{Z}\bar{Z}^T - I\| \leq O(u)\kappa(A)$ or $\|\bar{Z}\bar{Z}^T A - I\| \leq O(u)\kappa(A)$. The bounds (7) and (9) do not depend on the conditioning of $Z^{(0)}$ or $A^{1/2}Z^{(0)}$. Actually these matrices appear in the analysis only implicitly in $\bar{\Lambda}^{1/2}\bar{V}^T Z^{(0)}$ in the QR decompositon (4). Indeed the factor Z is computed as a product of two (nearly) orthogonal and one diagonal matrix with a condition number equal to $\kappa^{1/2}(A)$. This is probably the best approach one can get in finite precision arithmetic and in this sense the backward stable eigendecomposition-based implementation can be considered as an optimal algorithm. One can hardly expect that bounds on $\bar{Z}^T A\bar{Z} - I$, $A\bar{Z}\bar{Z}^T - I$ and $\bar{Z}\bar{Z}^T A - I$ will not depend on the conditioning of the matrix A at least for general symmetric positive definite A - of course in the case of the standard inner product $A = I$ all these quantities are the order of the roundoff unit u and they do not depend on $\kappa(Z^{(0)})$.

3 Modified Gram-Schmidt orthogonalization

Probably the most frequently used orthogonalization algorithm is the modified Gram-Schmidt process which represents a good compromise between the efficiency and numerical stability. In this section we show that in finite precision arithmetic for the modified Gram-Schmidt orthogonalization with the nonstandard inner product the error in the factorization $Z^{(0)} - \bar{Z}\bar{U}$ is small and independent of $\kappa(A^{1/2}Z^{(0)})$. This is no longer true for the loss of orthogonality $\bar{Z}^T A\bar{Z} - I$ where the condition number of the matrix $A^{1/2}Z^{(0)}$ plays a dominant role. Since there is a significant difference in the accuracy of the computed inner product $fl[\langle\cdot,\cdot\rangle_A]$ we will distinguish between the case of a general positive definite matrix A and the case when the inner product is induced with a positive and diagonal matrix A.

Given the matrix $Z^{(0)} = [z_1^{(0)}, \ldots, z_n^{(0)}]$ we consider the following algorithm for computing the factor $Z = [z_1, \ldots, z_n]$ such that for all $i = 1, \ldots n$ and $j = 1, \ldots, i-1$ we define the matrices $Z^{(j)} = [z_1^{(j)}, \ldots, z_n^{(j)}]$ with the recurrences

$$z_i^{(j)} = z_i^{(j-1)} - \alpha_{ji} z_j \equiv z_i^{(j-1)} - \langle z_i^{(j-1)}, z_j \rangle_A z_j. \tag{10}$$

The i-th column of Z is then given as $z_i \equiv z_i^{(i-1)}/\|z_i^{(i-1)}\|_A$. The orthogonalization coefficients α_{ji} form the upper triangular factor U together with the diagonal elements defined as $\alpha_{ii} = \|z_i^{(i-1)}\|_A$ for all $i = 1, \ldots n$ and $j = 1, \ldots, i-1$.

Due to rounding errors the computed vectors $\bar{z}_i^{(j)}$ satisfy after each projection the formula with the local errors $\Delta\delta_i^{(j)}$

$$\bar{z}_i^{(j)} = \bar{z}_i^{(j-1)} - \bar{\alpha}_{ji}\bar{z}_j + \Delta\delta_i^{(j)}, \ |\Delta\delta_i^{(j)}| \le u|\bar{z}_i^{(j-1)}| + 2u|\bar{\alpha}_{ji}||\bar{z}_j|. \tag{11}$$

Summarizing (11) for indices $j = 1, \ldots, i-1$ together with the definition of vectors $\bar{z}_i = \mathrm{fl}[\bar{z}_i^{(i-1)}/\bar{\alpha}_{ii}]$ implying $\bar{z}_i^{(i-1)} = \bar{\alpha}_{ii}\bar{z}_i - \Delta\delta_i^{(i)}$ with $|\Delta\delta_i^{(i)}| \le O(u)|\bar{\alpha}_{ii}||\bar{z}_i|$ gives

$$\bar{\alpha}_{ii}\bar{z}_i = z_i^{(0)} - \sum_{j=1}^{i-1} \bar{\alpha}_{ji}\bar{z}_j + \sum_{j=1}^{i} \Delta\delta_i^{(j)}, \|\sum_{j=1}^{i} \Delta\delta_i^{(j)}\| \le O(u) \left[\|z_i^{(0)}\| + \sum_{j=1}^{i} |\bar{\alpha}_{ji}|\|\bar{z}_j\| \right]. \tag{12}$$

The term $u|\bar{z}_i^{(j-1)}|$ in (11) can be bounded by $2u\left[|z_i^{(0)}| + \sum_{k=1}^{j-1}|\bar{\alpha}_{ki}||\bar{z}_k|\right]$. This leads to the bound for $\|\Delta\delta_i^{(j)}\| \le O(u)\left[\|z_i^{(0)}\| + \sum_{k=1}^{j}|\bar{\alpha}_{ki}|\|\bar{z}_k\|\right]$. Introducing then the matrix $\Delta E^{(1)}$ which contains the local errors $\sum_{j=1}^{i}\Delta\delta_i^{(j)}$ as its columns we obtain the first of two main results for the Gram-Schmidt orthogonalization

$$Z^{(0)} + \Delta E^{(1)} = \bar{Z}\bar{U}, \quad \|\Delta E^{(1)}\| \le O(u)\left[\|Z^{(0)}\| + \|\bar{Z}\|\|\bar{U}\|\right]. \tag{13}$$

In the modified Gram-Schmidt algorithm the computed coefficients are given as $\bar{\alpha}_{ji} = \mathrm{fl}[\langle \bar{z}_i^{(j-1)}, \bar{z}_j \rangle_A]$ and $\bar{\alpha}_{ii} = \mathrm{fl}[\|\bar{z}_i^{(i-1)}\|_A]$. Thus $u|\bar{\alpha}_{ki}| \le u\|\bar{z}_i^{(k-1)}\|_A \le u\|z_i^{(0)}\|_A$. Due to $\|\bar{Z}\| \lessapprox \|A^{-1}\|^{1/2}$ and $\|\bar{U}\| \le \|\bar{U}\|_F \lessapprox \|A^{1/2}Z^{(0)}\|$ we get a somewhat better bound than in (6). Indeed it follows that $\|\Delta\delta_i^{(j)}\| \le O(u)\|A^{-1}\|^{1/2}\|z_i^{(0)}\|_A$ and $\|\Delta E^{(1)}\| \le O(u)\|A^{-1}\|^{1/2}\|A^{1/2}Z^{(0)}\| \le O(u)\kappa^{1/2}(A)\|Z^{(0)}\|$. In addition we have $\|\Delta\delta_i^{(j)}\|_A \le \|A\|^{1/2}\|\Delta\delta_i^{(j)}\| \le O(u)\kappa^{1/2}(A)\|z_i^{(0)}\|_A$ for a general symmetric positive definite A. For A positive and diagonal we can improve this bound using (11) and show $\|\Delta\delta_i^{(j)}\|_A \le O(u)\|\bar{z}_i^{(j-1)}\|_A \le O(u)\|z_i^{(0)}\|_A$. This result is based on (21), see the discussion later in the text. Note also that the derivation of (13) does not depend on the way how we compute the coefficients $\bar{\alpha}_{ki}$. Therefore a similar result will hold also for the classical Gram-Schmidt (CGS) algorithm as well as for the AINV orthogonalization which will be discussed in next section.

Considering recursively the formula (11) for $j = k+1, \ldots, i-1$, rearranging the resulting identity and taking the A-inner product with the vector $\bar{z}_k$ we obtain

$$\sum_{j=k+1}^{i} \bar{\alpha}_{ji}\langle \bar{z}_k, \bar{z}_j \rangle_A = \langle \bar{z}_k, \bar{z}_i^{(k)} \rangle_A + \sum_{j=k+1}^{i} \langle \bar{z}_k, \Delta\delta_i^{(j)} \rangle_A. \tag{14}$$

The strongest property of the modified Gram-Schmidt process is that the local orthogonality between two consecutive computed vectors is well preserved. Indeed if we look at the inner product of the vector $\bar{z}_i^{(k)}$ with the vector $\bar{z}_k$ one can write $\langle \bar{z}_k, \bar{z}_i^{(k)} \rangle_A = -\langle \bar{z}_k, \Delta\eta_i^{(k)} \rangle_A + \langle \bar{z}_k, \Delta\delta_i^{(k)} \rangle_A$, where the local error $\Delta\eta_i^{(k)}$ is given as

$$\Delta\eta_i^{(k)} = \left(\mathrm{fl}[\langle \bar{z}_i^{(k-1)}, \bar{z}_k \rangle_A] - \langle \bar{z}_i^{(k-1)}, \bar{z}_k \rangle_A \right) \bar{z}_k + \left(\|\bar{z}_k\|_A^2 - 1 \right) \bar{z}_i^{(k-1)}. \tag{15}$$

