Exploiting variable precision in GMRES

Serge Gratton, Ehouarn Simon, David Titley-Peloquin, Philippe Toint

Exploiting variable precision in GMRES*

Serge Gratton* Ehouarn Simon† David Titley-Peloquin‡
Philippe Toint§

February 20, 2020

Abstract

We describe how variable precision floating-point arithmetic can be used to compute inner products in the iterative solver GMRES. We show how the precision of the inner products carried out in the algorithm can be reduced as the iterations proceed, without affecting the convergence rate or final accuracy achieved by the iterates. Our analysis explicitly takes into account the resulting loss of orthogonality in the Arnoldi vectors. We also show how inexact matrix-vector products can be incorporated into this setting.

Keywords— variable precision arithmetic, inexact inner products, inexact matrix-vector products, Arnoldi algorithm, GMRES algorithm
AMS Subject Codes— 15A06, 65F10, 65F25, 97N20

1 Introduction

As highlighted in a recent SIAM News article [17], there is growing interest in the use of variable precision floating-point arithmetic in numerical algorithms. (Other recent references include [4, 14, 15, 16, 18, 19] to cite only a few.) In this paper, we describe how variable precision arithmetic can be exploited in the iterative solver GMRES [27]. We show that the precision of some floating-point operations carried

*The work of Ehouarn Simon was partially supported by the French National program LEFE (Les Enveloppes Fluides et l'Environnement). The work of Serge Gratton and Phillipe Toint was partially supported by the 3IA Artificial and Natural Intelligence Toulouse Institute, French "Investing for the Future - PIA3" program under the Grant agreement ANR-19-PI3A-0004.

*INPT-IRIT-ENSEEIHT, Toulouse, France (serge.gratton@toulouse-inp.fr).
†INPT-IRIT-ENSEEIHT, Toulouse, France (ehouarn.simon@toulouse-inp.fr).
‡McGill University, Montreal, Canada (david.titley-peloquin@mcgill.ca).
§The University of Namur, Namur, Belgium (philippe.toint@unamur.be).

out in the algorithm can be reduced as the iterations proceed, without affecting the convergence rate or final accuracy achieved by the iterates.

There is already a literature on the use of inexact matrix-vector products in GMRES and other Krylov subspace methods; see, e.g., [28, 7, 3, 11, 29, 12] and the references therein. This work is not a simple extension of such results. To illustrate, suppose that all arithmetic operations are performed exactly, except the matrix-vector products. Then one obtains an inexact Arnoldi relation

$$AV_k + E_k = V_{k+1}H_k, \qquad V_k^T V_k = I. \tag{1}$$

On the other hand, if only inner products are performed inexactly, the Arnoldi relation continues to hold but the orthogonality of the Arnoldi vectors is lost:

$$AV_k = V_{k+1}H_k, \qquad V_k^T V_k = I - F_k. \tag{2}$$

Thus, to understand the convergence behaviour and maximum attainable accuracy of GMRES implemented with inexact inner products, it is absolutely necessary to understand the resulting loss of orthogonality in the Arnoldi vectors. We adapt techniques used in the rounding-error analysis of the Modified Gram-Schmidt (MGS) algorithm (see [1, 2] or [20] for a more recent survey) and of the MGS-GMRES algorithm (see [6, 13, 22]).

We focus on inexact inner products and matrix-vector products (as opposed to the other saxpy operations involved in the algorithm) because these are the two most time-consuming operations in parallel computations. The rest of the paper is organized as follows. We start with a brief discussion of GMRES in non-standard inner products in Section 2. Next, in Section 3, we analyze GMRES with inexact inner products. We then show how inexact matrix-vector products can be incorporated into this setting in Section 4. Some numerical examples are presented in Sections 5 and 6.

2 GMRES in weighted inner products

Shown below is the Arnoldi algorithm, with $\langle y, z\rangle = y^T z$ denoting the standard Euclidean inner product.

After k steps of the algorithm are performed in exact arithmetic, the output is $V_{k+1} = [v_1, \ldots, v_{k+1}] \in \mathbb{R}^{n\times(k+1)}$ and upper-Hessenberg $H_k \in \mathbb{R}^{(k+1)\times k}$ such that

$$v_1 = \frac{b}{\beta}, \qquad AV_k = V_{k+1}H_k, \qquad V_k^T V_k = I_k.$$

The columns of V_k form an orthonormal basis for the Krylov subspace

$$\mathcal{K}_k(A, b) = \text{span}\big\{\, b, Ab, A^2b, \ldots, A^{k-1}b \,\big\}.$$

In GMRES, we restrict x_k to this subspace: $x_k = V_k y_k$, where $y_k \in \mathbb{R}^k$ is the solution of

$$\min_y \|b - AV_k y\|_2 = \min_y \|V_{k+1}(\beta e_1 - H_k y)\|_2 = \min_y \|\beta e_1 - H_k y\|_2.$$

Algorithm 1 Arnoldi algorithm

Require: $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$

$\beta = \sqrt{\langle b, b\rangle}$
$v_1 = b/\beta$
for $j = 1, 2, \ldots$ **do**
  $w_j = Av_j$
  for $i = 1, \ldots, j$ **do**
    $h_{ij} = \langle v_i, w_j\rangle$
    $w_j = w_j - h_{ij}v_i$
  end for
  $h_{j+1,j} = \sqrt{\langle w_j, w_j\rangle}$
  $v_{j+1} = w_j/h_{j+1,j}$
end for

It follows that

$$\begin{aligned} x_k &= V_k y_k = V_k(H_k^T H_k)^{-1} H_k^T(\beta e_1) = V_k H_k^\dagger(\beta e_1), \\ r_k &= b - Ax_k = V_{k+1}(\beta e_1 - H_k y_k) = V_{k+1}(I - H_k H_k^\dagger)\beta e_1. \end{aligned} \tag{3}$$

Any given symmetric positive definite matrix W defines a weighted inner product $\langle y, z\rangle_W = y^T W z$ and associated norm $\|z\|_W = \sqrt{\langle z, z\rangle_W}$. Suppose we use this inner product instead of the standard Euclidean inner product in the Arnoldi algorithm. We use tildes to denote the resulting quantities in the algorithm. After k steps, the result is $\widetilde{V}_{k+1} = [\widetilde{v}_1, \ldots, \widetilde{v}_{k+1}]$ and upper-Hessenberg $\widetilde{H}_k \in \mathbb{R}^{(k+1)\times k}$ such that

$$\widetilde{v}_1 = \frac{b}{\|b\|_W} = \frac{b}{\widetilde{\beta}}, \qquad A\widetilde{V}_k = \widetilde{V}_{k+1}\widetilde{H}_k, \qquad \widetilde{V}_k^T W \widetilde{V}_k = I_k.$$

The columns of $\widetilde{V}_k$ form a W-orthonormal basis for $\mathcal{K}_k(A, b)$. Let $\widetilde{x}_k = \widetilde{V}_k\widetilde{y}_k$, where $\widetilde{y}_k \in \mathbb{R}^k$ is the solution of

$$\min_y \|b - A\widetilde{V}_k y\|_W = \min_y \|\widetilde{V}_{k+1}(\widetilde{\beta}e_1 - \widetilde{H}_k y)\|_W = \min_y \|\widetilde{\beta}e_1 - \widetilde{H}_k y\|_2,$$

so that

$$\widetilde{x}_k = \widetilde{V}_k\widetilde{H}_k^\dagger(\widetilde{\beta}e_1), \qquad \widetilde{r}_k = b - A\widetilde{x}_k = \widetilde{V}_{k+1}(I - \widetilde{H}_k\widetilde{H}_k^\dagger)\widetilde{\beta}e_1.$$

We denote the above algorithm W-GMRES.

