Reorthogonalized Pythagorean variants of block classical Gram-Schmidt

Erin Carson* Kathryn Lund†,‡ Yuxin Ma* Eda Oktay*,§

*Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University, Sokolovská 49/83, 186 75 Praha 8, Czechia
Email: {carson, oktay}@karlin.mff.cuni.cz, Email: yuxin.ma@matfyz.cuni.cz,
ORCID: 0000-0001-9469-7467, ORCID: 0000-0003-0761-2184, ORCID: 0000-0002-2860-0134
†Computational Methods in Systems and Control Theory, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany
‡Computational Mathematics Theme, Building R71, STFC Rutherford Appleton Laboratory, Harwell Oxford, Didcot, Oxfordshire, OX11 0QX, United Kingdom
Email: kathryn.lund@stfc.ac.uk, ORCID: 0000-0001-9851-6061
§Department of Mathematics, Chemnitz University of Technology, Reichenhainer Str. 41, 09126 Chemnitz, Germany

Abstract: Block classical Gram-Schmidt (BCGS) is commonly used for orthogonalizing a set of vectors X in distributed computing environments due to its favorable communication properties relative to other orthogonalization approaches, such as modified Gram-Schmidt or Householder. However, it is known that BCGS (as well as recently developed low-synchronization variants of BCGS) can suffer from a significant loss of orthogonality in finite-precision arithmetic, which can contribute to instability and inaccurate solutions in downstream applications such as s-step Krylov subspace methods. A common solution to improve the orthogonality among the vectors is reorthogonalization. Focusing on the "Pythagorean" variant of BCGS, introduced in [E. Carson, K. Lund, & M. Rozložník. *SIAM J. Matrix Anal. Appl.* 42(3), pp. 1365–1380, 2021], which guarantees an $O(\varepsilon)\kappa^2(X)$ bound on the loss of orthogonality as long as $O(\varepsilon)\kappa^2(X) < 1$, where ε denotes the unit roundoff, we introduce and analyze two reorthogonalized Pythagorean BCGS variants. These variants feature favorable communication properties, with asymptotically two synchronization points per block column, as well as an improved $O(\varepsilon)$ bound on the loss of orthogonality. Our bounds are derived in a general fashion to additionally allow for the analysis of mixed-precision variants. We verify our theoretical results with a panel of test matrices and experiments from a new version of the `BlockStab` toolbox.

Keywords: Gram-Schmidt algorithm, low-synchronization, communication-avoiding, mixed precision, multiprecision, loss of orthogonality, stability

Mathematics subject classification: 65-04, 65F25, 65G50, 65Y20

Novelty statement: Bounds on the loss of orthogonality are proven for two variants of reorthogonalized block classical Gram-Schmidt, including new variants with asymptotically two synchronization points per block vector. These bounds are important for the design of scalable, iterative solvers in high-performance computing. We also examine mixed-precision variants of these methods.

1. Introduction

Interest in low-synchronization variants of the Gram-Schmidt method has been proliferating recently [5, 8, 9, 18, 22, 25, 29–31]. These methods are part of the more general trend of developing communication-reducing orthogonalization routines with the goal of reducing memory movement between levels of the memory hierarchy or nodes on a network, thereby improving scalability in high-performance, and especially exascale, computing [1,11,17]. In this manuscript, we concentrate on block Gram-Schmidt (BGS) methods, and in particular, on reorthogonalized versions of block classical Gram-Schmidt with Pythagorean inner product (`BCGS-PIP`) from [8], and how they can achieve loss of orthogonality on the order of unit roundoff with only two synchronization points per block of columns. In related work [7], we consider a generalization of BCGS2-type algorithms [3,4], which have four such synchronization points per block, and we use this generalization to study the stability of a reorthogonalized BGS variant with only one synchronization point per block.