The left-hand side of (14) represents the (k,i)-component of the matrix $\Delta E^{(3)}\bar{U}$, where $\Delta E^{(3)}$ is a strictly upper triangular containing the off-diagonal elements of the matrix

$$\bar{Z}^T A \bar{Z} - I = \Delta E^{(4)} + \Delta E^{(3)} + (\Delta E^{(3)})^T, \tag{16}$$

where $\Delta E^{(4)}$ is diagonal. The identity (14) can be then rewritten into simple matrix form $\Delta E^{(3)}\bar{U} = \Delta E^{(2)}$, where the matrix $\Delta E^{(2)}$ is defined by the elements $-\langle \bar{z}_k, \Delta\eta_i^{(k)} \rangle_A + \langle \bar{z}_k, \sum_{j=k}^{i} \Delta\delta_i^{(j)} \rangle_A$. The norm of $\Delta E^{(3)}$ from (16) can be bounded by $\|\Delta E^{(3)}\| \leq \|\Delta E^{(2)}\|_F \|\bar{U}^{-1}\|$. The computed factors $\bar{Z}$ and $\bar{U}$ satisfy (13) and (16) and therefore we can write

$$\bar{U}^T\bar{U} = (Z^{(0)} + \Delta E^{(1)})^T A (Z^{(0)} + \Delta E^{(1)}) - \bar{U}^T [\Delta E^{(4)} + \Delta E^{(3)} + (\Delta E^{(3)})^T] \bar{U}.$$

Since the exact factorization is $Z^{(0)} = ZU$ and $\Delta E^{(3)}\bar{U} = \Delta E^{(2)}$ the matrix $\bar{U}^T\bar{U}$ can be related to $U^TU = (Z^{(0)})^T A Z^{(0)}$ as follows

$$\begin{aligned}\bar{U}^T\bar{U} = U^T \Big[& I + Z^T A \Delta E^{(1)} U^{-1} + (Z^T A \Delta E^{(1)} U^{-1})^T + (\Delta E^{(1)} U^{-1})^T A (\Delta E^{(1)} U^{-1}) \\ & + (\bar{U}U^{-1})^T \Delta E^{(4)} (\bar{U}U^{-1}) + (\bar{U}U^{-1})^T \Delta E^{(2)} U^{-1} + ((\bar{U}U^{-1})^T \Delta E^{(2)} U^{-1})^T \Big] U.\end{aligned}$$

This gives rise to the equation $(\bar{U}U^{-1})^T(\bar{U}U^{-1}) = I + \Delta E^{(5)}$, where the norm of the error matrix $\Delta E^{(5)}$ satisfies the inequality

$$\begin{aligned}\|\Delta E^{(5)}\| \leq\ & 2\|A\|^{1/2} \|\Delta E^{(1)}\| \|U^{-1}\| + \|A\| \|\Delta E^{(1)}\|^2 \|U^{-1}\|^2 \\ & + \|\Delta E^{(4)}\| \|\Delta E^{(5)}\| + 2\|\Delta E^{(2)}\| \|U^{-1}\| (1 + \|\Delta E^{(5)}\|^2)^{1/2}.\end{aligned}$$

It is clear that if we assume that $2\|\Delta E^{(2)}\|\|U^{-1}\| + \|\Delta E^{(4)}\| < 1$ then $\|\Delta E^{(5)}\| \lessapprox 2(\|A\|^{1/2}\|\Delta E^{(1)}\| + \|\Delta E^{(2)}\|)\|U^{-1}\|$ and $\|\bar{U}^{-1}\|^2 \leq (1 - \|\Delta E^{(5)}\|)^{-1}\|U^{-1}\|^2$. The elements of the matrix $\Delta E^{(3)}$ thus depend significantly on the magnitude of matrices $\Delta E^{(1)}$ and $\Delta E^{(2)}$. The definition of $\Delta E^{(2)}$ indicates that the A-norms of local errors $\|\Delta\delta_i^{(j)}\|_A$ and $\|\Delta\eta_i^{(k)}\|_A$ play a decisive role here. From (15) it follows that we need to estimate the terms $\mathrm{fl}[\langle \bar{z}_i^{(k-1)}, \bar{z}_k \rangle_A] - \langle \bar{z}_i^{(k-1)}, \bar{z}_k \rangle_A \rangle$ and $\|\bar{z}_k\|_A^2 - 1$. In addition, the second term defines the elements of the matrix $\Delta E^{(4)}$.

If the inner product is induced by a general symmetric positive definite matrix A, then the error in computing $\mathrm{fl}[\langle \bar{z}_i^{(k-1)}, \bar{z}_k \rangle_A]$ can be bounded by

$$|\mathrm{fl}[\langle \bar{z}_i^{(k-1)}, \bar{z}_k \rangle_A] - \langle \bar{z}_i^{(k-1)}, \bar{z}_k \rangle_A| \leq O(u)\|A\| \|\bar{z}_i^{(k-1)}\| \|\bar{z}_k\|. \tag{17}$$

Note that (17) can be rather pessimistic and it may be worth to consider the details in computation of the $\langle \cdot, \cdot \rangle_A$. The result in finite precision arithmetic depends on the fact whether we multiply the first or the second argument by A (see the discussion in numerical experiments). Since the vector $\bar{z}_k = \mathrm{fl}[\bar{z}_k^{(k-1)}/\bar{\alpha}_{kk}]$ with $\bar{\alpha}_{kk} = \mathrm{fl}[\|\bar{z}_k^{(k-1)}\|_A]$

is just the computed result from the normalization of the vector $\bar{z}_k^{(k-1)}$ with respect to the inner product with a general A one can conclude that

$$\left| \|\bar{z}_k\|_A^2 - 1 \right| \leq O(u)\|A\|\|\bar{z}_k\|^2 \leq O(u)\kappa(A). \tag{18}$$

It follows from (15), (17) and (18) that $\|\Delta\eta_i^{(k)}\|_A \leq O(u)\|A\|\|\bar{z}_k\|(\|\bar{z}_i^{(k-1)}\|\|\bar{z}_k\|_A + \|\bar{z}_k\|\|\bar{z}_i^{(k-1)}\|_A) \leq O(u)\|A\|^{1/2}\kappa^{1/2}(A)\|\bar{z}_k\|\|z_i^{(0)}\|_A$. Since we have already $\|\Delta\delta_i^{(j)}\|_A \leq O(u)\kappa^{1/2}(A)\|z_i^{(0)}\|_A$ the Frobenius norm of the matrix $\Delta E^{(2)}$ and the 2-norm of the matrix $\Delta E^{(4)}$ can be then bounded by

$$\|\Delta E^{(2)}\|_F \leq O(u)\|A\|^{1/2}\|\bar{Z}\|\kappa^{1/2}(A)\|A^{1/2}\bar{Z}^{(0)}\|, \quad \|\Delta E^{(4)}\| \leq O(u)\|A\|\|\bar{Z}\|^2. \tag{19}$$

For a general symmetric positive definite A assuming $O(u)\kappa(A)\kappa(A^{1/2}Z^{(0)}) < 1$ we can conclude that the loss of orthogonality between the computed vectors $\bar{Z}$ is bounded by a quantity proportional not only to the condition number of the matrix $A^{1/2}Z^{(0)}$ but also to the condition number of the matrix A which is actually the upper bound for the size of local errors in the computation of associated inner products

$$\|\bar{Z}^T A\bar{Z} - I\| \leq O(u)\|A\|^{1/2}\|\bar{Z}\|\kappa^{1/2}(A)\|A^{1/2}Z^{(0)}\|\|\bar{U}^{-1}\| \leq \frac{O(u)\kappa(A)\kappa(A^{1/2}Z^{(0)})}{1 - O(u)\kappa(A)\kappa(A^{1/2}Z^{(0)})}. \tag{20}$$

The situation is more transparent when A is diagonal (and positive definite). Then for the difference of the computed $\mathrm{fl}[\langle\bar{z}_i^{(k-1)}, \bar{z}_k\rangle_A]$ and the exact inner product $\langle\bar{z}_i^{(k-1)}, \bar{z}_k\rangle_A$ we have the bound

$$|\mathrm{fl}[\langle\bar{z}_i^{(k-1)}, \bar{z}_k\rangle_A] - \langle\bar{z}_i^{(k-1)}, \bar{z}_k\rangle_A| \leq O(u)\|\bar{z}_i^{(k-1)}\|_A\|\bar{z}_k\|_A \tag{21}$$

and for the error in the normalization of the vector $\bar{z}_k^{(k-1)}$ one can write

$$\left| \|\bar{z}_k\|_A^2 - 1 \right| \leq O(u). \tag{22}$$

The previous two results lead to significantly better bounds for the local errors $\|\Delta\delta_i^{(j)}\|_A \leq O(u)\|z_i^{(0)}\|_A$ and $\|\Delta\eta_i^{(k)}\|_A \leq O(u)\|z_i^{(0)}\|_A$. In matrix notation we then have bounds

$$\|\Delta E^{(2)}\|_F \leq O(u)\|A^{1/2}Z^{(0)}\|, \quad \|\Delta E^{(4)}\| \leq O(u). \tag{23}$$

Indeed the relative local errors are small multiples of the roundoff unit and do not depend on the conditioning of the matrix A. For A diagonal we thus get the result analogous to the case with standard inner product [6]. Assuming that $O(u)\kappa(A^{1/2}Z^{(0)}) < 1$ the loss of orthogonality between the computed vectors $\bar{Z}$ is bounded by

$$\|\bar{Z}^T A\bar{Z} - I\| \leq O(u)\|A^{1/2}Z^{(0)}\|\|\bar{U}^{-1}\| \leq \frac{O(u)\kappa(A^{1/2}Z^{(0)})}{1 - O(u)\kappa(A^{1/2}Z^{(0)})}. \tag{24}$$

For a diagonal A the matrix $A^{1/2}Z^{(0)}$ is just the matrix $Z^{(0)}$ scaled by row. It is well-known that the orthogonality of computed vectors in the modified Gram-Schmidt process (with the standard inner product) is independent of the column-scaling of the original matrix $Z^{(0)}$. The effect of row-scaling thus seems to be similar to the application of weighted modified Gram-Schmidt process, i.e. the MGS algorithm with the A-inner product applied to the columns of $Z^{(0)}$. This process has been extensively studied by Gulliksson in [15], see also [14] and [16].