Let x_k and $\widetilde{x}_k$ denote the iterates computed by standard GMRES and W-GMRES, respectively, with corresponding residual vectors r_k and $\widetilde{r}_k$. It is well known that

$$1 \le \frac{\|\widetilde{r}_k\|_2}{\|r_k\|_2} \le \sqrt{\kappa_2(W)}. \tag{4}$$

See e.g. [26] for a proof. Thus, if $\kappa_2(W)$ is small, the Euclidean norm of the residual vector in W-GMRES converges at essentially the same rate as in standard GMRES. A similar result [8, Theorem 4] holds for the residual computed in the quasi-minimal residual method [10, 9].

3 GMRES with inexact inner products

3.1 Preliminary results

Suppose the inner products in the Arnoldi algorithm are computed inexactly, i.e., line 6 in Algorithm 1 is replaced by

$$h_{ij} = v_i^T w_j + \eta_{ij}, \tag{5}$$

with $|\eta_{ij}|$ bounded by some tolerance. Our main contribution is to show precisely how large each η_{ij} can be without affecting the convergence of GMRES. Throughout we assume that all arithmetic operations in GMRES are performed exactly, except for the above inner products.

It is straightforward to show that despite the inexact inner products in (5), the relation $AV_k = V_{k+1}H_k$ continues to hold. On the other hand, the orthogonality of the Arnoldi vectors is lost. We have

$$[b, AV_k] = V_{k+1}[\beta e_1, H_k], \qquad V_{k+1}^T V_{k+1} = I_{k+1} + F_k. \tag{6}$$

The relation between each η_{ij} and the overall loss of orthogonality F_k is very difficult to understand. To simplify the analysis we suppose that each v_j is normalized exactly. (This is not an uncommon assumption; see, e.g., [1] and [21].) Under this simplification, we have

$$F_k = \bar{U}_k + \bar{U}_k^T, \quad \bar{U}_k = \begin{bmatrix} 0_{k\times 1} & U_k \\ 0_{1\times 1} & 0_{1\times k} \end{bmatrix}, \quad U_k = \begin{bmatrix} v_1^T v_2 & \dots & v_1^T v_{k+1} \\ & \ddots & \vdots \\ & & v_k^T v_{k+1} \end{bmatrix}, \tag{7}$$

i.e., $U_k \in \mathbb{R}^{k\times k}$ contains the strictly upper-triangular part of F_k. Define

$$N_k = \begin{bmatrix} \eta_{11} & \dots & \eta_{1k} \\ & \ddots & \vdots \\ & & \eta_{kk} \end{bmatrix}, \quad R_k = \begin{bmatrix} h_{21} & \dots & h_{2k} \\ & \ddots & \vdots \\ & & h_{k+1,k} \end{bmatrix}. \tag{8}$$

Note that R_k must be invertible if $h_{j+1,j} \neq 0$ for $j = 1, \dots, k$, in other words, if GMRES has not terminated by step k. (We assume that GMRES does not breakdown by step k.) Following Björck's seminal rounding error analysis of MGS [1], it can be shown that

$$N_k = -[0, U_k]H_k = -U_k R_k. \tag{9}$$

For completeness, a proof of (9) is provided in the appendix.

Additionally, in order to understand how $\|F_k\|_2$ increases as the residual norm decreases, we will need the following rather technical lemma. The relationship (10) is well know (see for example [28, Lemma 5.1]) while (12) is essentially a special case of [23, Theorem 4.1]. We defer the proof to the appendix.

Lemma 1. *Let y_j and t_j be the least squares solution and residual vector of*

$$\min_y \|\beta e_1 - H_j y\|_2,$$

for $j = 1, \ldots, k$. Then

$$|e_j^T y_k| \leq \frac{\|t_{j-1}\|_2}{\sigma_{\min}(H_k)}. \tag{10}$$

In addition, given $\epsilon > 0$, let D_k be any nonsingular matrix such that

$$\|D_k\|_2 \leq \frac{\sigma_{\min}(H_k)\epsilon\|b\|_2}{\sqrt{2}\|t_k\|_2}. \tag{11}$$

Then

$$\frac{\|t_k\|_2}{\left(\epsilon^2\|b\|_2^2 + 2\|D_k y_k\|_2^2\right)^{1/2}} \leq \sigma_{\min}\left(\left[\epsilon^{-1}e_1, H_k D_k^{-1}\right]\right) \leq \frac{\|t_k\|_2}{\epsilon\|b\|_2}. \tag{12}$$

Finally, although the columns of V_{k+1} in (6) are not orthonormal in the standard Euclidean inner product, we will use the fact that there exists an inner product in which they are orthonormal. The proof of the following lemma is given in the appendix.

Lemma 2. *Consider a given matrix $Q \in \mathbb{R}^{n\times k}$ of rank k such that*

$$Q^T Q = I_k - F. \tag{13}$$

If $\|F\|_2 \leq \delta$ for some $\delta \in (0, 1)$, then there exists a matrix M such that $I_n + M$ is symmetric positive definite and

$$Q^T(I_n + M)Q = I_k. \tag{14}$$

In other words, the columns of Q are exactly orthonormal in an inner product defined by $I_n + M$. Furthermore,

$$\kappa_2(I_n + M) \leq \frac{1+\delta}{1-\delta}. \tag{15}$$

Note that $\kappa_2(I_n+M)$ remains small even for values of δ close to 1. For example, suppose $\|I_k - Q^T Q\|_2 = \delta = 1/2$, indicating an extremely severe loss of orthogonality. Then $\kappa_2(I_n + M) \leq 3$, so Q still has exactly orthonormal columns in an inner product defined by a very well-conditioned matrix.

Remark 1. *Paige and his coauthors [2, 21, 25] have developed an alternative measure of loss of orthogonality. Given $Q \in \mathbb{R}^{n\times k}$ with normalized columns, the measure is $\|S\|_2$, where $S = (I+U)^{-1}U$ and U is the strictly upper-triangular part of Q^TQ. Additionally, orthogonality can be recovered by augmentation: the matrix $P = \begin{bmatrix} S \\ Q(I-S) \end{bmatrix}$ has orthonormal columns. This measure was used in the groundbreaking rounding error analysis of the MGS-GMRES algorithm [22]. In the present paper, under the condition $\|F\|_2 \le \delta < 1$, we use the measure $\|F\|_2$ and recover orthogonality in the $(I+M)$ inner product. However, Paige's approach is likely to be the most appropriate for analyzing the Lanczos and conjugate gradient algorithms, in which orthogonality is quickly lost and $\|F\|_2 > 1$ long before convergence.*

3.2 A strategy for bounding the η_{ij}

We now show how to bound the error η_{ij} in (5) to ensure that the convergence of the GMRES is not affected by the inexact inner products.

The following theorem shows how the convergence of GMRES with inexact inner products relates to that of exact GMRES. The idea is similar to [22, Section 5], in which the quantity $\|E_kR_k^{-1}\|_F$ must be bounded, where R_k is the matrix in (8) and E_k is a matrix containing rounding errors.

Theorem 1. *Let $x_k^{(\mathrm{e})}$ denote the k-th iterate of standard GMRES, performed exactly, with residual vector $r_k^{(\mathrm{e})}$. Now suppose that the Arnoldi algorithm is run with inexact inner products as in (5), so that (6)–(9) hold, and let x_k and r_k denote the resulting GMRES iterate and residual vector. Let y_k and t_k be the least squares solution and residual vector of*

$$\min_y \|\beta e_1 - H_k y\|_2.$$

If for all steps $j = 1, \ldots, k$ of GMRES all inner products are performed inexactly as in (5) with tolerances bounded by

$$|\eta_{ij}| \le \eta_j \equiv \frac{\phi_j \epsilon \sigma_{\min}(H_k)}{\sqrt{2}} \frac{\|b\|_2}{\|t_{j-1}\|_2} \tag{16}$$

for any $\epsilon \in (0,1)$ and any positive numbers ϕ_j such that $\sum_{j=1}^k \phi_j^2 \le 1$, then at step k either

$$1 \le \frac{\|r_k\|_2}{\|r_k^{(\mathrm{e})}\|_2} \le \sqrt{3}, \tag{17}$$

or

$$\frac{\|t_k\|_2}{\|b\|_2} \le 6k\epsilon, \tag{18}$$

implying that GMRES has converged to a relative residual of $6k\epsilon$.

