BIT Numer Math (2012) 52:1035–1058
DOI 10.1007/s10543-012-0398-9

BIT

Numerical stability of orthogonalization methods with a non-standard inner product

Miroslav Rozložník · Miroslav Tůma · Alicja Smoktunowicz · Jiří Kopal

Received: 18 September 2011 / Accepted: 11 April 2012 / Published online: 25 August 2012

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.

Communicated by Michiel Hochstenbach.

The work of M. Rozložník and M. Tůma was supported by Grant Agency of the Czech Republic under the project 108/11/0853 and by the Grant Agency of the Academy of Sciences of the Czech Republic under the project IAA100300802. The work of J. Kopal was supported by the Ministry of Education of the Czech Republic under the project no. 7822/115.

M. Rozložník (✉) · M. Tůma
Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
e-mail: miro@cs.cas.cz

M. Tůma
e-mail: tuma@cs.cas.cz

A. 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

J. Kopal
Institute of Novel Technologies and Applied Informatics, Technical University of Liberec, Hálkova 6, 461 17 Liberec, Czech Republic
e-mail: jiri.kopal@tul.cz

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

Mathematics Subject Classification (2010) 65F25 · 65G50 · 15A23

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

$$\|U\| = \|A^{1/2} Z^{(0)}\| \leq \|A\|^{1/2} \|Z^{(0)}\|,$$
$$\sigma_n(U) = \sigma_n\big(A^{1/2} Z^{(0)}\big) \geq \lambda_m^{1/2}(A) \sigma_n\big(Z^{(0)}\big)$$

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

$$\begin{aligned} \|Z\| &\leq \|A^{-1/2}\| = \|A^{-1}\|^{1/2}, \\ \sigma_n(Z) &\geq \lambda_m^{1/2}\big(A^{-1}\big) = \|A\|^{-1/2}. \end{aligned} \tag{1.1}$$

Indeed, it follows from (1.1) that $\kappa(Z) \leq \kappa^{1/2}(A)$. Since $Z = Z^{(0)} U^{-1}$ the product ZZ^T can be written as $ZZ^T = Z^{(0)}[(Z^{(0)})^T AZ^{(0)}]^{-1}(Z^{(0)})^T$. 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 $n = m$ 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 $Z = U^{-1}$ is its triangular inverse satisfying $\kappa(Z) = \kappa(U)$.

The idea of computing the A-orthogonal vectors is heavily used in many applications. One of the important preconditioning classes involves computing an approximate inverse factorization such that ZZ^T is a sparse approximation of A^{-1}. Many practical schemes with incomplete factorizations of this type have been proposed and are widely used [6]. Another well-known examples are the symmetric definite generalized eigenvalue problem which can be transformed using such factors into the standard eigenproblem (in the well-conditioned case) [15, 25, 37], nonsymmetric eigenvalue problem [28], and the generalized least squares problems [10, 36] including the weighted least squares problem [18, 24] as a particular case. Sparse implementations of generalized orthogonalization schemes are efficiently used in linear scaling electronic theory [11], information retrieval [38] or solving complex systems from quantum chemistry or systems arising from Helmholtz's equation [26].

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 [16]. 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 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 [6] 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 [12, 19, 20, 22, 27, 31, 32]. 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 [6] 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 [5]) and in terms of appropriate computation of diagonal entries in U (one-sided versus stabilized versions of AINV [5, 7, 23]).

For the case of standard inner product $A = I$ there exists a complete rounding error analysis for all main orthogonalization schemes including Householder or Givens QR factorization and Gram–Schmidt process [8, 13, 14, 21, 33]. As for the Householder or Givens QR, we refer to classical results of J. Wilkinson, see the corresponding chapters in books [21, 37]. Regarding the Gram–Schmidt orthogonalization, the reader may first read the survey paper [9], but a real insight can be obtained from the fundamental paper of Å. Björck, who gave a bound for the loss of orthogonality in the MGS algorithm in terms of the condition number $\kappa(Z^{(0)})$. The proof of this result represents an important building block also in our analysis. For the rounding error analysis of CGS and in particular for its relation to the cross product matrix $(Z^{(0)})^T Z^{(0)}$, one should get acquainted with the results in [14, 33]. Several results on the numerical behavior of the Gram–Schmidt process with reorthogonalization can be found in [1, 9, 30]; the relevant sources are [2, 14] here.

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 especially be interested at the factorization error $Z^{(0)} - \bar{Z}\bar{U}$, the error in computing the Cholesky factor $(Z^{(0)})^T A Z^{(0)} - \bar{U}^T \bar{U}$ and, most importantly, the loss of A-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 $A^{-1} - \bar{Z}\bar{Z}^T$. We will formulate them mainly in terms of quantities proportional to the roundoff unit, 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 $(Z^{(0)})^T A Z^{(0)} \approx \bar{U}^T \bar{U}$. 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 [4].

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 $\sigma_k(X)$ stands for its kth largest singular value and if X has full column rank then $\kappa(X) = \sigma_1(X)/\sigma_n(X)$ is the condition number of the matrix X. The notation $|X|$ stands for the absolute value of the matrix X; $\|X\| = \sigma_1(X)$ is its 2-norm; $|x|$ is the absolute value of the vector x and $\|x\|$ is 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. The kth largest eigenvalue of such matrix A is denoted by $\lambda_k(A)$. For distinction with their exact arithmetic counterparts, we denote quantities computed in finite precision arithmetic using an extra upperbar. We assume the standard model for floating-point computations, and use also the notation $\mathrm{fl}[\cdot]$ for the computed result of some algebraic expression. The unit roundoff is denoted by u. The terms $\mathcal{O}(m, n)u$ 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 proportional to higher powers of u and also occasionally skip the technical details that would negatively affect the presentation of our results.

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 Sect. 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 Sect. 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 Sect. 2.

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 QR factorization $\Lambda^{1/2}V^TZ^{(0)} = QU$. The factor U is thus the triangular factor from the standard 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—see our further comments on the factorization error).

Theorem 2.1 *Let the eigenvalues $\bar{\Lambda}$ and eigenvectors $\bar{V}$ be computed by some backward stable algorithm applied to the matrix A; let $\bar{Q}$ and $\bar{U}$ be the orthogonal and triangular factors computed by the backward stable QR factorization of the matrix* $\mathrm{fl}[\bar{\Lambda}^{1/2}\bar{V}^TZ^{(0)}]$ *and* $\bar{Z} = \mathrm{fl}[\bar{V}\bar{\Lambda}^{-1/2}\bar{Q}]$. *Then the computed factors $\bar{Z}$ and $\bar{U}$ satisfy*

$$\left\|Z^{(0)} - \bar{Z}\bar{U}\right\| \le \mathcal{O}\left(m^{5/2}\right)u\|\bar{Z}\|\|A\|^{1/2}\left\|Z^{(0)}\right\|, \tag{2.1}$$

$$\left\|\bar{Z}^TA\bar{Z} - I\right\| \le \mathcal{O}\left(m^{5/2}\right)u\|A\|\|\bar{Z}\|^2, \tag{2.2}$$

$$\left\|\left(Z^{(0)}\right)^TAZ^{(0)} - \bar{U}^T\bar{U}\right\| \le \mathcal{O}\left(m^{5/2}\right)u\|A\|^{1/2}\left\|Z^{(0)}\right\|^2, \tag{2.3}$$

$$\left\|A^{-1} - \bar{Z}\bar{Z}^T\right\| \le \mathcal{O}\left(m^{5/2}\right)u\|A\|\|\bar{Z}\|^2\left\|A^{-1}\right\|. \tag{2.4}$$

Proof The backward stable eigendecomposition delivers the computed eigendecomposition $\bar{V}\bar{\Lambda}\bar{V}^T$ which is nearly the exact eigendecomposition of the nearby matrix $A + \Delta A = \hat{V}\bar{\Lambda}\hat{V}^T$ with $\|\Delta A\| \le \mathcal{O}(m^{5/2})u\|A\|$ and where $\hat{V} = \bar{V} + \Delta V$ is exactly orthogonal and $\|\Delta V\| \le \mathcal{O}(m^{5/2})u$ (see [30]). Multiplying the matrix $Z^{(0)}$ with $\bar{V}^T$ and $\bar{\Lambda}^{1/2}$ from the left and applying some backward stable QR decomposition (such as Householder QR [21]) to the product $\mathrm{fl}[\bar{\Lambda}^{1/2}\bar{V}^TZ^{(0)}]$ we get for the factor $\bar{U}$

$$\bar{\Lambda}^{1/2}\bar{V}^TZ^{(0)} = \hat{Q}\bar{U} + \Delta E_1, \quad \|\Delta E_1\| \le \mathcal{O}\left(mn^{3/2}\right)u\|\bar{\Lambda}\|^{1/2}\left\|Z^{(0)}\right\|, \tag{2.5}$$

where $\hat{Q} = \bar{Q} + \Delta Q$ is exactly orthogonal matrix with $\|\Delta Q\| \le \mathcal{O}(mn^{3/2})u$. The matrix $\bar{Z}$ is the computed product of two nearly orthogonal and one diagonal matrix satisfying

$$\bar{Z} = \hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2, \quad \|\Delta E_2\| \le \mathcal{O}\left(m^{1/2}n^{1/2}\right)u\left\|\bar{\Lambda}^{-1}\right\|^{1/2}. \tag{2.6}$$