4 Classical Gram-Schmidt and AINV orthogonalization

In the classical Gram-Schmidt algorithm the coefficients α_{ji} are computed as $\alpha_{ji} = \langle z_i^{(0)}, z_j \rangle_A$ for $j = 1, \dots i-1$. The computed coefficients $\bar{\alpha}_{ji} = fl[\langle z_i^{(0)}, \bar{z}_j \rangle_A]$ thus satisfy

$$|fl[\langle z_i^{(0)}, \bar{z}_j \rangle_A] - \langle z_i^{(0)}, \bar{z}_j \rangle_A| \le O(u)\|A\|\|z_i^{(0)}\|\|\bar{z}_j\|. \tag{25}$$

As we have already noted in the previous section the recurrence (11) for computed vectors $\bar{z}_i = \mathrm{fl}[\bar{z}_i^{(i-1)}/\bar{\alpha}_{ii}]$ will have the same form together with bounds (12) and (13). It was shown [31] that diagonal entries α_{ii} must be computed using the formula

$$\alpha_{ii} = (\|z_i^{(0)}\|_A^2 - \|Z_{i-1}^T A z_i^{(0)}\|^2)^{1/2} = \left(\|z_i^{(0)}\|_A^2 - \sum_{k=1}^{i-1} \alpha_{ki}^2 \right)^{1/2}. \tag{26}$$

The computed elements $\bar{\alpha}_{ki}$ for $k = 1, \dots, i$ then satisfy the bound

$$\left| \|z_i^{(0)}\|_A^2 - \sum_{j=1}^{i} \bar{\alpha}_{ji}^2 \right| \le O(u)\|A\|\|z_i^{(0)}\|^2. \tag{27}$$

The bound for the matrix $\Delta E^{(1)}$ is even more straightforward since from (25) we have

$$\|\Delta E^{(1)}\| \le O(u) \left[\|Z^{(0)}\| + \|\bar{Z}\|\|\bar{U}\| \right] \le O(u)(\|Z^{(0)}\| + \|\bar{Z}\|\|A^{1/2} Z^{(0)}\|). \tag{28}$$

The worst-case bounds $\|\bar{Z}\| \lessapprox \|A^{-1}\|^{1/2}$ and $\|\bar{U}\| \lessapprox \|A^{1/2} Z^{(0)}\|$ imply $\|\Delta E^{(1)}\| \le O(u)\|A^{-1}\|^{1/2}\|A^{1/2} Z^{(0)}\| \le O(u)\kappa^{1/2}(A)\|Z^{(0)}\|$. From (12) for each $j = 1, \dots, i-1$ we have $\bar{\alpha}_{jj}\bar{z}_j = z_j^{(0)} - \sum_{k=1}^{j-1} \bar{\alpha}_{ki}\bar{z}_k + \sum_{k=1}^{j} \Delta\delta_j^{(k)}$. Taking the A-inner product with $z_i^{(0)}$ and after some rearranging we get

$$\bar{\alpha}_{jj}\langle z_i^{(0)}, \bar{z}_j \rangle_A = \langle z_i^{(0)}, z_j^{(0)} \rangle_A - \sum_{k=1}^{j-1} \bar{\alpha}_{kj}\langle z_i^{(0)}, \bar{z}_k \rangle_A + \langle z_i^{(0)}, \sum_{k=1}^{j} \Delta\delta_j^{(k)} \rangle_A,$$

$$\sum_{k=1}^{j} \bar{\alpha}_{kj}\langle z_i^{(0)}, \bar{z}_k \rangle_A = \langle z_i^{(0)}, z_j^{(0)} \rangle_A + \langle z_i^{(0)}, \sum_{k=1}^{j} \Delta\delta_j^{(k)} \rangle_A,$$

$$\sum_{k=1}^{j} \bar{\alpha}_{kj}\bar{\alpha}_{ki} = \langle z_i^{(0)}, z_j^{(0)} \rangle_A + \sum_{k=1}^{j} \bar{\alpha}_{kj} \left(fl[\langle z_i^{(0)}, \bar{z}_k \rangle_A] - \langle z_i^{(0)}, \bar{z}_k \rangle_A \right) + \langle z_i^{(0)}, \sum_{k=1}^{j} \Delta\delta_j^{(k)} \rangle_A.$$

Let the last two terms of this equation define the (i,j)-th element of the error matrix ΔE. Considering (25), (27) and the bound for the local errors developed in the previous section $\|\Delta\delta_j^{(k)}\|_A \le O(u)\|A\|^{1/2} \left[\|z_j^{(0)}\| + \sum_{s=1}^{k} |\bar{\alpha}_{sj}|\|\bar{z}_s\| \right]$ we obtain the matrix identity (the analogous result for the standard inner product can be found in [12])

$$(Z^{(0)})^T A Z^{(0)} + \Delta E = \bar{U}^T \bar{U}, \|\Delta E\| \le O(u)\|A\|\|Z^{(0)}\|(\|Z^{(0)}\| + \|\bar{Z}\|\|A^{1/2} Z^{(0)}\|) \tag{29}$$

Indeed the computed factor $\bar{U}$ is the exact Cholesky factor of the matrix $(Z^{(0)})^T A Z^{(0)} + \Delta E$, where $\|\Delta E\| \le O(u)\kappa^{1/2}(A)\|A\|^{1/2}\|Z^{(0)}\|\|A^{1/2} Z^{(0)}\|$. In other words, up to the factor $\kappa^{1/2}(A)$ estimating the local errors, the classical Gram-Schmidt algorithm is a

way how to compute a backward stable Cholesky factor of the cross-product matrix $(Z^{(0)})^T A Z^{(0)}$ (see also [11], [12]). Using (29) and (13) with (28) we derive

$$\bar{U}^T(I - \bar{Z}^T A \bar{Z})\bar{U} = -(\Delta E^{(1)})^T A Z^{(0)} - (Z^{(0)})^T A \Delta E^{(1)} + (\Delta E^{(1)})^T A \Delta E^{(1)} + \Delta E,$$

which gives rise to the bound for the loss of orthogonality that depends quadratically on the minimal singular value of $A^{1/2}Z^{(0)}$. Indeed using (28) we have

$$\|I - \bar{Z}^T A \bar{Z}\| \leq O(u)\|A\|\|\bar{Z}\|\|Z^{(0)}\|\|\bar{U}\|\|\bar{U}^{-1}\|^2 \leq \frac{O(u)\kappa(A)\kappa(A^{1/2}Z^{(0)})\kappa(Z^{(0)})}{1 - O(u)\kappa(A)\kappa(A^{1/2}Z^{(0)})}, \tag{30}$$

where the size of local errors is reflected similarly as in the modified Gram-Schmidt algorithm with the worst-case bound $\|A\|^{1/2}\|\bar{Z}\| \leq \kappa^{1/2}(A)$. In the case of a diagonal A one can show due to $|fl[\langle z_i^{(0)}, \bar{z}_k\rangle_A] - \langle z_i^{(0)}, \bar{z}_k\rangle_A| \leq O(u)\|z_i^{(0)}\|_A\|\bar{z}_k\|_A$ that $\|A^{1/2}\Delta E^{(1)}\| \leq O(u)\|A^{1/2}Z^{(0)}\|$ and $\|\Delta E\| \leq \|A^{1/2}Z^{(0)}\|^2$ which give rise to a significantly better bound in the form

$$\|I - \bar{Z}^T A \bar{Z}\| \leq O(u)\|A^{1/2}Z^{(0)}\|^2\|\bar{U}^{-1}\|^2 \leq \frac{O(u)\kappa^2(A^{1/2}Z^{(0)})}{1 - O(u)\kappa^2(A^{1/2}Z^{(0)})\kappa(Z^{(0)})}. \tag{31}$$

In the following we will analyze the AINV orthogonalization scheme and show that its numerical behavior is very similar to CGS. Indeed, the coefficients α_{ji} in the recurrence for the vectors $z_i^{(j)}$ can be also determined using oblique projection as $\alpha_{ji} = \langle z_i^{(j-1)}, z_j^{(0)}\rangle_A / \langle z_j, z_j^{(0)}\rangle_A$. Where $\langle z_j, z_j^{(0)}\rangle_A = \langle z_j, \sum_{k=1}^{j} \alpha_{kj} z_k\rangle_A = \alpha_{jj}$. The coefficients α_{jj} are computed with the nonstandard formula

$$\alpha_{jj} = \left(\|z_j^{(0)}\|_A^2 - \sum_{k=1}^{j-1} \alpha_{kj}^2 \right)^{1/2}. \tag{32}$$