Proof. If (16) holds, then in (8)

$$|N_k| \leq \begin{bmatrix} \eta_1 & \eta_2 & \dots & \eta_k \\ & \eta_2 & \dots & \eta_k \\ & & \ddots & \vdots \\ & & & \eta_k \end{bmatrix} = E_k D_k,$$

where E_k is an upper-triangular matrix containing only ones in its upper-triangular part, so that $\|E_k\|_2 \leq k$, and $D_k = \operatorname{diag}(\eta_1, \dots, \eta_k)$. Then,

$$\begin{aligned} \|N_k R_k^{-1}\|_2 &\leq \|N_k D_k^{-1}\|_2 \|D_k R_k^{-1}\|_2 \\ &\leq \|E_k\|_2 \|D_k R_k^{-1}\|_2 \leq k \|(R_k D_k^{-1})^{-1}\|_2. \end{aligned} \tag{19}$$

Let h_k^T denote the first row of H_k, so that $H_k = \begin{bmatrix} h_k^T \\ R_k \end{bmatrix}$. For any $\epsilon > 0$ we have

$$\begin{aligned} \sigma_{\min}(R_k D_k^{-1}) &= \min_{\|u\|_2 = \|v\|_2 = 1} u^T R_k D_k^{-1} v \\ &= \min_{\|u\|_2 = \|v\|_2 = 1} [0, u^T] \begin{bmatrix} \epsilon^{-1} & h_k^T D_k^{-1} \\ 0 & R_k D_k^{-1} \end{bmatrix} \begin{bmatrix} 0 \\ v \end{bmatrix} \\ &\geq \min_{\|u\|_2 = \|v\|_2 = 1} u^T \begin{bmatrix} \epsilon^{-1} & h_k^T D_k^{-1} \\ 0 & R_k D_k^{-1} \end{bmatrix} v \\ &= \sigma_{\min}\left(\left[\epsilon^{-1} e_1, H_k D_k^{-1} \right] \right). \end{aligned}$$

Therefore,

$$\|(R_k D_k^{-1})^{-1}\|_2 = \frac{1}{\sigma_{\min}(R_k D_k^{-1})} \leq \frac{1}{\sigma_{\min}\left(\left[\epsilon^{-1} e_1, H_k D_k^{-1} \right] \right)}.$$

Notice that if the η_j are chosen as in (16), D_k automatically satisfies (11). Using the lower bound in (12), then (10) and (16), we obtain

$$\begin{aligned} \|(R_k D_k^{-1})^{-1}\|_2 &\leq \frac{\left(\epsilon^2 \|b\|_2^2 + 2 \|D_k y_k\|_2^2 \right)^{1/2}}{\|t_k\|_2} \\ &= \frac{\left(\epsilon^2 \|b\|_2^2 + 2 \sum_{j=1}^k \eta_j^2 (e_j^T y_k)^2 \right)^{1/2}}{\|t_k\|_2} \\ &\leq \frac{\left(\epsilon^2 \|b\|_2^2 + \sum_{j=1}^k \phi_j^2 \epsilon^2 \|b\|_2^2 \right)^{1/2}}{\|t_k\|_2} = \frac{\sqrt{2}\epsilon \|b\|_2}{\|t_k\|_2}. \end{aligned}$$

Therefore, in (19),

$$\|N_k R_k^{-1}\|_2 \leq \frac{\sqrt{2} k \epsilon \|b\|_2}{\|t_k\|_2} \leq \frac{6 k \epsilon \|b\|_2}{\|t_k\|_2} \frac{1}{4}.$$

If (18) does not hold, then $\|N_k R_k^{-1}\|_2 \leq 1/4$. From (7) and (9), we have

$$\|F_k\|_2 \leq 2\|U_k\|_2 = 2\|N_k R_k^{-1}\|_2, \tag{20}$$

with the matrix F_k defined in (6). Thus, $\|F_k\|_2 \leq \frac{1}{2} < 1$ and we can apply Lemma 2 with $Q = V_{k+1}$ and $\delta = \frac{1}{2}$. There is a symmetric positive definite matrix $W = I_n + M$ such that

$$[b, AV_k] = V_{k+1}[\beta e_1, H_k], \quad V_{k+1}^T W V_{k+1} = I_{k+1}, \quad \kappa_2(W) \leq \frac{1+\delta}{1-\delta} = 3.$$

The Arnoldi algorithm implemented with inexact inner products has computed an W-orthonormal basis for the Krylov subspace $\mathcal{K}_k(A, b)$. The iterate x_k is the same as the iterate that would have been obtained by running W-GMRES exactly, and (4) implies (17).

Therefore, if the $|\eta_{ij}|$ are bounded by tolerances η_j chosen as in (16), either (17) holds, or (18) holds. □

Theorem 1 can be interpreted as follows. If at all steps $j = 1, 2, \ldots$ of GMRES the inner products are computed inaccurately with tolerances η_j in (16), then convergence at the same rate as exact GMRES is achieved until a relative residual of essentially $k\epsilon$ is reached. Notice that η_j is inversely proportional to the residual norm. This allows the inner products to be computed more and more inaccurately as as the iterations proceed.

3.3 Practical considerations

If no more than $K_{\max}$ iterations are to be performed, we can let $\phi_j = K_{\max}^{-1/2}$ (although more elaborate choices for ϕ_j could be considered; see for example [12]). Then the factor $\phi_j/\sqrt{2}$ in (16) can be absorbed along with the k in (18).

One important difficulty with (16) is that $\sigma_{\min}(H_k)$ is required to pick η_j at the start of step j, but H_k is not available until the final step k. A similar problem occurs in GMRES with inexact matrix-vector products; see [28, 7] and the comments in Section 4. In our experience, is often possible to replace $\sigma_{\min}(H_k)$ in (16) by 1, without significantly affecting the convergence of GMRES. This leads to following:

$$\text{Aggressive threshhold :} \qquad \eta_j = \epsilon \frac{\|b\|_2}{\|t_{j-1}\|_2}, \quad j = 1, 2, \ldots. \tag{21}$$

In exact arithmetic, $\sigma_{\min}(H_k)$ is bounded below by $\sigma_{\min}(A)$. If the smallest singular value of A is known, one can estimate $\sigma_{\min}(H_k) \approx \sigma_{\min}(A)$ in (16), leading to the following:

$$\text{Conservative threshhold :} \qquad \eta_j = \epsilon\, \sigma_{\min}(A) \frac{\|b\|_2}{\|t_{j-1}\|_2}, \quad j = 1, 2, \ldots. \tag{22}$$

This prevents potential early stagnation of the residual norm, but is often unnecessarily stringent. (It goes without saying that if the conservative threshold is less than $u\|A\|_2$, where u is the machine precision, then the criterion is vacuous: according to this criterion no inexact inner products can be carried out at iteration j.) Numerical examples are given in Sections 5 and 6.

4 Incorporating inexact matrix-vector products

As mentioned in the introduction, there is already a literature on the use of inexact matrix-vector products in GMRES. These results are obtained by assuming that the Arnoldi vectors are orthonormal and analyzing the inexact Arnoldi relation

$$AV_k + E_k = V_{k+1}H_k, \qquad V_k^T V_k = I.$$

In practice, however, the computed Arnoldi vectors are very far from being orthonormal, even when all computations are performed in double precision arithmetic; see for example [6, 13, 22].