From (2.5) 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 (2.6) we get the factorization (2.1) for the computed factors $\bar{Z}$ and $\bar{U}$. The bound for the factorization error follows from the fact that $\|\bar{\Lambda}\| \lessapprox \|A\|$ and $\|\bar{\Lambda}^{-1}\|^{1/2} \lessapprox \|\bar{Z}\|$. Note that if $Z^{(0)} = I$, then we have $\|I - \bar{Z}\bar{U}\| \le \mathcal{O}(m^{5/2})u\|\bar{Z}\|\|\bar{U}\|$. The loss of A-orthogonality between the columns of the computed factor $\bar{Z}$ expressed as $\bar{Z}^TA\bar{Z} - I = (\hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2)^T(\hat{V}\bar{\Lambda}\hat{V}^T - \Delta A)(\hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2) - I$ which gives then the bound (2.2). Considering (2.5) we get $(Z^{(0)})^T\bar{V}\bar{\Lambda}\bar{V}^TZ^{(0)} = (\hat{Q}\bar{U} + \Delta E_1)^T(\hat{Q}\bar{U} + \Delta E_1)$ which together with $(\bar{V} + \Delta V)\bar{\Lambda}(\bar{V} + \Delta V)^T = A + \Delta A$ leads to the bound (2.3). The error in the inverse factorization can be bounded using $A^{-1} - \bar{Z}\bar{Z}^T = \hat{V}\bar{\Lambda}^{-1/2}(I - \bar{\Lambda}^{-1/2}\hat{V}^T\Delta A\hat{V}\bar{\Lambda}^{-1/2})^{-1}\bar{\Lambda}^{-1/2}\hat{V}^T - (\hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2)(\hat{V}\bar{\Lambda}^{-1/2}\hat{Q} + \Delta E_2)^T$. □

Corollary 2.1 *If in addition to assumptions of Theorem 2.1 the matrix A is diagonal, then we have $\|Z^{(0)} - \bar{Z}\bar{U}\| \leq \mathcal{O}(mn^{3/2})u\|\bar{Z}\|\|\bar{U}\|$ and $\|\bar{Z}^T A\bar{Z} - I\| \leq \mathcal{O}(mn^{3/2})u$.*

Proof If A is diagonal, there is no need for its eigendecomposition and one can rewrite (2.5) as $A^{1/2}Z^{(0)} = \hat{Q}\bar{U} + \Delta E_1$, $\|\Delta E_1\| \leq \mathcal{O}(mn^{3/2})u\|A^{1/2}Z^{(0)}\|$. The factor $\bar{Z}$ can be computed directly from A and $\bar{Q}$ as $\bar{Z} = A^{-1/2}\bar{Q} + \Delta E_2$ with $|\Delta E_2| \leq u|A^{-1/2}||\bar{Q}|$ and we get $Z^{(0)} = \bar{Z}\bar{U} - \Delta E_2\bar{U} + A^{-1/2}\Delta E_1$. This leads then to the better bound $\|Z^{(0)} - \bar{Z}\bar{U}\| \leq \mathcal{O}(mn^{3/2})u\|A^{-1}\|^{1/2}\|A^{1/2}Z^{(0)}\|$. Again, for a diagonal A we can write $\bar{Z}^T A\bar{Z} - I = (A^{-1/2}\bar{Q} + \Delta E_2)^T A(A^{-1/2}\bar{Q} + \Delta E_2) - I = \bar{Q}^T\bar{Q} - I + \bar{Q}A^{1/2}\Delta E_2 + (\Delta E_2)^T A^{1/2}\bar{Q} + (\Delta E_2)^T A\Delta E_2$. Since $\|A^{1/2}\Delta E_2\| \leq \mathcal{O}(n^{1/2})u$ this identity then gives rise to the bound $\|\bar{Z}^T A\bar{Z} - I\| \leq \mathcal{O}(mn^{3/2})u$. □

It is clear from (2.5) and (2.6) that $\|\bar{U}\| \approx \|\bar{A}^{1/2}\bar{V}^T Z^{(0)}\| \approx \|A^{1/2}Z^{(0)}\|$ and $\|\bar{Z}\| \approx \|\bar{A}^{-1}\|^{1/2} \approx \|A^{-1}\|^{1/2}$. For a general A this leads to the bounds $\|Z^{(0)} - \bar{Z}\bar{U}\| \leq \mathcal{O}(m^{5/2})u\kappa^{1/2}(A)\|Z^{(0)}\|$ and $\|\bar{Z}^T A\bar{Z} - I\| \leq \mathcal{O}(m^{5/2})u\kappa(A)$. The bound (2.1) can be however improved into $\|Z^{(0)} - \bar{Z}\bar{U}\| \leq \mathcal{O}(n^2)u\|\bar{Z}\|\|\bar{U}\| \leq O(n^2)u\|A^{-1}\|^{1/2} \times \|A^{1/2}Z^{(0)}\|$ provided that we have computed the factor $\bar{Z}$ via triangular inverse of the matrix $\bar{U}$ and the right-hand side $Z^{(0)}$ (see Sect. 14.2 of [21]). This however leads to the weaker bound for the loss of A-orthogonality.

If we compute the factor $\bar{Z}$ as a product of two (nearly) orthogonal and one diagonal matrix with a condition number equal to $\kappa^{1/2}(A)$ the bounds (2.2), (2.3), and (2.4) do not depend on the conditioning of $Z^{(0)}$ or $A^{1/2}Z^{(0)}$. This is probably the best approach one can get in finite precision arithmetic if we do not have any other information on A. In this sense the backward stable eigendecomposition-based implementation can be considered as a reference approach. One can hardly expect that bounds on $\|\bar{Z}^T A\bar{Z} - I\|$ and $\|A^{-1} - \bar{Z}\bar{Z}^T\|$ will not depend on the conditioning of the matrix A at least for general symmetric positive definite A. However, in the case of the standard inner product $A = I$ or in the case of the weighted inner-product (with a diagonal A) all these quantities are small multiples of the roundoff unit u.

3 Modified Gram–Schmidt orthogonalization

Probably the most frequently used orthogonalization algorithm is the modified Gram–Schmidt process which represents a good compromise between efficiency and numerical stability. Assuming the use of the non-standard inner product in this section we show that the factorization error $Z^{(0)} - \bar{Z}\bar{U}$ is not small in all implementations of the Gram–Schmidt process. It depends on the quantity $\|\bar{Z}\|\|\bar{U}\| \lessapprox \|A^{-1}\|^{1/2}\|A^{1/2}Z^{(0)}\|$ which can be significantly larger than $\|Z^{(0)}\|$. Therefore the factorization error depends on the conditioning of the matrix A, but it does not depend on the condition number of the matrix $A^{1/2}Z^{(0)}$. This is no longer true for the loss of A-orthogonality $\bar{Z}^T A\bar{Z} - I$ where $\kappa(A^{1/2}Z^{(0)})$ plays a dominant role in the Gram–Schmidt orthogonalization process. Since there is a significant difference in the accuracy of the computed inner product $\mathrm{fl}[\langle\cdot,\cdot\rangle_A]$ we will distinguish between the case of a general positive definite matrix A and the case when the inner product is induced by a positive and diagonal matrix A. As we will see later, the size of local errors that arise in the

computation of all these innerproducts (and associated norms) may also affect the A-orthogonality between the computed vectors contributing to the bound in a way similar as in (2.2).

Given the matrix $Z^{(0)} = [z_1^{(0)}, \ldots, z_n^{(0)}]$ we consider the following formula 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 have the recurrences

$$z_i^{(j)} = z_i^{(j-1)} - \alpha_{ji} z_j, \tag{3.1}$$

where 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 $j = 1, \ldots, i-1$. The i-th column of Z is then given as $z_i \equiv z_i^{(i-1)}/\|z_i^{(i-1)}\|_A$.

Theorem 3.1 *Due to rounding errors the factors $\bar{Z}$ and $\bar{U}$ computed by the Gram–Schmidt recurrence (3.1) satisfy the first of two main results*

$$Z^{(0)} + \Delta E^{(1)} = \bar{Z}\bar{U}, \quad \|\Delta E^{(1)}\| \leq \mathcal{O}\big(n^{3/2}\big)u\big[\|Z^{(0)}\| + \|\bar{Z}\|\|\bar{U}\|\big]. \tag{3.2}$$

Proof The computed vectors $\bar{z}_i^{(j)}$ satisfy after each projection the recurrence with the local errors $\Delta\delta_i^{(j)}$

$$\bar{z}_i^{(j)} = \bar{z}_i^{(j-1)} - \bar{\alpha}_{ji}\bar{z}_j + \Delta\delta_i^{(j)}, \quad \big|\Delta\delta_i^{(j)}\big| \leq u\big|\bar{z}_i^{(j-1)}\big| + 2u|\bar{\alpha}_{ji}||\bar{z}_j|. \tag{3.3}$$

The term $u|\bar{z}_i^{(j-1)}|$ in (3.3) can be bounded by $2u[|z_i^{(0)}| + \sum_{k=1}^{j-1}|\bar{\alpha}_{ki}||\bar{z}_k|]$. This leads to the bound for $\|\Delta\delta_i^{(j)}\| \leq 2u[\|z_i^{(0)}\| + \sum_{k=1}^{j}|\bar{\alpha}_{ki}|\|\bar{z}_k\|]$. Summarizing (3.3) 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 $|\Delta\delta_i^{(i)}| \leq u|\bar{\alpha}_{ii}||\bar{z}_i|$ gives

$$\bar{\alpha}_{ii}\bar{z}_i = z_i^{(0)} - \sum_{j=1}^{i-1}\bar{\alpha}_{ji}\bar{z}_j + \Delta e_i^{(1)},$$

$$\big|\Delta e_i^{(1)}\big| \leq (i+1)u\left[\big|z_i^{(0)}\big| + \sum_{j=1}^{i}|\bar{\alpha}_{ji}||\bar{z}_j|\right]. \tag{3.4}$$

Introducing then the matrix $\Delta E^{(1)}$ which contains the local errors $\Delta e_i^{(1)} = \sum_{j=1}^{i}\Delta\delta_i^{(j)}$ as its columns we obtain the desired statement. □

In the modified Gram–Schmidt algorithm we have the coefficients in (3.1) defined as $\alpha_{ji} \equiv \langle z_i^{(j-1)}, z_j\rangle_A$. The computed coefficients are then given as $\bar{\alpha}_{ji} = \mathrm{fl}[\langle\bar{z}_i^{(j-1)}, \bar{z}_j\rangle_A]$ and $\bar{\alpha}_{ii} = \mathrm{fl}[\|\bar{z}_i^{(i-1)}\|_A]$. It follows that $u|\bar{\alpha}_{ji}| \leq u\|\bar{z}_i^{(j-1)}\|_A$ (see also (3.5)). Considering the near-monotonicity $\|\bar{z}_i^{(j-1)}\|_A \lessapprox \|z_i^{(0)}\|_A$, which is true under some mild assumptions, we see that (3.2) is a somewhat better bound than (2.1). Indeed it follows that $\|\bar{U}\| \lessapprox \|A^{1/2}Z^{(0)}\|$ and $\|\Delta E^{(1)}\| \leq \mathcal{O}(n^{3/2})u\|\bar{Z}\|\|A^{1/2}Z^{(0)}\|$. Note that the derivation of (3.2) does not depend on the way how we compute the

coefficients α_{ji}. Therefore besides the modified Gram–Schmidt (MGS) algorithm it holds for the classical Gram–Schmidt (CGS) algorithm as well as for the AINV orthogonalization which will be discussed in next section. Theorem 3.1 holds true for any symmetric positive definite matrix A. However, as we will see in experimental section, we haven't managed to reproduce the $\|\bar{Z}\|\|\bar{U}\|$ dependence for A diagonal, while for general A, the bound (3.2) seems to be sharp.