This algorithm is a modification of the modified Gram-Schmidt „towards" the classical Gram-Schmidt algorithm and in the context of A-orthogonalization it is known and widely used as the AINV preconditioner. Its analogue for the case of the standard inner product is not used since it is clearly not competitive with the MGS algorithm. Indeed the recurrence (13) will remain true also for quantities computed in the AINV orthogonalization algorithm, whereas the norm of the matrix $\Delta E^{(1)}$ can be bounded as $\|\Delta E^{(1)}\| \leq O(u)[\|Z^{(0)}\| + \|\bar{Z}\|\|\bar{U}\|] \leq O(u)\|A^{-1}\|^{1/2}\|A^{1/2}Z^{(0)}\|$. The last bound can be shown only under assumption that the orthogonality is not lost completely. Otherwise the error in computing these oblique projections may be significantly larger than for orthogonal projections in CGS or MGS. The computed orthogonalization coefficients $\bar{\alpha}_{ji}$ then can be expressed as $\bar{\alpha}_{ji} = fl[\langle \bar{z}_i^{(j-1)}, \bar{z}_j^{(0)}\rangle_A]$, where $\bar{z}_j^{(0)} = fl[z_j^{(0)}/\bar{\alpha}_{jj}]$. From (11) we can write that $\bar{z}_i^{(j)} = z_i^{(0)} - \sum_{k=1}^{j-1} \bar{\alpha}_{ki}\bar{z}_k + \sum_{k=1}^{j-1} \Delta\delta_i^{(k)}$. Taking the A-inner product with the initial vector $z_j^{(0)}$ we obtain successively

$$\bar{\alpha}_{jj}\langle \bar{z}_i^{(j-1)}, \frac{z_j^{(0)}}{\bar{\alpha}_{jj}}\rangle_A = \langle z_i^{(0)}, z_j^{(0)}\rangle_A - \sum_{k=1}^{j-1} \bar{\alpha}_{ki}\langle \bar{z}_k, z_j^{(0)}\rangle_A + \sum_{k=1}^{j-1} \langle \Delta\delta_i^{(k)}, z_j^{(0)}\rangle_A$$

$$\bar{\alpha}_{jj}\langle \bar{z}_i^{(j-1)}, \bar{z}_j^{(0)}\rangle_A = \langle z_i^{(0)}, z_j^{(0)}\rangle_A - \sum_{k=1}^{j-1} \bar{\alpha}_{ki}\langle \bar{z}_k, z_j^{(0)}\rangle_A + \sum_{k=1}^{j-1} \langle \Delta\delta_i^{(k)}, z_j^{(0)}\rangle_A$$

$$+ \bar{\alpha}_{jj}\langle \bar{z}_i^{(j-1)}, \bar{z}_j^{(0)} - \frac{z_j^{(0)}}{\bar{\alpha}_{jj}}\rangle_A, \tag{33}$$

$$\sum_{k=1}^{j-1} \bar{\alpha}_{kj}\hat{\alpha}_{ki} = \langle z_i^{(0)}, z_j^{(0)}\rangle_A + \sum_{k=1}^{j-1}\langle \Delta\delta_i^{(k)}, z_j^{(0)}\rangle_A + \bar{\alpha}_{jj}\left(fl[\langle \bar{z}_i^{(j-1)}, \bar{z}_j^{(0)}\rangle_A] - \langle \bar{z}_i^{(j-1)}, \frac{z_j^{(0)}}{\bar{\alpha}_{jj}}\rangle_A \right),$$

where $\hat{\alpha}_{ki} = \langle z_i^{(0)}, \bar{z}_k\rangle_A$ for $k = 1, \ldots, i-1$ are the coefficients of the upper triangular matrix $\hat{U}$ (the diagonal of $\hat{U}$ will be identical to the diagonal of $\bar{U}$) and where

$$\left| fl[\langle \bar{z}_i^{(j-1)}, \bar{z}_j^{(0)}]\rangle_A] - \langle \bar{z}_i^{(j-1)}, \frac{z_j^{(0)}}{\bar{\alpha}_{jj}}\rangle_A \right| \leq O(u)\|A\|\|\bar{z}_i^{(j-1)}\| \frac{\|z_j^{(0)}\|}{|\bar{\alpha}_{jj}|}. \tag{34}$$

Since the left-hand side of the recurrence (34) is just the (j,i)-element of the matrix $\bar{U}^T\hat{U}$ it can be rewritten in matrix notation into an identity with the strictly upper triangular part of the matrix ΔF satisfying

$$\begin{aligned} &striu(\bar{U}^T\hat{U}) = striu((Z^{(0)})^T AZ^{(0)} + \Delta F), \\ &\|striu(\Delta F)\| \leq O(u)\ \kappa^{1/2}(A)\|A\|^{1/2}\|Z^{(0)}\|\|A^{1/2}Z^{(0)}\|. \end{aligned} \tag{35}$$

The diagonal elements of α_{ii} are computed with the formula (32). The computed quantities $\bar{\alpha}_{ki}$ satisfy $|\|z_i^{(0)}\|_A^2 - \sum_{k=1}^{i}\bar{\alpha}_{ki}^2| \leq O(u)\|A\|\|z_i^{(0)}\|^2$. The diagonal entries of the matrix ΔF thus satisfy the bound

$$diag(\bar{U}^T\hat{U}) = diag((Z^{(0)})^T AZ^{(0)} + \Delta F),\ \|diag(\Delta F)\| \leq O(u)\|A\|\|Z^{(0)}\|^2. \tag{36}$$

From (12) for each $j = 1, \ldots, i-1$ we have $\bar{\alpha}_{jj}\bar{z}_j = z_j^{(0)} - \sum_{k=1}^{j-1}\bar{\alpha}_{ki}\bar{z}_k + \sum_{k=1}^{j}\Delta\delta_j^{(k)}$. Taking the A-inner product with $z_i^{(0)}$ and after some rearranging we get

$$\begin{aligned} \bar{\alpha}_{jj}\hat{\alpha}_{ji} &= \langle z_i^{(0)}, z_j^{(0)}\rangle_A - \sum_{k=1}^{j-1}\bar{\alpha}_{kj}\hat{\alpha}_{ki} + \langle z_i^{(0)}, \sum_{k=1}^{j}\Delta\delta_j^{(k)}\rangle_A, \\ \sum_{k=1}^{j}\bar{\alpha}_{kj}\hat{\alpha}_{ki} &= \langle z_i^{(0)}, z_j^{(0)}\rangle_A + \langle z_i^{(0)}, \sum_{k=1}^{j}\Delta\delta_j^{(k)}\rangle_A. \end{aligned}$$

In matrix notation this leads to the bound for the strictly lower triangular part of the matrix ΔF (i.e. the strictly upper triangular part of $(\Delta F)^T$)

$$striu(\hat{U}^T\bar{U}) = striu((Z^{(0)})^T AZ^{(0)} + (\Delta F)^T), \|stril(\Delta F)\| \leq O(u)\ \kappa^{1/2}(A)\|A^{1/2}Z^{(0)}\|^2.$$

The matrix $\hat{U}^T$ and the computed upper triangular factor $\bar{U}$ are thus the exact lower and upper triangular factors in the triangular decomposition of the matrix $(Z^{(0)})^T AZ^{(0)}$ perturbed by the nonsymmetric matrix ΔF so that $(Z^{(0)})^T AZ^{(0)} + \Delta F = \hat{U}^T\bar{U}$, where $\|\Delta F\| \leq O(u)\ \kappa(A)^{1/2}\|A\|^{1/2}\|Z^{(0)}\|\|A^{1/2}Z^{(0)}\|$. In addition $\hat{U}$ and $\bar{U}$ have the same diagonal entries, a fact which appears to be very important for further considerations. If we introduce matrices $\Delta\bar{U} = \bar{U} - U$ and $\Delta\hat{U} = \hat{U} - U$ whereas the matrix U is the exact Cholesky factor of the matrix $(Z^{(0)})^T AZ^{(0)} = U^T U$, they

must then satisfy $\Delta F = \Delta\bar{U}^T U + U^T\Delta\hat{U} + \Delta\bar{U}^T\Delta\hat{U}$. Multiplying this identity by U^{-T} and U^{-1} from the left and from the right, respectively, we get by modyfing the approach from [1], [27]

$$U^{-T}\Delta F U^{-1} = U^{-T}\Delta\bar{U}^T + \Delta\hat{U}U^{-1} + U^{-T}\Delta\bar{U}^T\Delta\hat{U}U^{-1}.$$

The matrices $\Delta\bar{U}U^{-1}$ and $\Delta\hat{U}U^{-1}$ are upper triangular. Due to $diag(\bar{U}) = diag(\hat{U})$ their Frobenius norms can be bounded by a system of two inequalities

$$\begin{aligned} 2\|\Delta\bar{U}U^{-1}\|_F &\le \|U^{-T}\Delta F U^{-1}\|_F + \|\Delta\bar{U}U^{-1}\|_F\|\Delta\hat{U}U^{-1}\|_F, \\ 2\|\Delta\hat{U}U^{-1}\|_F &\le \|U^{-T}\Delta F U^{-1}\|_F + \|\Delta\bar{U}U^{-1}\|_F\|\Delta\hat{U}U^{-1}\|_F. \end{aligned}$$

Assuming $\|U^{-T}\Delta F U^{-1}\|_F \ll 1$ we obtain after some manipulation $\|\Delta\bar{U}U^{-1}\|_F \le \|U^{-T}\Delta F U^{-1}\|_F$. Due to $Z^{(0)} + \Delta E^{(1)} = \bar{Z}\bar{U}$ and $(Z^{(0)})^T A Z^{(0)} = U^T U$ we can write

$$\begin{aligned} &(\bar{U}U^{-1})^T(\bar{Z}^T A\bar{Z} - I)(\bar{U}U^{-1}) = \Delta\bar{U}U^{-1} + (\Delta\bar{U}U^{-1})^T + (\Delta\bar{U}U^{-1})^T(\Delta\bar{U}U^{-1}) \\ &+Z^T A\Delta E^{(1)}U^{-1} + (Z^T A\Delta E^{(1)}U^{-1})^T + U^{-T}(\Delta E^{(1)})^T A\Delta E^{(1)}U^{-1}. \end{aligned} \tag{37}$$