The purpose of this section is to show that the framework used in [28] and [7] to analyze inexact matrix-vector products in GMRES is still valid when the orthogonality of the Arnoldi vectors is lost, i.e., under the inexact Arnoldi relation

$$AV_k + E_k = V_{k+1}H_k, \qquad V_k^T V_k = I - F_k. \tag{23}$$

We assume that the errors η_{ij} in computing the inner products is sufficiently small that $\|F_k\|_2 \leq \delta < 1$, as per Section 3. Then from Lemma 2 there exists a symmetric positive definite matrix $W = I_n + M \in \mathbb{R}^{n\times n}$ such that $V_{k+1}^T W V_{k+1} = I_{k+1}$, and with singular values bounded as in (35).

4.1 Bounding the residual gap

As in previous sections, we use $x_k = V_k y_k$ to denote the computed GMRES iterate, with $r_k = b - Ax_k$ for the actual residual vector and $t_k = \beta_1 e_1 - H_k y_k$ for the residual vector updated in the GMRES iterations. From

$$\|r_k\|_2 \leq \|r_k - V_{k+1}t_k\|_2 + \|V_{k+1}t_k\|_2,$$

if

$$\max\{\, \|r_k - V_{k+1}t_k\|_2,\, \|V_{k+1}t_k\|_2 \,\} \leq \frac{\epsilon}{2}\|b\|_2 \tag{24}$$

then

$$\|r_k\|_2 \leq \epsilon\|b\|_2. \tag{25}$$

From the fact that the columns of $W^{1/2}V_{k+1}$ are orthonormal as well as (35), we obtain

$$\|V_{k+1}t_k\|_2 \leq \|W^{-1/2}\|_2\|W^{1/2}V_{k+1}t_k\|_2 = \|W\|_2^{-1/2}\|t_k\|_2 \leq \sqrt{1+\delta}\|t_k\|_2.$$

In GMRES, $\|t_k\|_2 \to 0$ with increasing k, which implies that $\|V_{k+1}t_k\|_2 \to 0$ as well. Therefore, we focus on bounding the residual gap $\|r_k - V_{k+1}t_k\|_2$ in order to satisfy (24) and (25).

Suppose the matrix-vector products in the Arnoldi algorithm are computed inexactly, i.e., line 4 in Algorithm 1 is replaced by

$$w_j = (A + \mathcal{E}_j)v_j, \tag{26}$$

where $\|\mathcal{E}_j\|_2 \leq \epsilon_j$ for some given tolerance ϵ_j. Then in (23),

$$E_k = \left[\mathcal{E}_1 v_1, \mathcal{E}_2 v_2, \ldots, \mathcal{E}_k v_k\right]. \tag{27}$$

The following proposition bounds the residual gap at step k in terms of the tolerances ϵ_j, for $j = 1, \ldots, k$. This is a direct corollary of results in [28] and [7].

Proposition 1. *Suppose that the inexact Arnoldi relation* (23) *holds, where* E_k *is given in* (27) *with* $\|\mathcal{E}_j\|_2 \leq \epsilon_j$ *for* $j = 1, \ldots, k$. *Then the resulting residual gap satisfies*

$$\|r_k - V_{k+1} t_k\|_2 \leq \|H_k^\dagger\|_2 \sum_{j=1}^{k} \epsilon_j \|t_{j-1}\|_2. \tag{28}$$

4.2 A strategy for picking the ϵ_j

Proposition 1 suggests the following strategy for picking the tolerances ϵ_j that bound the level of inexactness $\|\mathcal{E}_j\|_2$ in the matrix-vector products in (26). Similarly to Theorem 1, let ϕ_j be any positive numbers such that $\sum_{j=1}^{k} \phi_j = 1$. If for all steps $j = 1, \ldots, k$,

$$\epsilon_j \leq \frac{\phi_j \epsilon \sigma_{\min}(H_k)}{2} \frac{\|b\|_2}{\|t_{j-1}\|_2}, \tag{29}$$

then from (28) the residual gap in (24) satisfies

$$\|r_k - V_{k+1} t_k\|_2 \leq \frac{\epsilon}{2} \|b\|_2.$$

Interestingly, this result is independent of the accuracy of the inner products. Similarly to (16), the criterion for picking ϵ_j at step j involves H_k that is only available at the final step k. A large number of numerical experiments [7, 3] indicate that $\sigma_{\min}(H_k)$ can often be replaced by 1. Absorbing the factor $\phi_j/2$ into ϵ in (29) and replacing $\sigma_{\min}(H_k)$ by 1 or by $\sigma_{\min}(A)$ leads, respectively, to the same aggressive and conservative thresholds for ϵ_j as we obtained for η_j in (21) and in (22). This suggests that matrix-vector products and inner products in GMRES can be computed with the same level of inexactness. We illustrate this with numerical examples in the next section.

5 Numerical examples with emulated accuracy

We illustrate our results with a few numerical examples. We run GMRES with different matrices A and right-hand sides b, and compute the inner products and matrix-vector products inexactly as in (5) and (26), as described in Algorithm 2 below.

Note that the inner product $h_{j+1,j}$ in line 17 of Algorithm 2 is also computed inexactly. In Section 3, to simplify the analysis, we supposed that each v_{j+1} was

normalized exactly. However, our numerical experiments indicate that $h_{j+1,j}$ can be computed with the same level of inexactness as the other inner products at step j.

We pick η_{ij} randomly, uniformly distributed between $-\eta_j$ and η_j, and pick $\mathcal{E}_j$ to be a matrix of independent standard normal random variables, scaled to have norm ϵ_j. Thus we have

$$|\eta_{ij}| \le \eta_j, \qquad \|\mathcal{E}_j\|_2 \le \epsilon_j,$$

for chosen tolerances η_j and ϵ_j. Throughout this first set of experiments, we use the same level of inexactness for inner products and matrix-vector products, i.e., $\eta_j = \epsilon_j$.

In the associated figures, the solid curve is the relative residual $\|b - Ax_k\|_2/\|b\|_2$. For reference, the dashed curve is the relative residual if GMRES is run in double precision. The crossed curve corresponds to the loss of orthogonality $\|F_k\|_2$ in (6). The dotted curve is the chosen tolerance η_j.

5.1 Relationship between η_{ij} and loss of orthogonality

Our first example illustrates the relationship between the errors η_{ij} in the inner products and the loss of orthogonality in the GMRES algorithm.

In this example, A is the 100×100 Grcar matrix of order 5. This is a highly non-normal Toeplitz matrix. The right hand side is $b = A[\sin(1), \ldots, \sin(100)]^T$. Results are shown in Figure 1.

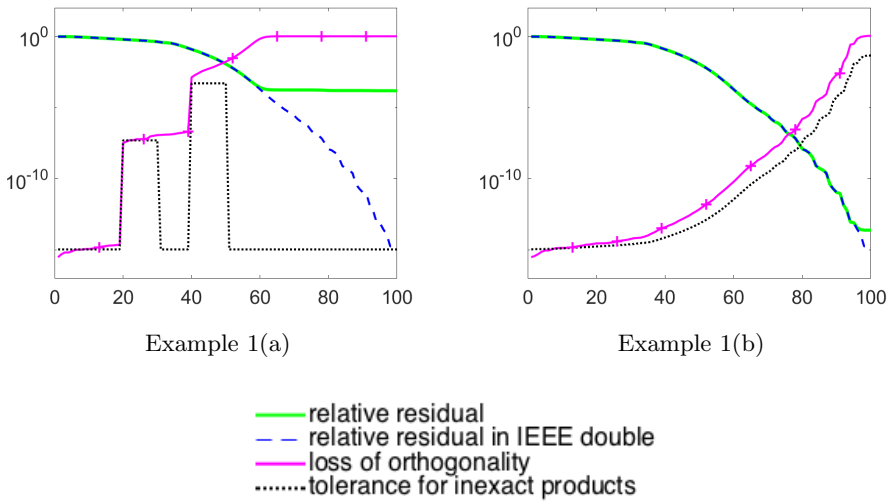

Figure 1: GMRES in variable precision: Grcar matrix.