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

$$\left|\mathrm{fl}\left[\left\langle \bar{z}_i^{(j-1)}, \bar{z}_j \right\rangle_A\right] - \left\langle \bar{z}_i^{(j-1)}, \bar{z}_j \right\rangle_A\right| \leq \mathcal{O}\big(m^{3/2}\big)u\|A\|\left\|\bar{z}_i^{(j-1)}\right\|\|\bar{z}_j\|. \tag{3.5}$$

Note that (3.5) can be rather pessimistic and it may be worth to consider the details in computation of the scalar product $\langle \cdot, \cdot \rangle_A$. The results in finite precision arithmetic may also depend on the fact whether we multiply the first or the second argument by A. We will not discuss this issue here. Since the vector $\bar{z}_i = \mathrm{fl}[\bar{z}_i^{(i-1)}/\bar{\alpha}_{ii}]$ with $\bar{\alpha}_{ii} = \mathrm{fl}[\|\bar{z}_i^{(i-1)}\|_A]$ is just the computed result from the normalization of the vector $\bar{z}_i^{(i-1)}$ with respect to the inner product with a general A one can conclude that

$$\left|\|\bar{z}_i\|_A^2 - 1\right| \leq \mathcal{O}\big(m^{3/2}\big)u\|A\|\|\bar{z}_i\|^2. \tag{3.6}$$

If we consider the splitting of the matrix $\bar{Z}^T A\bar{Z} - I = \Delta E^{(4)} + \Delta E^{(3)} + (\Delta E^{(3)})^T$, where $\Delta E^{(4)}$ is diagonal and $\Delta E^{(3)}$ is a strictly upper triangular containing the off-diagonal elements, then it is clear from (3.6) that $\|\Delta E^{(4)}\| \leq \mathcal{O}(m^{3/2})u\|A\|\|\bar{Z}\|^2$. Note that this bound is similar to the bound (2.2) obtained for the implementation based on eigendecomposition.

Theorem 3.2 *For a general symmetric positive definite A satisfying the assumption $\mathcal{O}(m^{3/2}n)u\kappa(A)\kappa(A^{1/2}Z^{(0)}) < 1$, the loss of A-orthogonality between the vectors computed $\bar{Z}$ in the modified Gram–Schmidt algorithm is bounded by*

$$\left\|\bar{Z}^T A\bar{Z} - I\right\| \leq \frac{\mathcal{O}(m^{3/2}n)u\|A\|\|\bar{Z}\| \max_{j\leq i} \frac{\|\bar{z}_i^{(j-1)}\|}{\|\bar{z}_i^{(j-1)}\|_A}\kappa(A^{1/2}Z^{(0)})}{1 - \mathcal{O}(m^{3/2}n)u\|A\|\|\bar{Z}\| \max_{j\leq i} \frac{\|\bar{z}_i^{(j-1)}\|}{\|\bar{z}_i^{(j-1)}\|_A}\kappa(A^{1/2}Z^{(0)})}.$$

Proof In this proof, we essentially follow the analysis of Å. Björck for the MGS algorithm with the standard inner product given in his pioneering work [8]. Considering recursively the formula (3.3) for $k = j+1, \ldots, i-1$, rearranging the resulting identity and taking the A-inner product with the vector $\bar{z}_j$ we obtain

$$\sum_{k=j+1}^{i} \bar{\alpha}_{ki}\langle \bar{z}_j, \bar{z}_k \rangle_A = \left\langle \bar{z}_j, \bar{z}_i^{(j)} \right\rangle_A + \sum_{k=j+1}^{i} \left\langle \bar{z}_j, \Delta\delta_i^{(k)} \right\rangle_A. \tag{3.7}$$

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 $\bar{z}_i^{(j)}$ with the vector $\bar{z}_j$ one can write $\langle \bar{z}_j, \bar{z}_i^{(j)} \rangle_A = -\langle \bar{z}_j, \Delta\eta_i^{(j)} \rangle_A + \langle \bar{z}_j, \Delta\delta_i^{(j)} \rangle_A$, where the local error $\Delta\eta_i^{(j)}$ is given as

$$\Delta\eta_i^{(j)} = \big(\mathrm{fl}\big[\langle \bar{z}_i^{(j-1)}, \bar{z}_j \rangle_A\big] - \langle \bar{z}_i^{(j-1)}, \bar{z}_j \rangle_A\big)\bar{z}_j + \big(\|\bar{z}_j\|_A^2 - 1\big)\bar{z}_i^{(j-1)}. \tag{3.8}$$

The left-hand side of (3.7) represents the (j, i)-component of the matrix $\Delta E^{(3)}\bar{U}$ and so the identity (3.7) 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 as $[\Delta E^{(2)}]_{ij} = -\langle \bar{z}_j, \Delta\eta_i^{(j)} \rangle_A + \langle \bar{z}_j, \sum_{k=j}^{i} \Delta\delta_i^{(k)} \rangle_A$. Thus the norm of $\Delta E^{(3)}$ can be bounded by $\|\Delta E^{(3)}\| \le \|\Delta E^{(2)}\|_F \|\bar{U}^{-1}\|$. The elements of the matrix $\Delta E^{(2)}$ depend on the A-norms of local errors $\|\Delta\delta_i^{(k)}\|_A$ and $\|\Delta\eta_i^{(j)}\|_A$ which play a decisive role here. From (3.8) it follows that we need the estimate (3.5) and (3.6) for the terms $\mathrm{fl}[\langle \bar{z}_i^{(j-1)}, \bar{z}_j \rangle_A] - \langle \bar{z}_i^{(j-1)}, \bar{z}_j \rangle_A\rangle$ and $\|\bar{z}_j\|_A^2 - 1$. Then it follows that $\|\Delta\eta_i^{(j)}\|_A \le \mathcal{O}(m^{3/2})u\|A\| \times \|\bar{z}_j\|(\|\bar{z}_i^{(j-1)}\|\|\bar{z}_j\|_A + \|\bar{z}_j\|\|\bar{z}_i^{(j-1)}\|_A) \le \mathcal{O}(m^{3/2})u\|A\|\|\bar{z}_j\|(\|\bar{z}_i^{(j-1)}\|/\|\bar{z}_i^{(j-1)}\|_A + \|\bar{z}_j\|)\|\bar{z}_i^{(j-1)}\|_A$. Since for the local error $\Delta\delta_i^{(j)}$ from previous theorem we have already $\|\Delta\delta_i^{(j)}\|_A \le \|A\|^{1/2}\|\Delta\delta_i^{(j)}\| \le 4u\|A\|^{1/2}\sum_{k=1}^{j}\|\bar{z}_k\|\|z_i^{(0)}\|_A$ the Frobenius norm of the matrix $\Delta E^{(2)}$ can be bounded by $\|\Delta E^{(2)}\|_F \le \mathcal{O}(m^{3/2}n)u\|A\| \times \|\bar{Z}\| \max_{j\le i}\|\bar{z}_i^{(j-1)}\|/\|\bar{z}_i^{(j-1)}\|_A$. The computed factors $\bar{Z}$ and $\bar{U}$ satisfy (3.2) and therefore we can write

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

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 $U^TU = (Z^{(0)})^TAZ^{(0)}$ as follows

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

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

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

It is clear that if we assume that $2\|\Delta E^{(2)}\|_F\|U^{-1}\| + \|\Delta E^{(4)}\| < 1$ then $\|\Delta E^{(5)}\| \lessapprox 2(\|A\|^{1/2}\|\Delta E^{(1)}\| + \|\Delta E^{(2)}\|_F)\|U^{-1}\|$ and we get

$$\big\|\bar{U}^{-1}\big\| \le \big(1 - \big\|\Delta E^{(5)}\big\|\big)^{-1/2}\big\|U^{-1}\big\|. \qquad \square$$

Indeed the loss of A-orthogonality in the modified Gram–Schmidt algorithm 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 given as $\mathcal{O}(m^{3/2})u\|A\|\|\bar{Z}\|\max_{j\le i}\|\bar{z}_i^{(j-1)}\|/\|\bar{z}_i^{(j-1)}\|_A \le \mathcal{O}(m^{3/2})u\kappa(A)$.

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

$$\left|\mathrm{fl}\left[\left\langle \bar{z}_i^{(j-1)}, \bar{z}_j\right\rangle_A\right] - \left\langle \bar{z}_i^{(j-1)}, \bar{z}_j\right\rangle_A\right| \le \mathcal{O}(m)u\left\|\bar{z}_i^{(j-1)}\right\|_A\|\bar{z}_j\|_A \tag{3.9}$$

and for the error in the normalization of the vector $\bar{z}_i^{(i-1)}$ it follows that

$$\left|\|\bar{z}_i\|_A^2 - 1\right| \le \mathcal{O}(m)u. \tag{3.10}$$

The previous two results lead to significantly better bounds for the local errors $\|\Delta\delta_i^{(j)}\|_A \le 4u\|z_i^{(0)}\|_A$ and $\|\Delta\eta_i^{(k)}\|_A \le \mathcal{O}(m)u\|z_i^{(0)}\|_A$. In matrix notation we then have $\|\Delta E^{(2)}\|_F \le \mathcal{O}(mn^{1/2})u\|A^{1/2}Z^{(0)}\|$ and $\|\Delta E^{(4)}\| \le \mathcal{O}(m)u$. Indeed the relative local errors are small multiples of the roundoff unit and do not depend on the conditioning of the matrix A. Our results for A diagonal generalizes the case with standard inner product [8].