Considering that $\|U\bar{U}^{-1}\| \le [1 - \|U^{-T}\Delta F U^{-1}\|]^{-1}$, $\|U^{-T}\Delta F U^{-1}\| \le \|\Delta F\|\|U^{-1}\|^2$ and $\|\Delta F\| \le O(u)\ \kappa(A)^{1/2}\|A\|^{1/2}\|Z^{(0)}\|\|A^{1/2}Z^{(0)}\|$ we finally get the bound for the loss of orthogonality between the computed vectors $\bar{Z}$ having identical form as (30), i.e.

$$\|\bar{Z}^T A\bar{Z} - I\| \le \frac{O(u)\kappa(A)\kappa(A^{1/2}Z^{(0)})\kappa(Z^{(0)})}{1 - O(u)\kappa(A)\kappa(A^{1/2}Z^{(0)})\kappa(Z^{(0)})}. \tag{38}$$

In the case of a diagonal A the bound can be improved by a factor of $\kappa^{1/2}(A)$ and we can get the bound identical to (31). These results clearly indicate that the numerical behavior of CGS and AINV is quite similar and the loss between computed vectors in these two schemes in the worst-case is proportional to $\kappa(A)\kappa(A^{1/2}Z^{(0)})\kappa(Z^{(0)})$ for a general A and to $\kappa^2(A^{1/2}Z^{(0)})$ for a diagonal A.

5 Classical Gram-Schmidt with reorthogonalization

We have shown that the orthogonality between computed vectors $\bar{Z}$ in MGS, CGS and AINV (besides the condition number of A bounding the local errors) depends significantly on the condition number of the matrix $A^{1/2}Z^{(0)}$, while in the implementation based on eigendecomposition we have the bound $\|\bar{Z}^T A\bar{Z} - I\| \le O(u)\|A\|\|\bar{Z}\|^2 \le O(u)\kappa(A)$. In this section we consider the classical Gram-Schmidt algorithm with reorthogonalization (i.e. classical Gram-Schmidt where the orthogonalization of the current vector $z_i^{(0)}$ is performed exactly twice). Provided we have already the vectors $Z_{i-1} = [z_1, \ldots, z_{i-1}]$ at the i-th step we generate the vectors

$$z_i^{(1)} = z_i^{(0)} - \sum_{j=1}^{i-1}\alpha_{ji}^{(1)} z_j = (I - Z_{i-1}Z_{i-1}^T A)z_i^{(0)}, \tag{39}$$

$$z_i^{(2)} = z_i^{(1)} - \sum_{j=1}^{i-1}\alpha_{ji}^{(2)} z_j = (I - Z_{i-1}Z_{i-1}^T A)z_i^{(1)}. \tag{40}$$

The new vector z_i is just the result from the normalization of $z^{(2)}$ given as $z_i = z^{(2)}/\alpha_{ii}$ with $\alpha_{ii} = \|z^{(2)}\|_A$. The new column of the triangular factor is given by elements $\alpha_{ji} = \alpha_{ji}^{(1)} + \alpha_{ji}^{(2)}$. It is clear that in exact arithmetic $z_i^{(2)} = z_i^{(1)}$, while the computed vectors satisfy the following identities

$$\bar{z}_i^{(1)} = z_i^{(0)} - \sum_{j=1}^{i-1} \bar{\alpha}_{ji}^{(1)} \bar{z}_j + \Delta\delta_i^{(1)}, \|\Delta\delta_i^{(1)}\| \leq O(u)(\|z_i^{(0)}\| + \sum_{j=1}^{i-1} |\bar{\alpha}_{ji}^{(1)}| \|\bar{z}_j\|), \quad (41)$$

$$\bar{z}_i^{(2)} = \bar{z}_i^{(1)} - \sum_{j=1}^{i-1} \bar{\alpha}_{ji}^{(2)} \bar{z}_j + \Delta\delta_i^{(2)}, \|\Delta\delta_i^{(2)}\| \leq O(u)(\|\bar{z}_i^{(1)}\| + \sum_{j=1}^{i-1} |\bar{\alpha}_{ji}^{(2)}| \|\bar{z}_j\|), \quad (42)$$

with local errors that can be further bounded as $\|\Delta\delta_i^{(1)}\|_A \leq \|A\|^{1/2} \|\Delta\delta_i^{(1)}\| \leq O(u)\kappa^{1/2}(A)\|z_i^{(0)}\|_A$ and $\|\Delta\delta_i^{(1)}\|_A \leq O(u)\kappa^{1/2}(A)\|\bar{z}_i^{(1)}\|_A \leq O(u)\kappa^{1/2}(A)\|z_i^{(0)}\|_A$ due to the (near-) monotonicity $\|\bar{z}_i^{(1)}\|_A \lesssim \|z_i^{(0)}\|_A$. The recurrences (41)-(42) can be rewritten as $z_i^{(0)} + \Delta\delta_i^{(1)} + \Delta\delta_i^{(2)} = \sum_{j=1}^{i-1} (\bar{\alpha}_{ji}^{(1)} + \bar{\alpha}_{ji}^{(2)}) \bar{z}_j + \bar{z}_i^{(2)}$. The vector $\bar{z}_i^{(2)}$ can be written as $\bar{z}_i^{(2)} = \bar{\alpha}_{ii}\bar{z}_i + \Delta\delta_i^{(i)}$. In matrix form this gives the identity for the first i vectors stored as columns of $\bar{Z}_i = [\bar{z}_1, \ldots, \bar{z}_i]$

$$Z_i^{(0)} + \Delta E_i^{(1)} = \bar{Z}_i(\bar{U}_i^{(1)} + \bar{U}_i^{(2)}), \|A^{1/2}\Delta E_i^{(1)}\| \leq O(u)\kappa^{1/2}(A)\|A^{1/2}Z^{(0)}\|. \quad (43)$$

It is clear that the computed coefficients $\bar{\alpha}_{ji}^{(1)}$ and $\bar{\alpha}_{ji}^{(2)}$ satisfy $|fl[\langle z_i^{(0)}, \bar{z}_j\rangle_A] - \langle z_i^{(0)}, \bar{z}_j\rangle_A| \leq O(u)\|A\|\|z_i^{(0)}\|\|\bar{z}_j\|$ and $|fl[\langle \bar{z}_i^{(1)}, \bar{z}_j\rangle_A] - \langle \bar{z}_i^{(1)}, \bar{z}_j\rangle_A| \leq O(u)\|A\|\|\bar{z}_i^{(1)}\|\|\bar{z}_j\|$ leading to reformulation of the recurrences (41)-(42) in the form

$$\begin{aligned} \bar{z}_i^{(1)} &= (I - \bar{Z}_{i-1}\bar{Z}_{i-1}^T A) z_i^{(0)} + \Delta\eta_i^{(1)}, & (44)\\ \|\Delta\eta_i^{(1)}\|_A &\leq O(u)\ \kappa^{1/2}(A)\|A\|^{1/2}\|\bar{Z}_{i-1}\|\|z_i^{(0)}\|_A, \\ \bar{z}_i^{(2)} &= (I - \bar{Z}_{i-1}\bar{Z}_{i-1}^T A)\bar{z}_i^{(1)} + \Delta\eta_i^{(2)}, & (45)\\ \|\Delta\eta_i^{(2)}\|_A &\leq O(u)\kappa^{1/2}(A)\|A\|^{1/2}\|\bar{Z}_{i-1}\|\|\bar{z}_i^{(1)}\|_A. \end{aligned}$$

We will use an incremental approach and assume at step $i-1$ that the loss of orthogonality between the vectors $\bar{z}_1, \ldots, \bar{z}_{i-1}$ is bounded by

$$\|\bar{Z}_{i-1}^T A \bar{Z}_{i-1}\| \leq O(u)\kappa^{1/2}(A)\|A\|^{1/2}\|\bar{Z}_{i-1}\| \leq O(u)\kappa(A) \quad (46)$$

and show that this statement will remain true also at step i. Multiplication of (44) from the left by $\bar{Z}_{i-1}A^T$ leads to the identity

$$\bar{Z}_{i-1}^T A \bar{z}_i^{(1)} = (I - \bar{Z}_{i-1}^T A \bar{Z}_{i-1})\bar{Z}_{i-1}^T A z_i^{(0)} + \bar{Z}_{i-1}^T A \Delta\eta_i^{(1)}.$$

Taking the norm, dividing by the A-norm of the vector $\bar{z}_i^{(1)}$ and taking into account (44) and (46) leads to the bound for the quantity $\|\bar{Z}_{i-1}^T A(\bar{z}_i^{(1)}/\|\bar{z}_i^{(1)}\|_A)\|$

$$\frac{\|\bar{Z}_{i-1}^T A \bar{z}_i^{(1)}\|}{\|\bar{z}_i^{(1)}\|_A} \leq O(u)\kappa^{1/2}(A)\|A\|^{1/2}\|\bar{Z}_{i-1}\|\|A^{1/2}\bar{Z}_{i-1}\| \frac{\|z_i^{(0)}\|_A}{\|\bar{z}_i^{(1)}\|_A}. \quad (47)$$