Algorithm 2 A variable precision GMRES

Require: $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, $\epsilon > 0$, $K_{\max} \in \mathbb{N}$, Conservative $\in \{0, 1\}$

if Conservative **then**
  Compute or estimate $\sigma_{\min}(A)$
end if
$\beta = \sqrt{b^T b}$
$v_1 = b/\beta$
for $j = 1, 2, \ldots, K_{\max}$ **do**
  if Conservative **then**
    Compute η_j and ϵ_j according to the bound (22)
  else
    Compute η_j and ϵ_j according to the bound (21)
  end if
  Compute $w_j = (A + \mathcal{E}_j)v_j$ with $\|\mathcal{E}_j\|_2 \leq \epsilon_j$
  for $i = 1, \ldots, j$ **do**
    Compute $h_{ij} = v_i^T w_j + \eta_{i,j}$ with $|\eta_{i,j}| \leq \eta_j$
    $w_j = w_j - h_{ij} v_i$
  end for
  Compute $h_{j+1,j} = \sqrt{w_j^T w_j + \eta_{j+1,j}}$ with $|\eta_{j+1,j}| \leq \eta_j$
  if $h_{j+1,j} = 0$ **then**
    Goto 27
  end if
  $v_{j+1} = w_j / h_{j+1,j}$
  Compute y_j and $\|t_j\|_2$, the solution and residual of $\min_{y\in\mathbb{R}^j} \|\beta e_1 - H_j y\|_2$.
  if $\|t_j\|_2 < \epsilon$ **then**
    Goto 27
  end if
end for
Set $x_j = V_j y_j$

In Example 1(a),

$$\eta_j = \epsilon_j = \begin{cases} 10^{-8}\|A\|_2, & \text{for } 20 \le j \le 30, \\ 10^{-4}\|A\|_2, & \text{for } 40 \le j \le 50, \\ 2^{-52}\|A\|_2, & \text{otherwise.} \end{cases}$$

The large increase in the inexactness of the inner products at iterations 20 and 40 immediately leads to a large increase in $\|F_k\|_2$. This clearly illustrates the connection between the inexactness of the inner products and the loss of orthogonality in the Arnoldi vectors. As proven in Theorem 1, until $\|F_k\|_2 \approx 1$, the residual norm is the same as it would have been had all computations been performed in double precision. Due to its large increases at iterations 20 and 40, $\|F_k\|_2$ approaches 1, and the residual norm starts to stagnate, long before the relative residual norm reaches the double precision machine precision.

In Example 1(b), the tolerances are chosen according to the aggressive criterion (21) with $\epsilon = 2^{-52}\|A\|_2$. With this choice, $\|F_k\|_2$ does not reach 1, and the residual norm does not stagnate until convergence.

5.2 Conservative vs aggressive thresholds

In our second example, A is the matrix 494_bus from the SuiteSparse matrix collection [5]. This is a 494×494 matrix with condition number $\kappa_2(A) \approx 10^6$. The right hand side is once again $b = A[\sin(1), \ldots, \sin(100)]^T$.

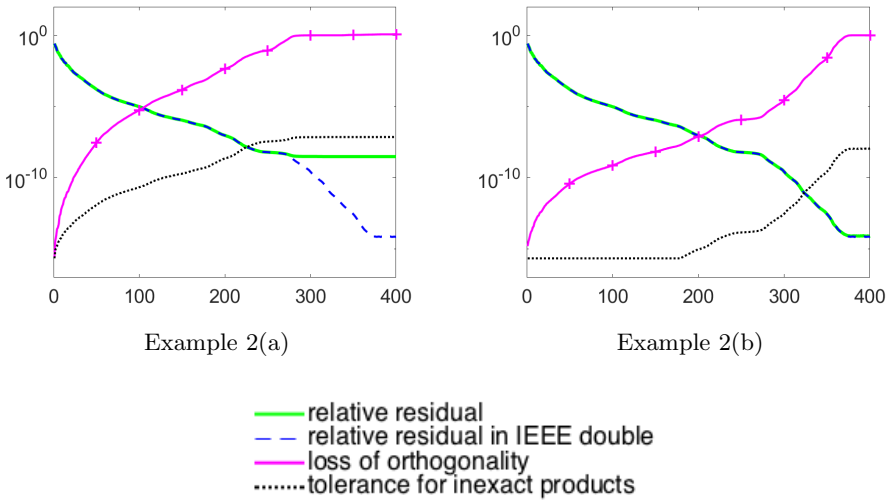

Figure 2: GMRES in variable precision: 494_bus matrix

Results are shown in Figure 2. In Example 2(a), tolerances are chosen according

to the aggressive threshhold (21) with $\epsilon = 2^{-52}\|A\|_2$. In this more ill-conditioned problem, the residual norm starts to stagnate before full convergence to double precision. In Example 2(b), the tolerances are chosen according to the conservative threshhold (22) with $\epsilon = 2^{-52}\|A\|_2$, and there is no more such stagnation. Because of these lower tolerances, the inner products and matrix-vector products have to be performed in double precision until about iteration 200. This example illustrates the tradeoff between the level of inexactness and the maximum attainable accuracy. The more ill-conditioned the matrix A is, the less opportunity there is for performing floating-point operations inexactly in GMRES.

6 Numerical experiments using variable floating-point arithmetic

In this section, we run variable floating-point arithmetic in order to assess the performances of the approach in the context of half, simple and double precisions. These experiments are done with the Julia[1] language which allows to switch on demand between different floating-point types (Float16, Float32 and Float64). We compute the inner products and matrix-vector products in lower floating-point precision $u^{(low)}$ once either the aggressive threshold in (21) or the conservative threshold in (22) has increased above $u^{(low)}\|A\|_2$.

Compared to the previous experiments, we must now take into account the magnitudes of both the vectors and matrix perturbations in order to select the precision of the computation. When no floating-point overflow arise, the perturbation of the matrix is simply computed from the difference between the norm of the matrix stored in Float64 and its conversion to Float16 and Float32. Regarding the inner products, we estimate their magnitude based on the sum of the exponents of the two vectors involved (plus 1 for the product of the mantissa). This explains the oscillating behavior of the accuracies observed in Figures 3 and 4: even if the precision has not changed, the estimated amplitude of the value of the inner products induces changes in the associated perturbations (21) and (22).

We focus only on the 494_bus matrix using both the conservative and agressive thresholds. The tolerances are chosen equal to $\epsilon = 10^{-6}\|A\|$ and $\epsilon = 10^{-12}\|A\|$ in order to illustrate the potential of the algorithm when moderate and high accuracies are required. In Figure 3, when a moderate decrease of the internally-recurred residual is required, we note that the conservative threshold results in a quick degradation of the precision for both the matrix-vector and inner products. All the matrix-vector products are computed in simple precision after 20 iterations, while the inner products start to be computed in simple precision after 30 iterations, precision that is mostly used after 90 iterations. We note a jump in the loss of orthogonality when the simple precision is triggered in the computation of the inner products. However, as expected from the theory, this does not degrade the decrease of the residual which is similar to the one observed with GMRES in double

[1] https://julialang.org/

precision. When the aggressive threshold is used, we note that the simple precision is triggered after a few iterations for both the matrix-vector and inner products. The matrix-vector products are then computed in half precision from iteration 90 to convergence, while the precision of the inner products oscillate between half and single depending on the amplitude of the vectors. The consequences are a complete loss of orthogonality after 100 iterations, which results in a slowing down in the decrease of the internally-recurred residual and a stagnation of the residual.