Corollary 3.1 *Assuming a positive diagonal A such that $\mathcal{O}(mn)u\kappa(A^{1/2}Z^{(0)}) < 1$, the loss of A-orthogonality between the computed vectors $\bar{Z}$ in the modified Gram–Schmidt algorithm is bounded by*

$$\left\|\bar{Z}^T A\bar{Z} - I\right\| \le \frac{\mathcal{O}(mn)u\kappa(A^{1/2}Z^{(0)})}{1-\mathcal{O}(mn)u\kappa(A^{1/2}Z^{(0)})}. \tag{3.11}$$

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 [17], see also [16] and [18]. Thomas and Zahar in [36] (see also [35] and [34]) studied the MGS algorithm with a non-standard inner product where the matrix A is given as $A = B^T B$ and so its Cholesky factor is known a priori. Under assumptions on the size of local errors similar to (3.9) and (3.10) they proved that the loss of A-orthogonality between the computed vectors is bounded by a term analogous to (3.11).

4 Classical Gram–Schmidt and AINV orthogonalization

In this section we analyze the CGS and AINV orthogonalization algorithms. Since they use the recurrences (3.1) the result on the factorization error developed in Theorem 3.1 holds true also here. We are going to show that the loss of A-orthogonality in

CGS and AINV can be significantly worse than in MGS. Nevertheless this quantity remains bounded by a term proportional to the factor $\kappa(A^{1/2}Z^{(0)})\kappa^{1/2}(A)\kappa(Z^{(0)})$, which essentially means the square of the condition number of the matrix $A^{1/2}Z^{(0)}$. In other words, the matrix $(Z^{(0)})^T AZ^{(0)}$ will play an important role here. Although these two schemes are quite different in formulation their numerical behavior is very similar.

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

$$\left|\mathrm{fl}\left[\left\langle z_i^{(0)}, \bar{z}_j\right\rangle_A\right] - \left\langle z_i^{(0)}, \bar{z}_j\right\rangle_A\right| \le \mathcal{O}\left(m^{3/2}\right)u\|A\|\left\|z_i^{(0)}\right\|\|\bar{z}_j\|. \tag{4.1}$$

It was shown in [33] that the diagonal entries $\alpha_{ii} = (\|z_i^{(0)}\|_A^2 - \|Z_{i-1}^T A z_i^{(0)}\|^2)^{1/2}$ in the CGS algorithm must be computed with the formula

$$\alpha_{ii} = \left(\left\|z_i^{(0)}\right\|_A^2 - \sum_{j=1}^{i-1}\alpha_{ji}^2\right)^{1/2}. \tag{4.2}$$

The computed elements $\bar{\alpha}_{ji}$ for $j = 1, \ldots, i$ in finite precision arithmetic then satisfy the bound

$$\left|\left\|z_i^{(0)}\right\|_A^2 - \sum_{j=1}^{i}\bar{\alpha}_{ji}^2\right| \le \mathcal{O}\left(m^{3/2}\right)u\|A\|\left\|z_i^{(0)}\right\|^2 \tag{4.3}$$

which is not true if we use the standard formula $\bar{\alpha}_{ii} = \mathrm{fl}[\|\bar{z}_i^{(i-1)}\|_A]$. For details we refer to [33].

Theorem 4.1 *Assuming that* $\mathcal{O}(m^{3/2}n)u\kappa(A)\kappa(A^{1/2}Z^{(0)})\kappa(Z^{(0)}) < 1$ *holds for a general symmetric positive definite* A, *the loss of* A*-orthogonality in the classical Gram–Schmidt algorithm is bounded by*

$$\left\|I - \bar{Z}^T A\bar{Z}\right\| \le \frac{\mathcal{O}(m^{3/2}n)u\|A\|^{1/2}\|\bar{Z}\|\kappa(A^{1/2}Z^{(0)})\kappa^{1/2}(A)\kappa(Z^{(0)})}{1 - \mathcal{O}(m^{3/2}n)u\|A\|^{1/2}\|\bar{Z}\|\kappa(A^{1/2}Z^{(0)})\kappa^{1/2}(A)\kappa(Z^{(0)})}. \tag{4.4}$$

Proof As we have already noted the recurrence (3.3) for the vectors $\bar{z}_i = \mathrm{fl}[\bar{z}_i^{(i-1)}/\bar{\alpha}_{ii}]$ will have the same form together with the bounds (3.2) and (3.4). Considering (3.4) 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 manipulations we get consecutively the following identities

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

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

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

Let the last two terms of this equation define the (i, j)-th element of the error matrix ΔE. Considering (4.1), (4.3) and the bound (3.4) for the local errors developed in the previous section we obtain the matrix identity (the analogous result for the standard inner product can be found in [14])

$$\begin{aligned} &\big(Z^{(0)}\big)^T A Z^{(0)} + \Delta E = \bar{U}^T \bar{U}, \\ &\|\Delta E\| \le \mathcal{O}\big(m^{3/2}n\big)u\|A\|\big\|Z^{(0)}\big\|\big(\big\|Z^{(0)}\big\| + \|\bar{Z}\|\big\|A^{1/2}Z^{(0)}\big\|\big). \end{aligned} \tag{4.5}$$

Indeed the computed factor $\bar{U}$ is the exact Cholesky factor of the matrix $(Z^{(0)})^T A Z^{(0)}$ perturbed by the matrix ΔE. In other words, up to the factor $\|A\|\|\bar{Z}\|\|Z^{(0)}\|/\|A^{1/2}Z^{(0)}\|$, the classical Gram–Schmidt algorithm is a way to compute a backward stable Cholesky factor of the cross-product matrix $(Z^{(0)})^T A Z^{(0)}$ (see also [13, 14]). Using (4.5) and (3.2) we have

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

which gives rise to the bound for the loss of A-orthogonality that depends quadratically on the reciprocal of the minimal singular value of $A^{1/2}Z^{(0)}$. Indeed using the bounds from (3.2) and (4.5) we have the statement of our theorem whereas the size of local errors is reflected similarly as in the modified Gram–Schmidt algorithm with the worst-case bound $\|A\|^{1/2}\|\bar{Z}\| \lesssim \kappa^{1/2}(A)$. □

Corollary 4.1 *Assuming a diagonal positive matrix A such that*

$$\mathcal{O}(mn)u\kappa^2\big(A^{1/2}Z^{(0)}\big) < 1,$$

the loss of A-orthogonality in the classical Gram–Schmidt algorithm is bounded by

$$\big\|I - \bar{Z}^T A\bar{Z}\big\| \le \frac{\mathcal{O}(mn)u\kappa^2(A^{1/2}Z^{(0)})}{1 - \mathcal{O}(mn)u\kappa^2(A^{1/2}Z^{(0)})}. \tag{4.6}$$

Proof In the case of a diagonal positive A one can show due to $|\mathrm{fl}[\langle z_i^{(0)}, \bar{z}_k\rangle_A] - \langle z_i^{(0)}, \bar{z}_k\rangle_A| \le \mathcal{O}(m)u\|z_i^{(0)}\|_A\|\bar{z}_k\|_A$ that $\|A^{1/2}\Delta E^{(1)}\| \le \mathcal{O}(n^{3/2})u\|A^{1/2}Z^{(0)}\|$ and $\|\Delta E\| \le \mathcal{O}(mn)u\|A^{1/2}Z^{(0)}\|^2$ which give rise to a significantly better bound (4.6). □

In the following we will analyze the AINV orthogonalization scheme and show that its numerical behavior is very similar to CGS. The coefficients α_{ji} in

the recurrence (3.1) 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 diagonal coefficients α_{ii} are again computed with the non-standard formula (4.2). 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.

Theorem 4.2 *Assuming that* $\mathcal{O}(m^{3/2}n)u\kappa(A)\kappa(A^{1/2}Z^{(0)})\kappa(Z^{(0)}) < 1$ *holds for a general symmetric positive definite* A, *the loss of* A*-orthogonality in the AINV orthogonalization process is bounded by*

$$\left\| I - \bar{Z}^T A \bar{Z} \right\| \leq \frac{\mathcal{O}(m^{3/2}n)u\|A\|^{1/2}\|\bar{Z}\|\kappa(A^{1/2}Z^{(0)})\kappa^{1/2}(A)\kappa(Z^{(0)})}{1 - \mathcal{O}(m^{3/2}n)u\|A\|^{1/2}\|\bar{Z}\|\kappa(A^{1/2}Z^{(0)})\kappa^{1/2}(A)\kappa(Z^{(0)})}.$$