The factor $\|z_i^{(0)}\|_A/\|\bar{z}_i^{(1)}\|_A$ can be estimated using the recurrence (43) for the first $i-1$ steps together with (41) which can be written after multiplication by $A^{1/2}$ from the left in the form

$$A^{1/2}Z_i^{(0)} + A^{1/2}[\Delta E_{i-1}^{(1)}, \Delta\delta_i^{(1)} - \bar{z}_i^{(1)}] = A^{1/2}\bar{Z}_{i-1}[\bar{U}_{i-1}^{(1)} + \bar{U}_{i-1}^{(2)}, \bar{u}_i^{(1)}],$$

where $[\bar{U}_{i-1}^{(1)} + \bar{U}_{i-1}^{(2)}, \bar{u}_i^{(1)}]$ is the $(i-1)\times i$ matrix that contains the sums of computed coefficients (at step i we consider only the first sweep of the algorithm). The matrix $A^{1/2}Z_i^{(0)} + A^{1/2}[\Delta E_{i-1}^{(1)}, \Delta\delta_i^{(1)} - \bar{z}_i^{(1)}]$ has rank $i-1$ and the matrix $A^{1/2}Z_i^{(0)}$ has full column rank. Therefore the distance from $A^{1/2}Z_i^{(0)}$ to the set of matrices having rank $i-1$ is less than the norm of $A^{1/2}[\Delta E_{i-1}^{(1)}, \Delta\delta_i^{(1)} - \bar{z}_i^{(1)}]$. Indeed the minimal singular value of $A^{1/2}Z_i^{(0)}$ can be then bounded by the Frobenius norm of the perturbation which can be bounded further as

$$\sigma_i(A^{1/2}Z_i^{(0)}) \le \sqrt{\|A^{1/2}\Delta E_{i-1}^{(1)}\|^2 + \|\Delta\delta_i^{(1)}\|_A^2) + \|\bar{z}_i^{(1)}\|_A^2}.$$

Using the bounds from (41) and (43) and assuming that $O(u)\kappa^{1/2}(A)\kappa(A^{1/2}Z_i^{(0)}) < 1$ we can give a lower bound for the A-norm of the vector $\bar{z}_i^{(1)}$

$$\|\bar{z}_i^{(1)}\|_A \ge \sigma_i(A^{1/2}Z_i^{(0)})\left(1 - O(u)\kappa^{1/2}(A)\kappa(A^{1/2}Z_i^{(0)})\right). \tag{48}$$

Due to $\|A\|^{1/2}\|\bar{Z}_{i-1}\| \le \kappa^{1/2}(A)$ and (48) for the left-hand side of (47) we get

$$\frac{\|\bar{Z}_{i-1}^T A\bar{z}_i^{(1)}\|}{\|\bar{z}_i^{(1)}\|_A} \le \frac{O(u)\kappa(A)\kappa(A^{1/2}Z_i^{(0)})}{1 - O(u)\kappa^{1/2}(A)\kappa(A^{1/2}Z_i^{(0)})}. \tag{49}$$

Similarly as before we consider (45), multiply it from the left by $\bar{Z}_{i-1}^T A$ and obtain the identity

$$\bar{Z}_{i-1}^T A\bar{z}_i^{(2)} = (I - \bar{Z}_{i-1}^T A\bar{Z}_{i-1})\bar{Z}_{i-1}^T A\bar{z}_i^{(1)} + \bar{Z}_{i-1}^T A\Delta\eta_i^{(2)}.$$

which is treated similarly as in (50), i.e. using (46) and (45) we get

$$\frac{\|\bar{Z}_{i-1}^T A\bar{z}_i^{(2)}\|}{\|\bar{z}_i^{(2)}\|_A} \le O(u)\kappa^{1/2}(A)\|A\|^{1/2}\|\bar{Z}_{i-1}\|\|A^{1/2}\bar{Z}_{i-1}\|\|\frac{\|\bar{z}_i^{(1)}\|_A}{\|\bar{z}_i^{(2)}\|_A}. \tag{50}$$

The factor $\|\bar{z}_i^{(1)}\|_A/\|\bar{z}_i^{(2)}\|_A$ can be bounded from below reconsidering (45) once again as follows

$$\frac{\|\bar{z}_i^{(2)}\|_A}{\|\bar{z}_i^{(1)}\|_A} \ge \frac{\|\bar{z}_i^{(1)}\|_A}{\|\bar{z}_i^{(1)}\|_A} - \|A^{1/2}\bar{Z}_{i-1}\|\frac{\|\bar{Z}_{i-1}^T A\bar{z}_i^{(1)}\|}{\|\bar{z}_i^{(1)}\|_A} - \frac{\|\Delta\eta_i^{(2)}\|_A}{\|\bar{z}_i^{(1)}\|_A}$$

which under stronger assumption $O(u)\kappa(A)\kappa(A^{1/2}Z_i^{(0)}) < 1$ leads to the final bound $\|\bar{z}_i^{(1)}\|_A/\|\bar{z}_i^{(2)}\|_A \le [1 - O(u)\kappa(A)\kappa(A^{1/2}Z_i^{(0)})]^{-1}$. Considering $\bar{z}_i^{(2)} = \bar{\alpha}_{ii}\bar{z}_i + \Delta\delta_i^{(i)}$ with $|\bar{\alpha}_{ii} - \|\bar{z}_i^{(2)}\|_A| \le O(u)\|A\|\|\bar{z}_i^{(2)}\|^2$ we can relate the left-hand side of (50) with the quantity $\|\bar{Z}_{i-1}^T A\bar{z}_i\|$. Taking into account also the error from the normalization $|1 - \|\bar{z}_i\|_A^2| \le O(u)\|A\|\|\bar{z}_i\|^2$ we end up with the statement

$$\|I - \bar{Z}_i^T A\bar{Z}_i\| \le O(u)\kappa^{1/2}(A)\|A\|^{1/2}\|\bar{Z}_i\| \le O(u)\kappa(A). \tag{51}$$

Again, the bound (51) can be significantly improved for diagonal A. Assuming only $O(u)\kappa(A^{1/2}Z_i^{(0)}) < 1$ one can show using the same approach that the orthogonality between the computed vectors $\bar{Z}_i$ in this case does not depend on the condition number of the matrix $A^{1/2}Z^{(0)}$ and it is preserved on the roundoff unit level with $\|I - \bar{Z}_i^T A\bar{Z}_i\| \leq O(u)$. It is interesting to note that a similar result holds also for the EIG implementation based on backward stable eigendecomposition. However, the assumption $O(u)\kappa(A^{1/2}Z_i^{(0)}) < 1$ is crucial for the CGS2 algorithm, while for EIG this result holds without any requirement on the initial vectors stored in $Z^{(0)}$. In practical situations, both EIG and CGS2 behave very similarly as it is also illustrated in our numerical examples.

6 Numerical experiments

In this section we illustrate our theoretical results. All experiments are performed using MATLAB with $u = 1.1\ 10^{-16}$. We consider three sequences of test examples A_i with increasing condition number $\kappa(A_i) \approx 10^i, i = 0, \ldots, 15$ and show that our bounds for the error in factorization and the loss of orthogonality are realistic. In all the figures we depict the loss of orthogonality $\| I - \bar{Z}_i^T A_i \bar{Z}_i \|$ and the 2-norm of the error of the factorization $\| Z_i^{(0)} - \bar{Z}_i \bar{U}_i \|$ with respect to the condition number $\kappa(A_i)$ for the eigenvalue decomposition-based (EIG) implementation (solid lines), the modified Gram-Schmidt (MGS) algorithm (dashed lines with bold dots), the classical Gram-Schmidt (CGS) algorithm (dash-dotted lines), the AINV orthogonalization (dashed lines) and the classical Gram-Schmidt algorithm with reorthogonalization (CGS2, solid lines with bold dots). The dotted lines in figures always correspond to relevant bounds (i.e. $u\kappa(A)$, $u\kappa(A)\kappa(A^{1/2}Z^{(0)})$ and $u\kappa(A)\kappa(A^{1/2}Z^{(0)})\kappa(Z^{(0)})$ for the loss of orthogonality and $u\|Z^{(0)}\|$, $u\|\bar{Z}\|\|A^{1/2}Z^{(0)}\|$ and $u\kappa^{1/2}(A)\|Z^{(0)}\|$ for the error of the factorization). For all Gram-Schmidt algorithms we consider two computational variants which differ only in the floating-point evaluation of the inner product induced by the matrix A. The variant „a" assumes that in the i-th step we first compute matrix-vector multiplication with the first argument of the inner product and then compute the dot product with the second argument (e.g. in MGS we compute the coefficient as $\mathrm{fl}(\langle \bar{z}_i^{(j-1)}, \mathrm{fl}(A\bar{z}_j)\rangle)$). The variant "b" assumes the reverse order of the computation (in MGS it corresponds to $\mathrm{fl}(\langle \mathrm{fl}(Az_i^{(j-1)}), \bar{z}_j\rangle)$).

The first sequence of matrices A_i with dimension $n = 8$ (denoted as Problem 1) is generated as powers of the inverse Hilbert matrix $A = \mathrm{invhilb}(8) = V\Lambda V^T$ ($\kappa(A) \approx 10^{10}$) such that $A_i = V\Lambda^{i/10}V^T$ with $\kappa(A_i) \approx 10^i, i = 0, \ldots, 15$. The matrix $Z_i^{(0)}$ is constructed as $Z_i^{(0)} = V\Lambda^{-i/20}(L_i)^T$, where L_i is the Cholesky factor of A_i satisfying $A_i = L_i(L_i)^T$ so that $\kappa(Z_i^{(0)}) \approx 10^i$ and $\kappa(A_i^{1/2}Z_i^{(0)}) \approx 10^{i/2}$ for $i = 0, \ldots, 15$. It is clear from the definition that in exact arithmetic the orthogonal factor is equal to $Z_i = V\Lambda^{-i/20}$ and the triangular factor is identical to the transpose of the Cholesky factor $U_i = (L_i)^T$. Moreover we assume that the columns of Z_i are ordered with respect to increasing eigenvalue of A.