The effect of requiring a higher accuracy is mainly a delay in the exploitation of the multi-arithmetic. The results shown in Figure 4 are similar to those obtained with a coarser tolerance, except for the delay in triggering the computation in simple, and half precisions. We note again a similar decrease in the residuals compared to the GMRES algorithm in double precision. The loss of orthogonality is more severe when the agressive threshold is used, due to an earlier use of the simple precision in the inner products, as well as the use of the half precision in the latest iterations. The residual does not decrease anymore until the maximum number of iterations is reached.

This once again illustrates the tradeoff between the level of inexactness of the computations and the maximum attainable accuracy.

7 Conclusion

We have shown how inner products can be performed inexactly in MGS-GMRES without affecting the convergence or final achievable accuracy of the algorithm. We have also shown that a known framework for inexact matrix-vector products is still valid despite the loss of orthogonality in the Arnoldi vectors. It would be interesting to investigate the impact of scaling or preconditioning on these results. Additionally, in future work, we plan to address the question of how much computational savings can be achieved by this approach on massively parallel computer architectures.

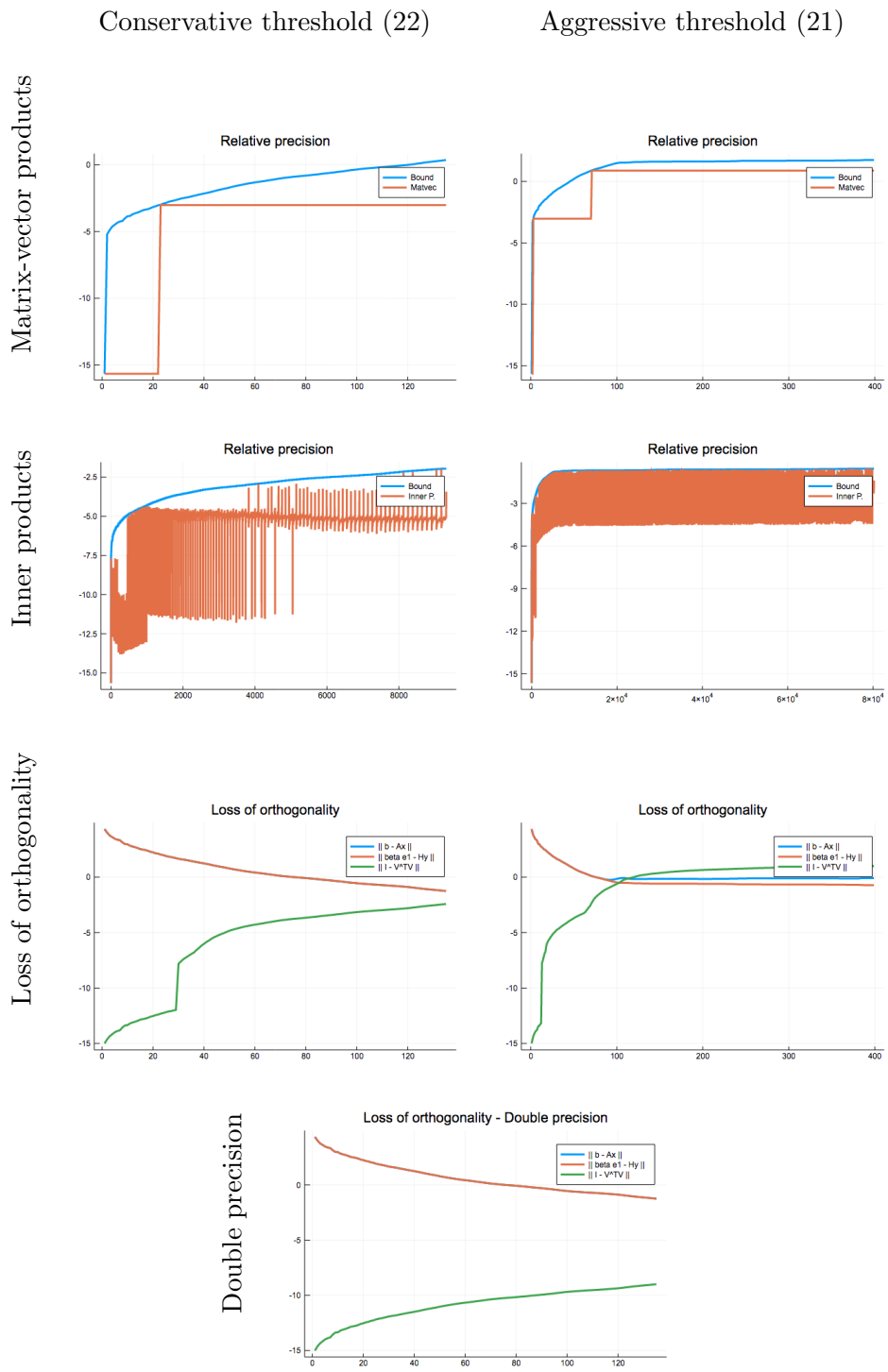

Figure 3: GMRES with variable precision floating-point arithmetic: Experiments with the 494_bus matrix and $\epsilon = 10^{-6}\|A\|_2$. The bottom figure corresponds to GMRES in double precision.