Proof The computed orthogonalization coefficients $\bar{\alpha}_{ji}$ are given as $\bar{\alpha}_{ji} = \mathrm{fl}[\langle \bar{z}_i^{(j-1)}, \bar{z}_j^{(0)}\rangle_A]$, where $\bar{z}_j^{(0)} = \mathrm{fl}[z_j^{(0)}/\bar{\alpha}_{jj}]$. From (3.3) we can write that $\bar{z}_i^{(j-1)} = z_i^{(0)} - \sum_{k=1}^{j-1} \bar{\alpha}_{ki}\bar{z}_k + \sum_{k=1}^{j-1} \Delta\delta_i^{(k)}$. Taking the A-inner product with the vector $z_j^{(0)}$ we obtain successively

$$\begin{aligned}
\bar{\alpha}_{jj}\left\langle \bar{z}_i^{(j-1)}, \frac{z_j^{(0)}}{\bar{\alpha}_{jj}}\right\rangle_A &= \langle z_i^{(0)}, z_j^{(0)}\rangle_A - \sum_{k=1}^{j-1} \bar{\alpha}_{ki}\langle \bar{z}_k, z_j^{(0)}\rangle_A + \left\langle \sum_{k=1}^{j-1} \Delta\delta_i^{(k)}, z_j^{(0)}\right\rangle_A, \\
\bar{\alpha}_{jj}\langle \bar{z}_i^{(j-1)}, \bar{z}_j^{(0)}\rangle_A &= \langle z_i^{(0)}, z_j^{(0)}\rangle_A - \sum_{k=1}^{j-1} \bar{\alpha}_{ki}\langle \bar{z}_k, z_j^{(0)}\rangle_A + \left\langle \sum_{k=1}^{j-1} \Delta\delta_i^{(k)}, z_j^{(0)}\right\rangle_A \\
&\quad + \bar{\alpha}_{jj}\left\langle \bar{z}_i^{(j-1)}, \bar{z}_j^{(0)} - \frac{z_j^{(0)}}{\bar{\alpha}_{jj}}\right\rangle_A, \\
\sum_{k=1}^{j-1} \bar{\alpha}_{kj}\hat{\alpha}_{ki} &= \langle z_i^{(0)}, z_j^{(0)}\rangle_A + \left\langle \sum_{k=1}^{j-1} \Delta\delta_i^{(k)}, z_j^{(0)}\right\rangle_A \\
&\quad + \bar{\alpha}_{jj}\left(\mathrm{fl}\left[\langle \bar{z}_i^{(j-1)}, \bar{z}_j^{(0)}\rangle_A\right] - \left\langle \bar{z}_i^{(j-1)}, \frac{z_j^{(0)}}{\bar{\alpha}_{jj}}\right\rangle_A\right),
\end{aligned} \tag{4.7}$$

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

$$\left| \mathrm{fl}\left[\langle \bar{z}_i^{(j-1)}, \bar{z}_j^{(0)}\rangle_A\right] - \left\langle \bar{z}_i^{(j-1)}, \frac{z_j^{(0)}}{\bar{\alpha}_{jj}}\right\rangle_A \right| \leq \mathcal{O}\big(m^{3/2}\big)u\|A\|\,\big\|\bar{z}_i^{(j-1)}\big\|\frac{\|z_j^{(0)}\|}{|\bar{\alpha}_{jj}|}. \tag{4.8}$$

Since the left-hand side of the recurrence (4.7) 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

$$\begin{aligned}\operatorname{striu}\big(\bar{U}^T\hat{U}\big) &= \operatorname{striu}\big(\big(Z^{(0)}\big)^T AZ^{(0)} + \Delta F\big),\\ \big\|\operatorname{striu}(\Delta F)\big\| &\le \mathcal{O}\big(m^{3/2}n\big)u\|A\|\|\bar{Z}\|\big\|Z^{(0)}\big\|\big\|A^{1/2}Z^{(0)}\big\|.\end{aligned}$$

The diagonal elements of α_{ii} are computed with the formula (4.2). The computed quantities $\bar{\alpha}_{ki}$ satisfy $|\|z_i^{(0)}\|_A^2 - \sum_{k=1}^{i}\bar{\alpha}_{ki}^2| \le \mathcal{O}(m^{3/2})\|A\|\|z_i^{(0)}\|^2$. The diagonal entries of the matrix ΔF thus satisfy the bound

$$\begin{aligned}\operatorname{diag}\big(\bar{U}^T\hat{U}\big) &= \operatorname{diag}\big(\big(Z^{(0)}\big)^T AZ^{(0)} + \Delta F\big),\\ \big\|\operatorname{diag}(\Delta F)\big\| &\le \mathcal{O}\big(m^{3/2}\big)\|A\|\big\|Z^{(0)}\big\|^2.\end{aligned}$$

From (3.4) 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

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

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 $(\Delta F)^T$)

$$\operatorname{striu}\big(\hat{U}^T\bar{U}\big) = \operatorname{striu}\big(\big(Z^{(0)}\big)^T AZ^{(0)} + (\Delta F)^T\big),$$

with $\|\operatorname{stril}(\Delta F)\| \le \mathcal{O}(n^{3/2})u\|A\|^{1/2}\|\bar{Z}\|\|A^{1/2}Z^{(0)}\|^2$. 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 matrix ΔF satisfying

$$\begin{aligned}\big(Z^{(0)}\big)^T AZ^{(0)} + \Delta F &= \hat{U}^T\bar{U},\\ \|\Delta F\| &\le \mathcal{O}\big(m^{3/2}n\big)u\|A\|\|\bar{Z}\|\big\|Z^{(0)}\big\|\big\|A^{1/2}Z^{(0)}\big\|.\end{aligned}$$

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, and modifying the approach of [3, 29] we get

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

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

$$2\left\|\Delta\bar{U}U^{-1}\right\|_F \le \left\|U^{-T}\Delta F U^{-1}\right\|_F + \left\|\Delta\bar{U}U^{-1}\right\|_F \left\|\Delta\hat{U}U^{-1}\right\|_F,$$
$$2\left\|\Delta\hat{U}U^{-1}\right\|_F \le \left\|U^{-T}\Delta F U^{-1}\right\|_F + \left\|\Delta\bar{U}U^{-1}\right\|_F \left\|\Delta\hat{U}U^{-1}\right\|_F.$$

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

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

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 $\|\Delta F\| \le \mathcal{O}(m^{3/2}n)u\|A\|\|\bar{Z}\|\|Z^{(0)}\|\|A^{1/2}Z^{(0)}\|$ we finally get the bound for the loss of A-orthogonality between the computed vectors $\bar{Z}$ having identical form as (4.4). □

In the case of a diagonal A we can get the bound identical to (4.6). These results clearly indicate that the numerical behavior of CGS and AINV is quite similar, and in the worst-case when $\|A\|^{1/2}\|\bar{Z}\| \lesssim \kappa^{1/2}(A)$ the loss of A-orthogonality between computed vectors in these two schemes 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 A-orthogonality between computed vectors in MGS, CGS and AINV (besides $\kappa(A)$ bounding in the worst-case the local errors represented by the term $\|A\|\|\bar{Z}\|^2$) depends significantly on the condition number of the matrix $A^{1/2}Z^{(0)}$. On the other hand, for the implementation based on eigendecomposition we have the bound $\|\bar{Z}^T A\bar{Z} - I\| \le \mathcal{O}(m^{5/2})u\|A\|\|\bar{Z}\|^2 \le \mathcal{O}(m^{5/2})u\kappa(A)$ which does not depend explicitly on the matrix $Z^{(0)}$. In this section we consider the classical Gram–Schmidt algorithm with reorthogonalization (CGS2), for which we will get a similar result. The key idea here is that the A-norm of the projection computed after the first orthogonalization step is not infinitely small and it remains bounded by the minimal singular value of the matrix $A^{1/2}Z^{(0)}$. This result follows simply from the recurrence for computed factors $\bar{Z}$ and $\bar{U}$ (formulated in Theorem 5.1). Taking into account the effect of the second orthogonalization step it leads to the bound for the loss of A-orthogonality that does not depend on $A^{1/2}Z^{(0)}$. This phenomenon is often called as "two-steps-are-enough".

Indeed, in the CGS2 algorithm the orthogonalization of the current vector $z_i^{(0)}$ in the classical Gram–Schmidt process is performed exactly twice. Provided we have already the vectors $z_1, \ldots, z_{i-1}$ at the i-th step we generate

$$z_i^{(1)} = z_i^{(0)} - \sum_{j=1}^{i-1} \alpha_{ji}^{(1)} z_j, \qquad z_i^{(2)} = z_i^{(1)} - \sum_{j=1}^{i-1} \alpha_{ji}^{(2)} z_j$$

with $\alpha_{ji}^{(1)} = \langle z_i^{(0)}, z_j \rangle_A$ and $\alpha_{ji}^{(2)} = \langle z_i^{(1)}, z_j \rangle_A$ defined for $j = 1, \ldots, i-1$. 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 for the computed vectors $\bar{Z}$ and the coefficients $\bar{U}$ we have the following theorem.

Theorem 5.1 *The factors $\bar{Z}$ and $\bar{U}$ computed by the classical Gram–Schmidt algorithm with reorthogonalization satisfy the recurrence*

$$\begin{aligned} &Z^{(0)} + \Delta E^{(1)} = \bar{Z}\bar{U}, \\ &\big\|A^{1/2}\Delta E^{(1)}\big\| \le \mathcal{O}\big(n^{3/2}\big)u\|A\|^{1/2}\|\bar{Z}\|\big\|A^{1/2}Z^{(0)}\big\|. \end{aligned} \tag{5.1}$$

Proof In finite precision arithmetic for the computed vectors $\bar{z}_i^{(1)}$ and $\bar{z}_i^{(2)}$ we have

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

$$\big\|\Delta\delta_i^{(1)}\big\| \le \mathcal{O}(n)u\left(\big\|z_i^{(0)}\big\| + \sum_{j=1}^{i-1} \big|\bar{\alpha}_{ji}^{(1)}\big|\|\bar{z}_j\|\right), \tag{5.2}$$

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

$$\big\|\Delta\delta_i^{(2)}\big\| \le \mathcal{O}(n)u\left(\big\|\bar{z}_i^{(1)}\big\| + \sum_{j=1}^{i-1} \big|\bar{\alpha}_{ji}^{(2)}\big|\|\bar{z}_j\|\right). \tag{5.3}$$