It is clear from Figure 1 that the loss of orthogonality between computed vectors in the EIG implementation is proportional to $\kappa(A)$ and the same applies to the CGS2 algorithm. The behavior of CGS and AINV is very similar; they both generate vectors with loss of orthogonality approaching the theoretical bound $O(u)\kappa(A)\kappa(A^{1/2}Z^{(0)})\kappa(Z^{(0)})$

as predicted by the theory. Note that while there is no visible difference between the ,,a" and ,,b" variants in CGS, AINV or CGS2, the situation can be completely different for the MGS algorithm. Indeed the loss of orthogonality for the version ,,a" is very close to EIG and CGS2, while for the version ,,b" it approaches the theoretical bound $\kappa(A)\kappa(A^{1/2}Z^{(0)})$. We believe that this observation can be explained by our particular construction of the matrix $Z^{(0)}$. While in the first case we compute $A\bar{z}_j$ which in exact arithmetic should be $\lambda_{n-j+1}z_j$ and λ_{n-j+1} is the j-th smallest eigenvalue of A, in the second case one multiplies $z_i^{(j-1)}$ which is a combination of eigenvectors corresponding to $n-j$ largest eigenvalues. Due to the fact that the error in the dot product is proportional to the size of arguments, one can expect for Problem 1 (and also for Problem 2) more accurate results in the ,,a" version of the MGS algorithm. Figure 2 shows the 2-norm of the error in the factorization measured by $\| Z_i^{(0)} - \bar{Z}_i\bar{U}_i \|$. Since the behavior of all schemes does not differ significantly for ,,a" or ,,b" versions, we consider only the version ,,b" here. The results confirm that the EIG implementation is significantly worse in terms of the error and it approximately scales as $O(u)\kappa^{1/2}(A)\|Z^{(0)}\|$. All the other algorithms behave similarly and correspond to the significantly better bound $O(u)(\|Z^{(0)}\| + \|\bar{Z}\|\|A^{1/2}Z^{(0)}\|)$.

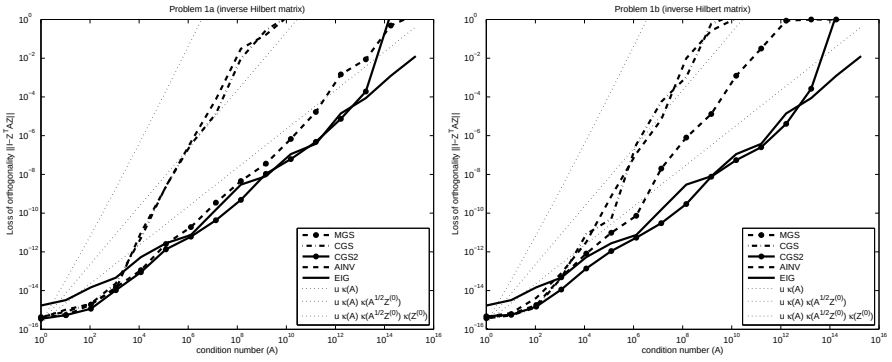

Fig. 1 Loss of orthogonality $\|\bar{Z}^T A\bar{Z} - I\|$ for Problem 1 (versions ,,a" and ,,b")

The definition of Problem 2 uses the Hilbert matrix of the dimension $n = 8$: $A = V\Lambda V^T$ ($\kappa(A) \approx 10^{10}$). The sequence of the matrices A_i was defined through $A_i = V\Lambda^{i/10}V^T$ with $\kappa(A_i) \approx 10^i, i = 0, \ldots, 15$. Again, we set $Z_i^{(0)} = V\Lambda^{-i/20}(L_i)^T$, where L_i stands now for the Cholesky factor of the matrix $A_i^3 = L_iL_i^T$. As we see from Figure 3 the results are qualitatively similar, but more ill-conditioned triangular factors U_i lead to significantly weaker orthogonality of computed vectors for CGS, AINV and MGS algorithms close to our theoretical bounds, whereas the error in the factorization behaves similarly as in Problem 1. It is also apparent from Figure 3 that the loss of orthogonality in the CGS2 algorithm can be significantly different from the loss of orthogonality in EIG. This may however happen only when $O(u)\kappa(A_i)\kappa(A_i^{1/2}Z_i^{(0)}) \approx 1$, see the second dotted line on Figure 3.

Finally we investigate the behavior of all five schemes in the case of a sequence of diagonal matrices A_i. As above, the dimension of the problem is $n = 8$, with the constant $Z_i^{(0)} = A^{1/2}$, where A is the inverse Hilbert matrix $\left(\kappa(Z^{(0)}) \approx 10^5\right)$ and $A^{(i)}$

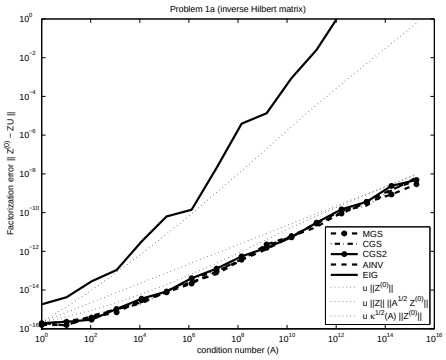

Fig. 2 Factorization error $\|Z^{(0)} - \bar{Z}\bar{U}\|$ for Problem 1 (version ,,b")

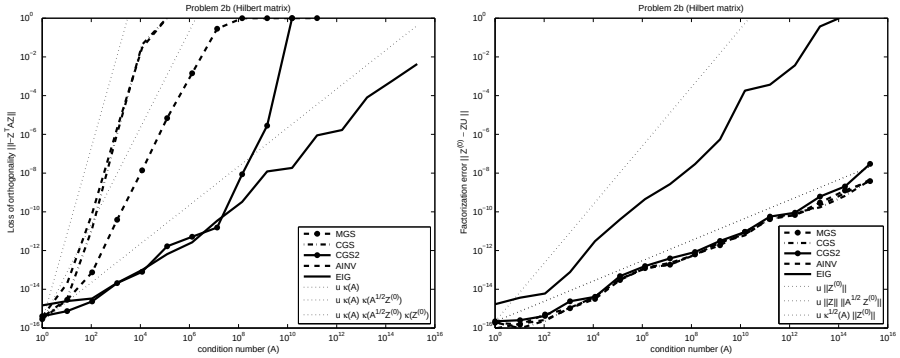

Fig. 3 Loss of orthogonality $\|\bar{Z}^T A\bar{Z} - I\|$ and factorization error $\|Z^{(0)} - \bar{Z}\bar{U}\|$ for Problem 2 (version ,,b")

is a diagonal matrix with $\kappa(A_i) \approx 10^i, i = 0, \ldots, 15$. The results are plotted on Figure 4 and clearly illustrate that all our theoretical bounds are tight. The only exception is that the factorization error seems to be independent of $\|\bar{Z}\|\|\bar{U}\|$, but we do not see how to prove that $\|Z^{(0)} - \bar{Z}\bar{U}\| \leq O(u)\|Z^{(0)}\|$ holds for a diagonal A. This would actually complete our analysis with the conclusion that the numerical behavior of the weighted Gram-Schmidt orthogonalization is similar to the numerical behavior of the Gram-Schmidt orthogonalization with the standard inner product.

7 Conclusions

In this paper we have presented several theoretical results on the factorization error and orthogonality of vectors computed by the most important schemes used for orthogonalization with respect to the non-standard inner product. Although they are mathematically equivalent, their numerical behavior in finite precision arithmetic may significantly differ. Our main results are summarized in Table 1. It follows for our analysis that while the factorization error is quite comparable for all these schemes (with exception of the EIG implementation), the orthogonality between computed vectors can

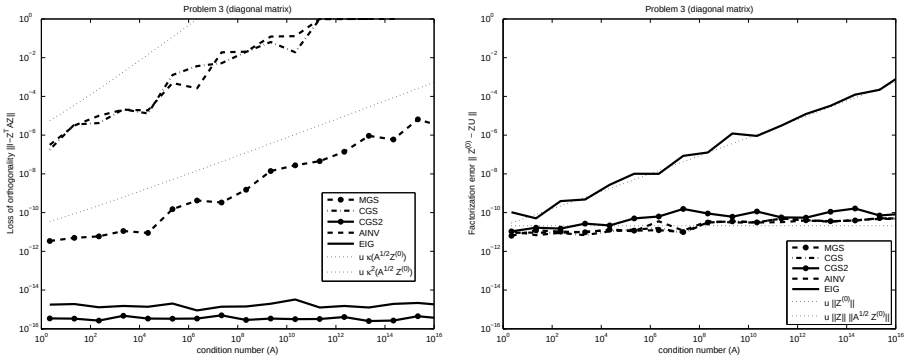

Fig. 4 Loss of orthogonality $\|\bar{Z}^T A\bar{Z} - I\|$ and factorization error $\|Z^{(0)} - \bar{Z}\bar{U}\|$ for Problem 3

be significantly lost and it depends linearly on the conditioning of the matrix inducing the inner product. This is the case also for the eigenvalue-based implementation and Gram-Schmidt with reorthogonalization. The classical Gram-Schmidt algorithm and AINV orthogonalization behave very similarly and compute vectors with the orthogonality that depends besides $\kappa(A)$ also on the factor $\kappa(A^{1/2}Z^{(0)})\kappa(Z^{(0)})$ essentially meaning the quadratic dependence on the condition number of the matrix $A^{1/2}Z^{(0)}$. Since the orthogonality in the modified Gram-Schmidt algorithm depends only linearly on $\kappa(A^{1/2}Z^{(0)})$ this algorithm appears to be a good compromise between expensive EIG or CGS2 and less accurate CGS or AINV. Indeed in the context of approximate inverse preconditioning the stabilization of AINV has lead to the SAINV algorithm which uses exactly the MGS orthogonalization. We have treated also the particular case of a diagonal A which is extremely useful for the context of weighted least squares problems. It appears then that local errors arising in the computation of non-standard inner products do not play an important role and numerical behavior of these schemes is similar to the behavior of the orthogonalization with the standard inner product.
The authors would like to thank for the fruitful discussion and useful comments to G. Meurant and S. Gratton.