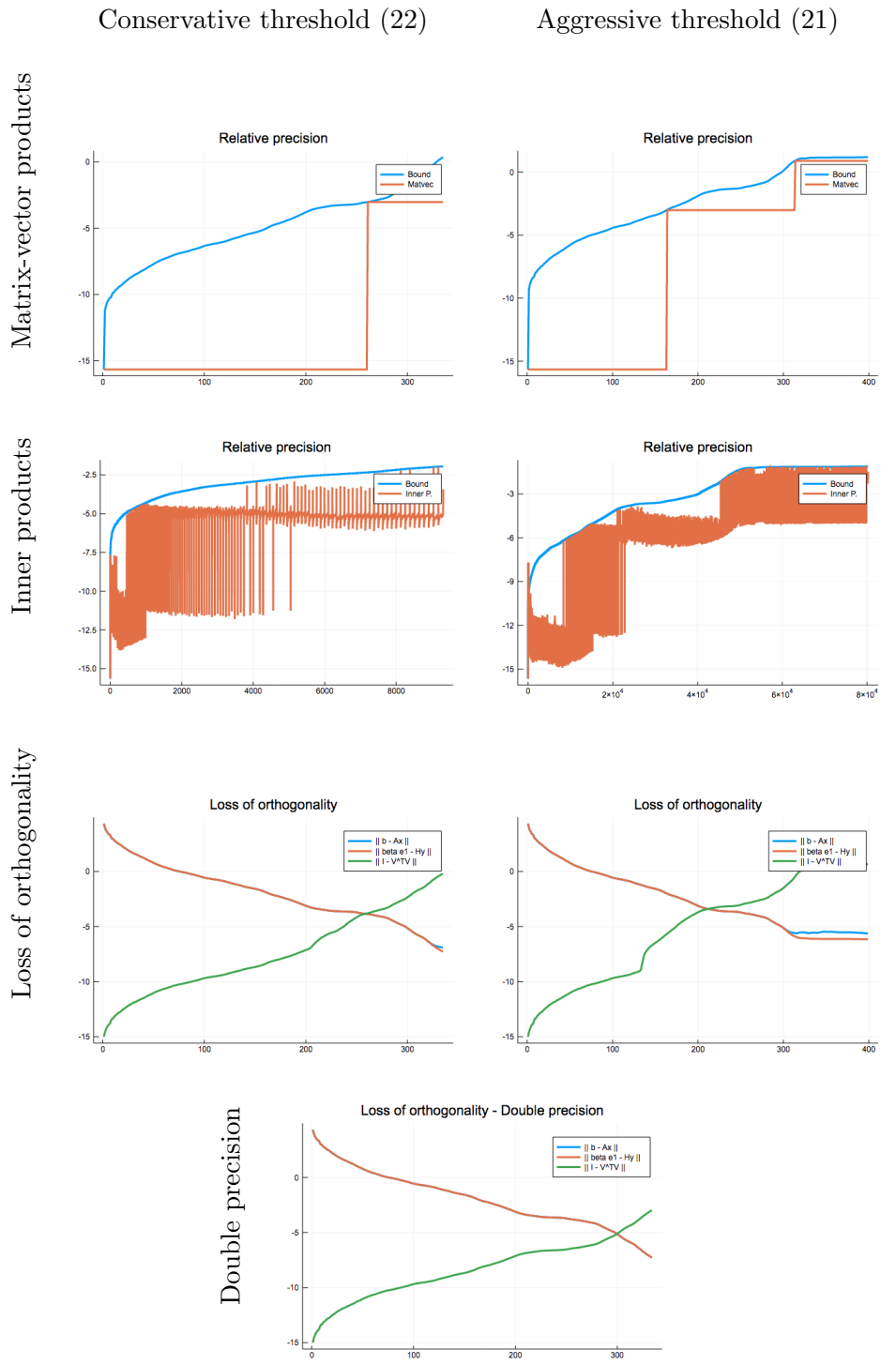

Figure 4: GMRES with variable precision floating-point arithmetic: Experiments with the 494_bus matrix and $\epsilon = 10^{-12}\|A\|_2$. The bottom figure corresponds to GMRES in double precision.

A Appendix

A.1 Proof of (9)

In line 7 of Algorithm 1, in the ℓth pass of the inner loop at step j, we have

$$w_j^{(\ell)} = w_j^{(\ell-1)} - h_{\ell j} v_\ell \tag{30}$$

for $\ell = 1, \ldots, j$ and with $w_j^{(0)} = Av_j$. Writing this equation for $\ell = i+1$ to j, we have

$$\begin{aligned} w_j^{(i+1)} &= w_j^{(i)} - h_{i+1,j} v_{i+1}, \\ w_j^{(i+2)} &= w_j^{(i+1)} - h_{i+2,j} v_{i+2}, \\ &\vdots \\ w_j^{(j)} &= w_j^{(j-1)} - h_{j,j} v_j. \end{aligned}$$

Summing the above and cancelling identical terms that appear on the left and right hand sides gives

$$w_j^{(j)} = w_j^{(i)} - \sum_{\ell=i+1}^{j} h_{\ell j} v_\ell.$$

Because $w_j^{(j)} = v_{j+1} h_{j+1,j}$, this reduces to

$$w_j^{(i)} = \sum_{\ell=i+1}^{j+1} h_{\ell j} v_\ell. \tag{31}$$

Because the inner products h_{ij} are computed inexactly as in (5), from (30) we have

$$\begin{aligned} w_j^{(i)} &= w_j^{(i-1)} - h_{ij} v_i \\ &= w_j^{(i-1)} - (v_i^T w_j^{(i-1)} + \eta_{ij}) v_i \\ &= (I - v_i v_i^T) w_j^{(i-1)} - \eta_{ij} v_i. \end{aligned}$$

Therefore,

$$v_i^T w_j^{(i)} = -\eta_{ij}.$$

Multiplying (31) on the left by $-v_i^T$ gives

$$\eta_{ij} = -\sum_{\ell=i+1}^{j+1} h_{\ell j} (v_i^T v_\ell), \tag{32}$$

which is the entry in position (i,j) of the matrix equation

$$\begin{bmatrix} \eta_{11} & \cdots & \eta_{1k} \\ & \ddots & \vdots \\ & & \eta_{kk} \end{bmatrix} = - \begin{bmatrix} v_1^T v_2 & \cdots & v_1^T v_{k+1} \\ & \ddots & \vdots \\ & & v_k^T v_{k+1} \end{bmatrix} \begin{bmatrix} h_{21} & \cdots & h_{2k} \\ & \ddots & \vdots \\ & & h_{k+1,k} \end{bmatrix},$$

i.e., (9).

A.2 Proof of Lemma 1

Equation (10) follows from

$$e_j^T H_k^\dagger \underbrace{\begin{bmatrix} H_{j-1} & 0 \\ 0 & 0 \end{bmatrix}}_{\in \mathbb{R}^{(k+1)\times k}} \begin{bmatrix} y_{j-1} \\ 0_{k-j+1} \end{bmatrix} = e_j^T H_k^\dagger H_k \begin{bmatrix} y_{j-1} \\ 0 \end{bmatrix} = e_j^T \begin{bmatrix} y_{j-1} \\ 0 \end{bmatrix} = 0,$$

and thus

$$\begin{aligned} |e_j^T y_k| = |e_j^T H_k^\dagger \beta_1 e_1| &= \left| e_j^T H_k^\dagger \left(\beta_1 e_1 - \begin{bmatrix} H_{j-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_{j-1} \\ 0 \end{bmatrix} \right) \right| \\ &= \left| e_j^T H_k^\dagger \begin{bmatrix} \beta_1 e_1 - H_{j-1} y_{j-1} \\ 0 \end{bmatrix} \right| \leq \|H_k^\dagger\|_2 \|t_{j-1}\|_2. \end{aligned}$$

As for (12), for any $\gamma > 0$, the smallest singular value of the matrix $\left[\beta\gamma e_1, H_k D_k^{-1}\right]$ is the scaled total least squares (STLS) distance [24] for the estimation problem $H_k D_k^{-1} z \approx \beta e_1$. As shown in [23], it can be bounded by the least squares distance

$$\min_z \|\beta e_1 - H_k D_k^{-1} z\|_2 = \|\beta e_1 - H_k D_k^{-1} z_k\|_2 = \|\beta e_1 - H_k y_k\|_2 = \|t_k\|_2,$$

where $z_k = D_k y_k$. From [23, Theorem 4.1], we have

$$\frac{\|t_k\|_2}{\left(\gamma^{-2} + \|D_k y_k\|_2^2/(1-\tau_k^2)\right)^{1/2}} \leq \sigma_{\min}\left(\left[\beta\gamma e_1, H_k D_k^{-1}\right]\right) \leq \gamma \|t_k\|_2, \tag{33}$$

provided $\tau_k < 1$, where

$$\tau_k \equiv \frac{\sigma_{\min}\left(\left[\beta\gamma e_1, H_k D_k^{-1}\right]\right)}{\sigma_{\min}\left(H_k D_k^{-1}\right)}.$$

We now show that if $\gamma = (\epsilon\|b\|_2)^{-1}$ and D_k satisfies (11), then $\tau_k \leq 1/\sqrt{2}$. From the upper bound in (33) we immediately have

$$\sigma_{\min}\left(\left[\beta\gamma e_1, H_k D_k^{-1}\right]\right) \leq \gamma \|t_k\|_2 = \frac{\|t_k\|_2}{\epsilon\|b\|_2}.$$

Also,

$$\sigma_{\min}\left(H_k D_k^{-1}\right) = \min_{z\neq 0} \frac{\|H_k D_k^{-1} z\|_2}{\|z\|_2} = \min_{z\neq 0} \frac{\|H_k z\|_2}{\|D_k z\|_2} \geq \min_{z\neq 0} \frac{\|H_k z\|_2}{\|D_k\|_2 \|z\|_2} = \frac{\sigma_{\min}(H_k)}{\|D_k\|_2}.$$

Therefore, if (11) holds,

$$\tau_k \leq \frac{\|t_k\|_2}{\epsilon\|b\|_2} \frac{\|D_k\|_2}{\sigma_{\min}(H_k)} \leq \frac{1}{\sqrt{2}}.$$

Substituting $\gamma = (\epsilon\|b\|_2)^{-1}$ and $\tau_k \leq 1/\sqrt{2}$ into (33) gives (12).