It is clear that the computed coefficients $\bar{\alpha}_{ji}^{(1)} = \mathrm{fl}[\langle z_i^{(0)}, \bar{z}_j \rangle_A]$ and $\bar{\alpha}_{ji}^{(2)} = \mathrm{fl}[\langle \bar{z}_i^{(1)}, \bar{z}_j \rangle_A]$ satisfy the bounds $|\bar{\alpha}_{ji}^{(1)} - \langle z_i^{(0)}, \bar{z}_j \rangle_A| \le \mathcal{O}(m^{3/2})u\|A\|\|z_i^{(0)}\|\|\bar{z}_j\|$ and $|\bar{\alpha}_{ji}^{(2)} - \langle \bar{z}_i^{(1)}, \bar{z}_j \rangle_A| \le \mathcal{O}(m^{3/2})u\|A\|\|\bar{z}_i^{(1)}\|\|\bar{z}_j\|$. The local errors $\Delta\delta_i^{(1)}$ and $\Delta\delta_i^{(2)}$ can be bounded as

$$\begin{aligned} \big\|\Delta\delta_i^{(1)}\big\|_A &\le \|A\|^{1/2}\big\|\Delta\delta_i^{(1)}\big\| \le \mathcal{O}(n)u\|A\|^{1/2}\big(\big\|z_i^{(0)}\big\| + \big\|z_i^{(0)}\big\|_A\|\bar{Z}_{i-1}\|\big), \\ \big\|\Delta\delta_i^{(2)}\big\|_A &\le \|A\|^{1/2}\big\|\Delta\delta_i^{(1)}\big\| \le \mathcal{O}(n)u\|A\|^{1/2}\big(\big\|\bar{z}_i^{(1)}\big\| + \big\|\bar{z}_i^{(1)}\big\|_A\|\bar{Z}_{i-1}\|\big) \\ &\le \mathcal{O}(n)u\|A\|^{1/2}\big(\big\|z_i^{(0)}\big\| + \big\|z_i^{(0)}\big\|_A\|\bar{Z}_{i-1}\|\big) \end{aligned}$$

due to the (near-) monotonicity $\|\bar{z}_i^{(1)}\|_A \lesssim i\|z_i^{(0)}\|_A$. The recurrences (5.2)–(5.3) can be rewritten as $z_i^{(0)} + \Delta\delta_i^{(1)} + \Delta\delta_i^{(2)} = \sum_{j=1}^{i-1}(\bar{\alpha}_{ji}^{(1)} + \bar{\alpha}_{ji}^{(2)})\bar{z}_j + \bar{z}_i^{(2)}$ with $\bar{z}_i^{(2)} = \bar{\alpha}_{ii}\bar{z}_i + \Delta\delta_i^{(0)}$. Setting the coefficients $\bar{\alpha}_{ji}$ of the upper triangular factor $\bar{U}$ as $\bar{\alpha}_{ji} = \bar{\alpha}_{ji}^{(1)} + \bar{\alpha}_{ji}^{(2)}$ we obtain the desired statement. □

Theorem 5.2 *Assuming* $\mathcal{O}(m^{3/2}n)u\kappa^{1/2}(A)\kappa(A^{1/2}Z^{(0)}) < 1$ *for a general symmetric positive definite* A, *the loss of* A*-orthogonality in the CGS2 algorithm is bounded by*

$$\left\|I - \bar{Z}^T A\bar{Z}\right\| \le \mathcal{O}\left(m^{3/2}n\right)u\|A\|\|\bar{Z}\|\|\tilde{Z}\| \le \mathcal{O}\left(m^{3/2}n\right)u\kappa(A). \tag{5.4}$$

Proof We will use an incremental approach and assume that the loss of A-orthogonality between the columns of the matrix $\bar{Z}_{i-1} = [\bar{z}_1, \ldots, \bar{z}_{i-1}]$ after $i-1$ steps is bounded by

$$\left\|I - \bar{Z}_{i-1}^T A\bar{Z}_{i-1}\right\| \le \mathcal{O}\left(m^{3/2}n\right)u\|A\|\|\bar{Z}_{i-1}\|\|\tilde{Z}_{i-1}\| \le \mathcal{O}\left(m^{3/2}n\right)u\kappa(A) \tag{5.5}$$

and show that this statement will remain true also at step i, whereas the matrix $\tilde{Z}_{i-1}$ is defined as $\tilde{Z}_{i-1} = [\frac{\bar{z}_1^{(1)}}{\|\bar{z}_1^{(1)}\|_A}, \ldots, \frac{\bar{z}_{i-1}^{(1)}}{\|\bar{z}_{i-1}^{(1)}\|_A}]$. The recurrences (5.2)–(5.3) can be reformulated into the form

$$\begin{aligned} \bar{z}_i^{(1)} &= \left(I - \bar{Z}_{i-1}\bar{Z}_{i-1}^T A\right)z_i^{(0)} + \Delta\eta_i^{(1)}, \\ \left\|\Delta\eta_i^{(1)}\right\|_A &\le \mathcal{O}\left(m^{3/2}n\right)u\|A\|\|\bar{Z}_{i-1}\|\left\|z_i^{(0)}\right\|, \end{aligned} \tag{5.6}$$

$$\begin{aligned} \bar{z}_i^{(2)} &= \left(I - \bar{Z}_{i-1}\bar{Z}_{i-1}^T A\right)\bar{z}_i^{(1)} + \Delta\eta_i^{(2)}, \\ \left\|\Delta\eta_i^{(2)}\right\|_A &\le \mathcal{O}\left(m^{3/2}n\right)u\|A\|\|\bar{Z}_{i-1}\|\left\|\bar{z}_i^{(1)}\right\|. \end{aligned} \tag{5.7}$$

Multiplication of (5.6) from the left by $\bar{Z}_{i-1}^T A$ leads to the identity

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

Taking the norm, dividing by the A-norm of the vector $\bar{z}_i^{(1)}$ and taking into account (5.6) and (5.5) leads to the bound for the quantity $\|\bar{Z}_{i-1}^T A(\bar{z}_i^{(1)}/\|\bar{z}_i^{(1)}\|_A)\|$

$$\frac{\|\bar{Z}_{i-1}^T A\bar{z}_i^{(1)}\|}{\|\bar{z}_i^{(1)}\|_A} \le \mathcal{O}\left(m^{3/2}n\right)u\|A\|\|\bar{Z}_{i-1}\|\|\tilde{Z}_{i-1}\|\left\|A^{1/2}\bar{Z}_{i-1}\right\|\frac{\|z_i^{(0)}\|_A}{\|\bar{z}_i^{(1)}\|_A}. \tag{5.8}$$

The factor $\|z_i^{(0)}\|_A/\|\bar{z}_i^{(1)}\|_A$ can be estimated using the recurrence (5.1) for the first $i-1$ steps together with (5.2) which can be written after multiplication by $A^{1/2}$ from the left in the form

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

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

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

Assuming that $\mathcal{O}(n^{3/2})u\kappa^{1/2}(A)\kappa(A^{1/2}Z_i^{(0)}) < 1$ and using the bounds from (5.2) and (5.1) we can give a lower bound for the A-norm of the vector $\bar{z}_i^{(1)}$

$$\big\|\bar{z}_i^{(1)}\big\|_A \ge \sigma_i\big(A^{1/2}Z_i^{(0)}\big)\big(1 - \mathcal{O}\big(n^{3/2}\big)u\|A\|^{1/2}\|\bar{Z}_i\|\kappa\big(A^{1/2}Z_i^{(0)}\big)\big). \tag{5.9}$$

Due to (5.8) and (5.9) we get the bound for the left-hand side of (5.8)

$$\frac{\|\bar{Z}_{i-1}^T A\bar{z}_i^{(1)}\|}{\|\bar{z}_i^{(1)}\|_A} \le \frac{\mathcal{O}(m^{3/2}n)u\|A\|\|\bar{Z}_{i-1}\|\|\tilde{Z}_{i-1}\|\kappa(A^{1/2}Z_i^{(0)})}{1 - \mathcal{O}(n^{3/2})u\|A\|^{1/2}\|\bar{Z}_i\|\kappa(A^{1/2}Z_i^{(0)})}.$$

Similarly as before we consider (5.7), multiply it from the left by $\bar{Z}_{i-1}^T A$ and obtain

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

which is treated similarly as in (5.8), i.e. using (5.5) and (5.7) we get

$$\frac{\|\bar{Z}_{i-1}^T A\bar{z}_i^{(2)}\|}{\|\bar{z}_i^{(2)}\|_A} \le \mathcal{O}\big(m^{3/2}n\big)u\|A\|\|\bar{Z}_{i-1}\|\left(\|\tilde{Z}_{i-1}\| + \frac{\|\bar{z}_i^{(1)}\|}{\|\bar{z}_i^{(1)}\|_A}\right) \times \big\|A^{1/2}\bar{Z}_{i-1}\big\|\frac{\|\bar{z}_i^{(1)}\|_A}{\|\bar{z}_i^{(2)}\|_A}. \tag{5.10}$$

The factor $\|\bar{z}_i^{(1)}\|_A/\|\bar{z}_i^{(2)}\|_A$ can be bounded from below reconsidering (5.7) as follows

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

which under the stronger assumption $\mathcal{O}(m^{3/2}n)u\kappa^{1/2}(A)\kappa(A^{1/2}Z_i^{(0)}) < 1$ leads to the final bound $\|\bar{z}_i^{(1)}\|_A/\|\bar{z}_i^{(2)}\|_A \le [1 - \mathcal{O}(m^{3/2}n)u\kappa^{1/2}(A)\kappa(A^{1/2}Z_i^{(0)})]^{-1}$. Considering $\bar{z}_i^{(2)} = \bar{\alpha}_{ii}\bar{z}_i + \Delta\delta_i^{(i)}$ with $|\bar{\alpha}_{ii} - \|\bar{z}_i^{(2)}\|_A| \le \mathcal{O}(m^{3/2})u\|A\|\|\bar{z}_i^{(2)}\|^2$ we can relate the left-hand side of (5.10) with the quantity $\|\bar{Z}_{i-1}^T A\bar{z}_i\|$. Taking into account also the error from the normalization $|1 - \|\bar{z}_i\|_A^2| \le \mathcal{O}(m^{3/2})u\|A\|\|\bar{z}_i\|^2$ we end up with the statement of our theorem. □

Again, the bound (5.4) can be significantly improved for diagonal A. Assuming only $\mathcal{O}(mn)u\kappa(A^{1/2}Z^{(0)}) < 1$ one can show using the same approach that the

A-orthogonality between the computed vectors $\bar{Z}$ 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 $\|I - \bar{Z}^T A\bar{Z}\| \leq \mathcal{O}(mn)u$. It is interesting to note that a similar result holds also for the EIG implementation based on backward stable eigendecomposition. However, the assumption $\mathcal{O}(m^{3/2}n)u\kappa^{1/2}(A)\kappa(A^{1/2}Z^{(0)}) < 1$ 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 important cases, where $\kappa(A) \ll \kappa(A^{1/2}Z^{(0)})$, $\kappa(A^{1/2}Z^{(0)}) \ll \kappa(A)$ or A is diagonal and show that our bounds are realistic. In all figures we depict the loss of A-orthogonality $\|I - \bar{Z}^T A\bar{Z}\|$ and the factorization error $\|Z^{(0)} - \bar{Z}\bar{U}\|$ with respect to the condition numbers $\kappa(A)$ or $\kappa(A^{1/2}Z^{(0)})$ 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 all figures correspond to relevant bounds developed in this paper (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 A-orthogonality and $u\|Z^{(0)}\|$, $u\|\bar{Z}\|\|A^{1/2}Z^{(0)}\|$ and $u\kappa^{1/2}(A)\|Z^{(0)}\|$ for the factorization error, respectively).