References

1. Barrlund, A.: Perturbation bounds for the LDL and LU decompositions. BIT, **31**(2), 358–363 (1991)
2. Benzi, M.: Preconditioning techniques for large linear systems: a survey. J. Comput. Phys., **182**(2), 418–477 (2002)
3. Benzi, M., Cullum, J.K., and Tůma, M.: Robust approximate inverse preconditioning for the conjugate gradient method. SIAM J. Sci. Comput., **22**(4), 1318–1332 (2000)
4. Benzi, M., Meyer, C.D., and Tůma, M.: A sparse approximate inverse preconditioner for the conjugate gradient method. SIAM J. Sci. Comput., **17**(5), 1135–1149 (1996)
5. Benzi, M. and Tůma, M.: A robust incomplete factorization preconditioner for positive definite matrices. Numer. Linear Algebra Appl., **10**(5-6), 385–400 (2003)
6. Björck, Å.: Solving linear least squares problems by Gram-Schmidt orthogonalization. BIT, **7**(1), 1–21 (1967)
7. Björck, Å.: Numerical methods for Least Squares Problems. SIAM, Philadelphia, PA (1996)
8. Challacombe, M.: A simplified density matrix minimization for linear scaling self-consistent field theory. J. Chem. Phys., **110**(5), 2332–2342 (1999)

Table 1 Comparison of upper bounds on the loss of orthogonality and factorization error for different orthogonalization schemes

Eigendecomposition based orthogonalization		
Assumption	$\|I - \bar{Z}^T A\bar{Z}\|$	$\|Z^{(0)} - \bar{Z}\bar{U}\|$
general A	$O(u)\kappa(A)$	$O(u)\kappa^{\frac{1}{2}}(A)\|Z^{(0)}\|$
diagonal A	$O(u)$	$O(u)\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$
Classical Gram-Schmidt with reorthogonalization		
Assumption	$\|I - \bar{Z}^T A\bar{Z}\|$	$\|Z^{(0)} - \bar{Z}\bar{U}\|$
general A $O(u)\kappa(A)\kappa(A^{\frac{1}{2}}Z^{(0)}) < 1$	$O(u)\kappa(A)$	$O(u)\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$
diagonal A $O(u)\kappa(A^{\frac{1}{2}}Z^{(0)}) < 1$	$O(u)$	$O(u)\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$
Modified Gram-Schmidt orthogonalization		
Assumption	$\|I - \bar{Z}^T A\bar{Z}\|$	$\|Z^{(0)} - \bar{Z}\bar{U}\|$
general A $O(u)\kappa(A)\kappa(A^{\frac{1}{2}}Z^{(0)}) < 1$	$\frac{O(u)\kappa(A)\kappa(A^{\frac{1}{2}}Z^{(0)})}{1-O(u)\kappa(A)\kappa(A^{\frac{1}{2}}Z^{(0)})}$	$O(u)\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$
diagonal A $O(u)\kappa(A^{\frac{1}{2}}Z^{(0)}) < 1$	$\frac{O(u)\kappa(A^{\frac{1}{2}}Z^{(0)})}{1-O(u)\kappa(A^{\frac{1}{2}}Z^{(0)})}$	$O(u)\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$
Classical Gram-Schmidt and AINV orthogonalizations		
Assumption	$\|I - \bar{Z}^T A\bar{Z}\|$	$\|Z^{(0)} - \bar{Z}\bar{U}\|$
general A $O(u)\kappa(A)\kappa(A^{\frac{1}{2}}Z^{(0)})\kappa(Z^{(0)}) < 1$	$\frac{O(u)\kappa(A)\kappa(A^{\frac{1}{2}}Z^{(0)})\kappa(Z^{(0)})}{1-O(u)\kappa(A)\kappa(A^{\frac{1}{2}}Z^{(0)})\kappa(Z^{(0)})}$	$O(u)\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$
diagonal A $O(u)\kappa^2(A^{\frac{1}{2}}Z^{(0)}) < 1$	$\frac{O(u)\kappa^2(A^{\frac{1}{2}}Z^{(0)})}{1-O(u)\kappa^2(A^{\frac{1}{2}}Z^{(0)})}$	$O(u)\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$

9. Duff, I.S., Grimes, R.G., and Lewis, J.G.: The Rutherford-Boeing Sparse Matrix Collection. Report RAL-TR-97-031, Rutherford Appleton Laboratory, Chilton, Oxfordshire, England (1997)
10. Fox, L., Huskey, H.D., and Wilkinson, J.H.: Notes on the solution of algebraic linear simultaneous equations. Quart. J. Mech. and Appl. Math., **1**(1), 149–173 (1948)
11. Giraud, L., Langou, J., and Rozložník, M.: The loss of orthogonality in the Gram-Schmidt orthogonalization process. Comput. Math. Appl., **50**(7), 1069–1075 (2005)
12. Giraud, L., Langou, J., Rozložník, M., and van den Eshof, J.: Rounding error analysis of the classical Gram-Schmidt orthogonalization process. Numer. Math. **101**(1), 97–100 (2005)
13. Golub, G.H. and Van Loan, C.F.: Matrix computations. Johns Hopkins Studies in the Mathematical Sciences, Third Edition. Johns Hopkins University Press, Baltimore, MD (1996)
14. Gulliksson, M.: Backward error analysis for the constrained and weighted linear least squares problem when using the weighted QR factorization. SIAM J. Matrix Anal. Appl.,

16(2), 675–687 (1995)
15. Gulliksson, M.: On the modified Gram-Schmidt algorithm for weighted and constrained linear least squares problems. BIT, **35**(4), 453–468 (1995)
16. Gulliksson, M. and Wedin, P.-Å.: Modifying the QR-decomposition to constrained and weighted linear least squares. SIAM J. Matrix Anal. Appl., **13**(4), 1298–1313 (1992)
17. Higham, N.J.: Accuracy and stability of numerical algorithms. Second edition. SIAM, Philadelphia, PA (2002)
18. Hestenes, M.R.: Inversion of matrices by biorthogonalization and related results. J. SIAM, **6**(1), 51–90 (1958)
19. Hestenes, M.R. and Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Nat. Bur. Standards, **49**(6), 409–435 (1952)
20. Householder, A.S.: Terminating and nonterminating iterations for solving linear systems. J. SIAM, **3**(2), 67–72 (1955)
21. Kharchenko, S.A., Kolotilina, L.Y., Nikishin, A.A., and Yeremin, A.Y.: A robust AINV-type method for constructing sparse approximate inverse preconditioners in factored form. Numer. Linear Algebra Appl., **8**(3), 165–179 (2001)
22. Lawson, C.L., and Hanson, R.J.: Solving least squares problems. Prentice-Hall Series in Automatic Computation. Prentice-Hall, Inc., Englewood Cliffs, N.J. (1974)
23. Martin, R.S. and Wilkinson, J.H.: Handbook Series Linear Algebra: Reduction of the symmetric eigenproblem $Ax = \lambda Bx$ and related problems to standard form. Numer. Math., **11**(2), 99–110 (1968)
24. Mazzia, A. and Pini, G.: Numerical performance of preconditioning techniques for the solution of complex sparse linear systems. Comm. Num. Meth. Engineering, **19**(1), 37–48 (2003)
25. Morris, J.: An escalator process for the solution of linear simultaneous equations. Philos. Mag., **37**(7), 106–120 (1946)
26. Saberi Najafi, H. and Ghazvini, H.: Weighted restarting method in the weighted Arnoldi algorithm for computing the eigenvalues of a nonsymmetric matrix. Appl. Math. Comput., **175**(2), 1276–1287 (2006)
27. Sun, J.-G.: Perturbation bounds for the Cholesky and QR factorizations. BIT, **31**, 341–352 (1991)
28. Parlett, B.N.: The Symmetric Eigenvalue Problem. Prentice-Hall Series in Computational Mathematics. Prentice-Hall, Inc., Englewood Cliffs, N.J. (1980)
29. Pietrzykowski, T.: Projection method. Prace ZAM Ser. A, No. 8, 9, (1960)
30. Purcell, E.W.: The vector method of solving simultaneous linear equations. J. Math. Phys., **32**, 150–153 (1953)
31. Smoktunowicz, A., Barlow, J.L., and Langou, J.: A note on the error analysis of classical Gram-Schmidt. Numer. Math., **105**(2), 299–313 (2006)
32. Wilkinson, J.H.: The algebraic eigenvalue problem. Clarendon Press, Oxford (1965)
33. Yin, J.-F., Yin, G.-J., and Ng, M.: On adaptively accelerated Arnoldi method for computing PageRank. Preprint (2010)