A.3 Proof of Lemma 2

Note from (13) that the singular values of Q satisfy

$$\big(\sigma_i(Q)\big)^2 = \sigma_i(Q^TQ) = \sigma_i(I_k - F), \qquad i = 1, \ldots, k.$$

Therefore,

$$\sqrt{1 - \|F\|_2} \le \sigma_i(Q) \le \sqrt{1 + \|F\|_2}, \qquad i = 1, \ldots, k. \tag{34}$$

Equation (14) is equivalent to the linear matrix equation

$$Q^TMQ = I_k - Q^TQ.$$

It is straightforward to verify that one matrix M satisfying this equation is

$$\begin{aligned} M &= (Q^\dagger)^T(I_k - Q^TQ)Q^\dagger \\ &= Q(Q^TQ)^{-1}(I_k - Q^TQ)(Q^TQ)^{-1}Q^T. \end{aligned}$$

Notice that the above matrix M is symmetric. It can also be verified using the singular value decomposition of Q that the eigenvalues and singular values of I_n+M are

$$\lambda_i(I_n + M) = \sigma_i(I_n + M) = \begin{cases} \big(\sigma_i(Q)\big)^{-2}, & i = 1, \ldots, k, \\ 1, & i = k+1, \ldots, n, \end{cases}$$

which implies that the matrix $I_n + M$ is positive definite. From the above and (34), provided $\|F\|_2 \le \delta < 1$,

$$\frac{1}{1+\delta} \le \frac{1}{\big(\sigma_{\max}(Q)\big)^2} \le \sigma_i(I_n + M) \le \frac{1}{\big(\sigma_{\min}(Q)\big)^2} \le \frac{1}{1-\delta}, \tag{35}$$

from which (15) follows.

Acknowledgments

The authors would like to thank two anonymous referees whose comments lead to significant improvements in the presentation.

References

[1] A. Björck, *Solving linear least squares problems by Gram-Schmidt orthogonalization*, BIT Numerical Mathematics, 7 (1967), pp. 1–21.

[2] A. Björck and C. Paige, *Loss and recapture of orthogonality in the Modified Gram-Schmidt algorithm*, SIAM Journal on Matrix Analysis and Applications, 13 (1992), pp. 176–190.

[3] A. Bouras and V. Frayssé, *Inexact matrix-vector products in Krylov methods for solving linear systems: A relaxation strategy*, SIAM Journal on Matrix Analysis and Applications, 26 (2006), pp. 660–678.

[4] E. Carson and N.J.Higham, *Accelerating the solution of linear systems by iterative refinement in three precisions*, SIAM Journal on Scientific Computing, 40 (2018), pp. 817–847.

[5] T. A. Davis and Y. Hu, *The University of Florida sparse matrix collection*, ACM Transactions on Mathematical Software, 38 (2011), pp. 1–25.

[6] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical stability of GMRES*, BIT Numerical Mathematics, 35 (1995), pp. 309–330.

[7] J. V. D. Eshof and G. Sleijpen, *Inexact Krylov subspace methods for linear systems*, SIAM Journal on Matrix Analysis and Applications, 26 (2004), pp. 125–153.

[8] R. Freund, *Quasi-kernel polynomials and convergence results for quasi-minimal residual iterations*, Numerical Methods in Approximation Theory, 9 (1992), pp. 77–95.

[9] R. Freund, *Quasi-kernel polynomials and their use in non-Hermitian matrix iterations*, Journal of Computational and Applied Mathematics, 43 (1992), pp. 135–158.

[10] R. Freund and N. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numerische Mathematik, 60 (1991), pp. 315–339.

[11] L. Giraud, S. Gratton, and J. Langou, *Convergence in backward error of relaxed GMRES*, SIAM Journal on Scientific Computing, 29 (2007), pp. 710–728.

[12] S. Gratton, E. Simon, and P. Toint, *Minimizing convex quadratic with variable precision Krylov methods*, arXiv, abs/1807.07476 (2018).

[13] A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical behaviour of the Modified Gram-Schmidt GMRES implementation*, BIT Numerical Mathematics, 37 (1997), pp. 706–719.

[14] A. Haidar, A. Abdelfattah, M. Zounon, P. Wu, S. Pradesh, S. Tomov, and J. Dongarra, *The design of fast and energy-efficient linear solvers: on the potential of half-precision arithmetic and iterative refinement techniques*, in Computational Science—ICCS 2018, Yong Shi, Haohuan Fu, Yingjie Tian, Valeria V. Krzhizhanovskaya, Michael Harold Lees, Jack Dongarra, and Peter M. A. Sloot editors, Springer, 2018, pp. 586–600.

[15] A. Haidar, S. Tomov, J. Dongarra, and N. J. Higham, *Harnessing GPU tensor cores for fast FP16 arithmetic to speed up mixed-precision iterative refinement solvers*, in Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, IEEE Press, 2018, pp. 47:1–47:11.

[16] A. Haidar, P. Wu, S. Tomov, and J. Dongarra, *Investigating half precision arithmetic to accelerate dense linear system solvers*, in Proceedings of the 8th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems, ScalA 17, 2017, pp. 10:1–10:8.

[17] N. J. Higham, *A multiprecision world*, SIAM News, 50 (2017).

[18] N. J. Higham and S. Pradesh, *Simulating low precision floating-point arithmetic*, SIAM Journal on Scientific Computing, 41 (2019), pp. 585–602.

[19] N. J. Higham and S. Pradesh, *Squeezing a matrix into half precision, with an application to solving linear systems*, SIAM Journal on Scientific Computing, 41 (2019), pp. 2536–2551.

[20] S. J. Leon, A. Björck, and W. Gander, *Gram-Schmidt orthogonalization: 100 years and more*, Numerical Linear Algebra with Applications, 20 (2013), pp. 492–532.

[21] C. Paige, *A useful form of unitary matrix obtained from any sequence of unit* 2*-norm* n*-vectors*, SIAM Journal on Matrix Analysis and Applications, 31 (2009), pp. 565–583.

[22] C. Paige, M. Rozložník, and Z. Strakoš, *Modified Gram-Schmidt (MGS), least squares, and backward stability of MGS-GMRES*, SIAM Journal on Matrix Analysis and Applications, 28 (2006), pp. 264–284.

[23] C. Paige and Z. Strakoš, *Bounds for the least squares distance using scaled total least squares*, Numerische Mathematik, 91 (2002), pp. 93–115.

[24] C. Paige and Z. Strakoš, *Scaled total least squares fundamentals*, Numerische Mathematik, 91 (2002), pp. 117–146.

[25] C. Paige and W. Wülling, *Properties of a unitary matrix obtained from a sequence of normalized vectors*, SIAM Journal on Matrix Analysis and Applications, 35 (2014), pp. 526–545.

[26] J. Pestana and A. J. Wathen, *On the choice of preconditioner for minimum residual methods for non-Hermitian matrices*, Journal of Computational and Applied Mathematics, 249 (2013), pp. 57–68.

[27] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on Scientific and Statistical Computing, 7 (1986), pp. 856–869.

[28] V. Simoncini and D. Szyld, *Theory of inexact Krylov subspace methods and applications to scientific computing*, SIAM Journal on Scientific Computing, 25 (2003), pp. 454–477.

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