The first sequence of problems with dimension $n = 8$ (denoted as Problem 1) is generated with the fixed matrix A given as a matrix square root of the Hilbert matrix $A = \text{sqrtm(hilb(8))} = V\Lambda V^T$ ($\kappa(A) \approx 1.2\ 10^5$) and with the matrices $Z_i^{(0)}$ constructed as $Z_i^{(0)} = V\Lambda^{-1/2}U_i$ such that U_i is upper triangular and $\kappa(U_i) \approx 10^i$, $i = 0, \ldots, 15$. It is clear from the definition that in exact arithmetic the orthogonal factor is equal to $Z_i = V\Lambda^{-1/2}$ and the triangular factor is identical to the matrix U_i with $\kappa(A^{1/2}Z_i^{(0)}) = \kappa(U_i)$. Moreover we assume that the columns of Z_i are ordered with respect to increasing eigenvalue of A. It is clear from Fig. 1 that the loss of A-orthogonality between computed vectors in the EIG implementation is on the level of $\kappa(A)u$ and it does not depend on the condition number of $A^{1/2}Z^{(0)}$. The same applies to the CGS2 algorithm. The behavior of CGS and AINV is very similar; they both generate vectors with loss of A-orthogonality approaching the theoretical bound $u\kappa(A)\kappa(A^{1/2}Z^{(0)})\kappa(Z^{(0)})$ as predicted by the theory, while for the MGS algorithm it approaches the theoretical bound $u\kappa(A)\kappa(A^{1/2}Z^{(0)})$. Figure 1 also shows the 2-norm of the error in the factorization measured by $\|Z_i^{(0)} - \bar{Z}_i\bar{U}_i\|$. The results confirm that the EIG implementation is significantly worse in terms of the error and it approximately scales as $u\kappa^{1/2}(A)\|Z_i^{(0)}\|$. All the other algorithms behave similarly and correspond to the significantly better bound $\mathcal{O}(n^{3/2})u\|\bar{Z}_i\|\|\bar{U}_i\|$.

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

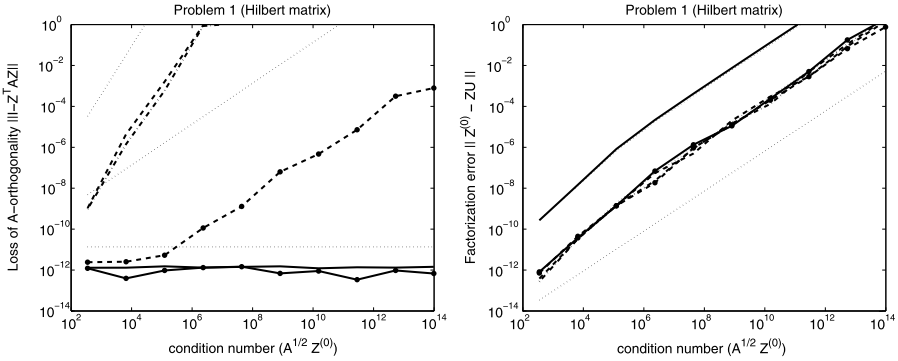

Fig. 1 Figure on the *left* depicts the loss of A-orthogonality $\|\bar{Z}^T A\bar{Z} - I\|$ for Problem 1: EIG implementation (*solid lines*), MGS algorithm (*dashed lines with bold dots*), CGS algorithm (*dash-dotted lines*), AINV orthogonalization (*dashed lines*) and CGS2 algorithm (*solid lines with bold dots*). The *dotted straight lines* represent the bounds $u\kappa(A)$, $u\kappa(A)\kappa(A^{1/2}Z^{(0)})$ and $u\kappa(A)\kappa(A^{1/2}Z^{(0)})\kappa(Z^{(0)})$. Figure on the *right* shows the factorization error $\|Z^{(0)} - \bar{Z}\bar{U}\|$ for Problem 1: EIG implementation (*solid lines*), MGS algorithm (*dashed lines with bold dots*), CGS algorithm (*dash-dotted lines*), AINV orthogonalization (*dashed lines*) and CGS2 algorithm (*solid lines with bold dots*). The *dotted straight lines* represent the bounds $u\|Z^{(0)}\|$, $u\|\bar{Z}\|\|A^{1/2}Z^{(0)}\|$ and $u\kappa^{1/2}(A)\|Z^{(0)}\|$

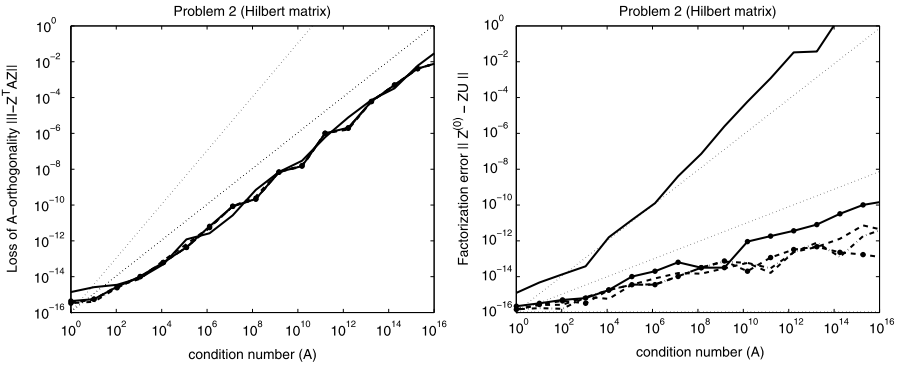

Fig. 2 Figure on the *left* depicts the loss of A-orthogonality $\|\bar{Z}^T A\bar{Z} - I\|$ for Problem 2: EIG implementation (*solid lines*), MGS algorithm (*dashed lines with bold dots*), CGS algorithm (*dash-dotted lines*), AINV orthogonalization (*dashed lines*) and CGS2 algorithm (*solid lines with bold dots*). The *dotted straight lines* represent the bounds $u\kappa(A)$, $u\kappa(A)\kappa(A^{1/2}Z^{(0)})$ and $u\kappa(A)\kappa(A^{1/2}Z^{(0)})\kappa(Z^{(0)})$. Figure on the *right* shows the factorization error $\|Z^{(0)} - \bar{Z}\bar{U}\|$ for Problem 2: EIG implementation (*solid lines*), MGS algorithm (*dashed lines with bold dots*), CGS algorithm (*dash-dotted lines*), AINV orthogonalization (*dashed lines*) and CGS2 algorithm (*solid lines with bold dots*). The *dotted straight lines* represent the bounds $u\|Z^{(0)}\|$, $u\|\bar{Z}\|\|A^{1/2}Z^{(0)}\|$ and $u\kappa^{1/2}(A)\|Z^{(0)}\|$

$A_i = V\Lambda^{i/10}V^T$ with $\kappa(A_i) \approx 10^i$, $i = 0, \ldots, 15$. We set now $Z_i^{(0)} = V\Lambda^{-i/20}$ leading to $U_i = A_i^{1/2}Z_i^{(0)} = I$. As we see from Fig. 2 the results are qualitatively different, well-conditioned triangular factors U_i lead to similar orthogonality of computed vectors for all algorithms that is close to our theoretical bound $u\kappa(A_i)$, whereas the factorization error (also depicted in Fig. 2) behaves similarly as in Problem 1.

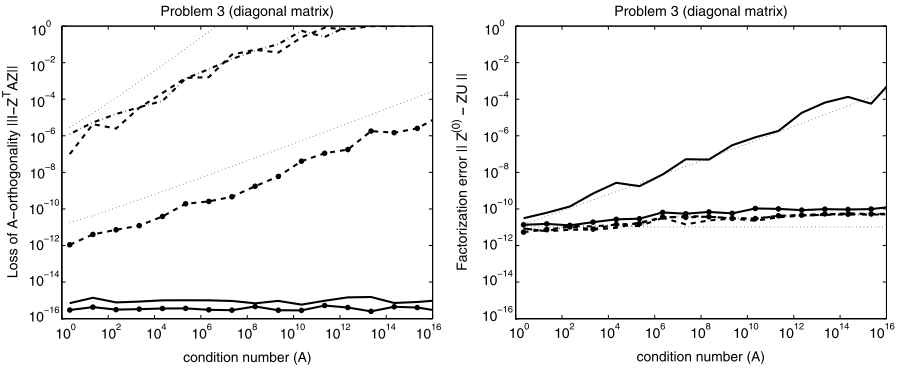

Fig. 3 Figure on the *left* depicts the loss of A-orthogonality $\|\bar{Z}^T A\bar{Z} - I\|$ for Problem 3: EIG implementation (*solid lines*), MGS algorithm (*dashed lines with bold dots*), CGS algorithm (*dash-dotted lines*), AINV orthogonalization (*dashed lines*) and CGS2 algorithm (*solid lines with bold dots*). The *dotted straight lines* represent the bounds $u\kappa(A)$, $u\kappa(A)\kappa(A^{1/2}Z^{(0)})$ and $u\kappa(A)\kappa(A^{1/2}Z^{(0)})\kappa(Z^{(0)})$. Figure on the *right* shows the factorization error $\|Z^{(0)} - \bar{Z}\bar{U}\|$ for Problem 3: EIG implementation (*solid lines*), MGS algorithm (*dashed lines with bold dots*), CGS algorithm (*dash-dotted lines*), AINV orthogonalization (*dashed lines*) and CGS2 algorithm (*solid lines with bold dots*). The *dotted straight lines* represent the bounds $u\|Z^{(0)}\|$, $u\|\bar{Z}\|\|A^{1/2}Z^{(0)}\|$ and $u\kappa^{1/2}(A)\|Z^{(0)}\|$

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 ($\kappa(Z^{(0)}) \approx 10^5$) and $A^{(i)}$ is a diagonal matrix with $\kappa(A_i) \approx 10^i$, $i = 0, \ldots, 15$. The results are plotted on Fig. 3. They 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 $\|Z^{(0)} - \bar{Z}\bar{U}\| \leq \mathcal{O}(n^{3/2})u\|Z^{(0)}\|$ 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 (for simplicity we omit the terms proportional to u^2). 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 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

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	$\mathcal{O}(m^{\frac{5}{2}})u\kappa(A)$	$\mathcal{O}(m^{\frac{5}{2}})u\kappa^{\frac{1}{2}}(A)\|Z^{(0)}\|$
Diagonal A	$\mathcal{O}(mn^{\frac{3}{2}})u$	$\mathcal{O}(mn^{\frac{3}{2}})u\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$
Classical Gram–Schmidt with reorthogonalization		
Assumption	$\|I - \bar{Z}^T A\bar{Z}\|$	$\|Z^{(0)} - \bar{Z}\bar{U}\|$
General A $\mathcal{O}(m^{\frac{3}{2}}n)u\kappa(A)\kappa(A^{\frac{1}{2}}Z^{(0)}) < 1$	$\mathcal{O}(m^{\frac{5}{2}})u\kappa(A)$	$\mathcal{O}(m^{\frac{5}{2}})u\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$
Diagonal A $\mathcal{O}(mn)u\kappa(A^{\frac{1}{2}}Z^{(0)}) < 1$	$\mathcal{O}(mn)u$	$\mathcal{O}(n^{\frac{3}{2}})u\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$
Modified Gram–Schmidt orthogonalization		
Assumption	$\|I - \bar{Z}^T A\bar{Z}\|$	$\|Z^{(0)} - \bar{Z}\bar{U}\|$
General A $\mathcal{O}(m^{\frac{3}{2}}n)u\kappa(A)\kappa(A^{\frac{1}{2}}Z^{(0)}) < 1$	$\mathcal{O}(m^{3/2}n)u\|A\|\,\|\bar{Z}\|$ $\times \max_{j\le i} \frac{\|\bar{z}_i^{(j-1)}\|}{\|\bar{z}_i^{(j-1)}\|_A}\kappa(A^{1/2}Z^{(0)})$	$\mathcal{O}(n^{\frac{3}{2}})u\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$
Diagonal A $\mathcal{O}(mn)u\kappa(A^{\frac{1}{2}}Z^{(0)}) < 1$	$\mathcal{O}(mn)u\kappa(A^{\frac{1}{2}}Z^{(0)})$	$\mathcal{O}(n^{\frac{3}{2}})u\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$
Classical Gram–Schmidt and AINV orthogonalizations		
Assumption	$\|I - \bar{Z}^T A\bar{Z}\|$	$\|Z^{(0)} - \bar{Z}\bar{U}\|$
General A $\mathcal{O}(m^{\frac{3}{2}}n)u\kappa(A)\kappa(A^{\frac{1}{2}}Z^{(0)}) < 1$	$\mathcal{O}(m^{3/2}n)u\|A\|^{1/2}\,\|\bar{Z}\|$ $\times \kappa(A^{1/2}Z^{(0)})\kappa^{1/2}(A)\kappa(Z^{(0)})$	$\mathcal{O}(n^{\frac{3}{2}})u\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$
Diagonal A $\mathcal{O}(mn)u\kappa(A^{\frac{1}{2}}Z^{(0)}) < 1$	$\mathcal{O}(mn)u\kappa^2(A^{\frac{1}{2}}Z^{(0)})$	$\mathcal{O}(n^{\frac{3}{2}})u\|A^{-1}\|^{\frac{1}{2}}\|A^{\frac{1}{2}}Z^{(0)}\|$

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.

Acknowledgements The authors would like to thank for the fruitful discussion and useful comments to G. Meurant and S. Gratton as well as to M. Hochstenbach and anonymous referee.

References

1. Abdelmalek, N.I.: Roundoff error analysis for Gram–Schmidt method and solution of linear least squares problems. BIT Numer. Math. **11**(4), 354–367 (1971)
2. Barlow, J.L., Smoktunowicz, A.: Reorthogonalized Block Classical Gram–Schmidt. Available electronically at http://arxiv.org/pdf/1108.4209.pdf
3. Barrlund, A.: Perturbation bounds for the LDL^T and LU decompositions. BIT Numer. Math. **31**(2), 358–363 (1991)
4. Benzi, M.: Preconditioning techniques for large linear systems: a survey. J. Comput. Phys. **182**(2), 418–477 (2002)
5. Benzi, M., Cullum, J.K., Tůma, M.: Robust approximate inverse preconditioning for the conjugate gradient method. SIAM J. Sci. Comput. **22**(4), 1318–1332 (2000)
6. Benzi, M., Meyer, C.D., Tůma, M.: A sparse approximate inverse preconditioner for the conjugate gradient method. SIAM J. Sci. Comput. **17**(5), 1135–1149 (1996)
7. Benzi, M., Tůma, M.: A robust incomplete factorization preconditioner for positive definite matrices. Numer. Linear Algebra Appl. **10**(5–6), 385–400 (2003)
8. Björck, Å.: Solving linear least squares problems by Gram–Schmidt orthogonalization. BIT Numer. Math. **7**(1), 1–21 (1967)
9. Björck, Å.: Numerics of Gram–Schmidt orthogonalization. Linear Algebra Appl. **197–198**, 297–316 (1994)
10. Björck, Å.: Numerical Methods for Least Squares Problems. SIAM, Philadelphia (1996)
11. Challacombe, M.: A simplified density matrix minimization for linear scaling self-consistent field theory. J. Chem. Phys. **110**(5), 2332–2342 (1999)
12. Fox, L., Huskey, H.D., Wilkinson, J.H.: Notes on the solution of algebraic linear simultaneous equations. Q. J. Mech. Appl. Math. **1**(1), 149–173 (1948)
13. Giraud, L., Langou, J., Rozložník, M.: The loss of orthogonality in the Gram–Schmidt orthogonalization process. Comput. Math. Appl. **50**(7), 1069–1075 (2005)
14. Giraud, L., Langou, J., Rozložník, M., van den Eshof, J.: Rounding error analysis of the classical Gram–Schmidt orthogonalization process. Numer. Math. **101**(1), 97–100 (2005)
15. Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences, 3rd edn. Johns Hopkins University Press, Baltimore (1996)
16. 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)
17. Gulliksson, M.: On the modified Gram–Schmidt algorithm for weighted and constrained linear least squares problems. BIT Numer. Math. **35**(4), 453–468 (1995)
18. Gulliksson, M., Wedin, P.-Å.: Modifying the QR-decomposition to constrained and weighted linear least squares. SIAM J. Matrix Anal. Appl. **13**(4), 1298–1313 (1992)
19. Hestenes, M.R.: Inversion of matrices by biorthogonalization and related results. J. SIAM **6**(1), 51–90 (1958)
20. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bur. Stand. **49**(6), 409–435 (1952)
21. Higham, N.J.: Accuracy and Stability of Numerical Algorithms, 2nd edn. SIAM, Philadelphia (2002)
22. Householder, A.S.: Terminating and nonterminating iterations for solving linear systems. J. SIAM **3**(2), 67–72 (1955)
23. Kharchenko, S.A., Kolotilina, L.Y., Nikishin, A.A., 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)

24. Lawson, C.L., Hanson, R.J.: Solving Least Squares Problems. Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs (1974)
25. Martin, R.S., Wilkinson, J.H.: Reduction of the symmetric eigenproblem $Ax = \lambda Bx$ and related problems to standard form. In: Handbook Series Linear Algebra. Numer. Math., vol. 11(2), pp. 99–110 (1968)
26. Mazzia, A., Pini, G.: Numerical performance of preconditioning techniques for the solution of complex sparse linear systems. Commun. Numer. Methods Eng. **19**(1), 37–48 (2003)
27. Morris, J.: An escalator process for the solution of linear simultaneous equations. Philos. Mag. **37**(7), 106–120 (1946)
28. Saberi Najafi, H., 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)
29. Sun, J.-G.: Perturbation bounds for the Cholesky and QR factorizations. BIT Numer. Math. **31**, 341–352 (1991)
30. Parlett, B.N.: The Symmetric Eigenvalue Problem. Prentice-Hall Series in Computational Mathematics. Prentice-Hall, Englewood Cliffs (1980)
31. Pietrzykowski, T.: Projection method. Prace ZAM Ser. A **8**, 9 (1960)
32. Purcell, E.W.: The vector method of solving simultaneous linear equations. J. Math. Phys. **32**, 150–153 (1953)
33. Smoktunowicz, A., Barlow, J.L., Langou, J.: A note on the error analysis of classical Gram–Schmidt. Numer. Math. **105**(2), 299–313 (2006)
34. Thomas, S.J.: A block algorithm for orthogonalization in elliptic norms. Lect. Notes Comput. Sci. **634**, 379–385 (1992)
35. Thomas, S.J., Zahar, R.V.M.: Efficient orthogonalization in the M-norm. Congr. Numer. **80**, 23–32 (1991)
36. Thomas, S.J., Zahar, R.V.M.: An analysis of orthogonalization in elliptic norms. Congr. Numer. **86**, 193–222 (1992)
37. Wilkinson, J.H.: The Algebraic Eigenvalue Problem. Clarendon Press, Oxford (1965)
38. Yin, J.-F., Yin, G.-J., Ng, M.: On adaptively accelerated Arnoldi method for computing PageRank. Numer. Linear Algebra Appl. **19**(1), 73–85 (2012)