We define a *block vector* $\boldsymbol{X} \in \mathbb{R}^{m\times s}$ with $m \gg s$ as a concatenation of s column vectors, i.e., a tall-skinny matrix. We are interested in computing an economic QR decomposition for the concatenation of p block vectors

$$\boldsymbol{\mathcal{X}} = \begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2 & \cdots \boldsymbol{X}_p\end{bmatrix} \in \mathbb{R}^{m\times ps}.$$

We achieve this via a BGS method that takes $\boldsymbol{\mathcal{X}}$ and a block size s as arguments and returns an orthonormal basis $\boldsymbol{\mathcal{Q}} \in \mathbb{R}^{m\times ps}$ along with an upper triangular $\mathcal{R} \in \mathbb{R}^{ps\times ps}$ such that $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{Q}}\mathcal{R}$. Both $\boldsymbol{\mathcal{Q}}$ and $\mathcal{R}$ are computed block-wise, meaning that s new columns of $\boldsymbol{\mathcal{Q}}$ are generated per iteration, as opposed to just one column at a time.

Blocking or batching data is a known technique for reducing the total number of synchronization points, or *sync points*. In a distributed setting, we define a sync point as an operation requiring all nodes to send and receive information to and from one other, such as an `MPI_Allreduce`. In an orthogonalization procedure like BGS, block inner products and *intraorthogonalization* routines (which orthogonalize vectors within a block column) like tall-skinny QR require sync points when block vectors are distributed row-wise across nodes. For the purposes of this manuscript, we will assume that a block inner product $\boldsymbol{X}^*\boldsymbol{Y}$ and an intraorthogonalization `IO` each constitute one sync point.

In addition to reducing sync points, we are also concerned with the stability of BGS, which we measure here in terms of the *loss of orthogonality (LOO)*,

$$\left\| I - \bar{\boldsymbol{\mathcal{Q}}}^T \bar{\boldsymbol{\mathcal{Q}}} \right\|, \tag{1}$$

where I is the $ps \times ps$ identity matrix and $\bar{\boldsymbol{\mathcal{Q}}} \in \mathbb{R}^{m\times ps}$ denotes the $\boldsymbol{\mathcal{Q}}$ factor computed in floating-point arithmetic. We take $\|\cdot\|$ as the induced matrix 2-norm. As is common in the relevant literature, we regard a rectangular matrix $\boldsymbol{\mathcal{Q}} \in \mathbb{R}^{m\times ps}$ as *orthogonal* when $I - \boldsymbol{\mathcal{Q}}^T\boldsymbol{\mathcal{Q}} = 0$, meaning not only are the columns of $\boldsymbol{\mathcal{Q}}$ orthogonal to one another but also each column has norm one. We regard the term *orthonormal* as synonymous with *orthogonal*.

Orthogonality is important for a number of downstream purposes, especially eigenvalue approximation; see, e.g., [15,26] and sources therein. Furthermore, having LOO close to working precision simplifies the backward error analysis of Krylov subspace methods like GMRES; see, e.g., the modular backward stability framework by Buttari et al. [6]. Orthogonality is also stronger than being well-conditioned, i.e., having $\kappa(\bar{\boldsymbol{\mathcal{Q}}}) \approx 1$, where κ denotes the 2-condition number of a matrix, i.e., the ratio between its largest and smallest singular values. Indeed, Gram-Schmidt methods frequently lose orthogonality while producing well-conditioned bases. For all these reasons, we concentrate on LOO as the primary metric of a method's stability.

We will also consider the standard residual

$$\|\boldsymbol{\mathcal{Q}}\mathcal{R} - \boldsymbol{\mathcal{X}}\|, \tag{2}$$

as well as the Cholesky residual,

$$\left\| \boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}} - \bar{\mathcal{R}}^T\bar{\mathcal{R}} \right\|, \tag{3}$$

where $\bar{\mathcal{R}}$ is the finite precision counterpart of $\mathcal{R}$. This latter residual measures how close a BGS method is to correctly computing a Cholesky decomposition of $\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}}$, which can provide insight into the stability pitfalls of a method; see, e.g., [8, 14, 23].

In the following section, we summarize BCGS-PIP and results from [8] and prove stability bounds on two reorthogonalized variants, BCGS-PIP+ and BCGS-PIPI+, which have $2p$ and $2p-1$ sync points, respectively, or roughly 2 sync points per block vector. Section 3 deals with mixed-precision variants of the new reorthogonalized methods, and Section 4 demonstrates the numerical behavior of all methods using the `BlockStab` toolbox. We summarize conclusions and future perspectives in Section 5.

A few remarks regarding notation are necessary before proceeding. Generally, uppercase Roman letters (R_{ij}, S_{ij}, T_{ij}) denote $s \times s$ block entries of a $ps \times ps$ matrix, which itself is usually denoted by uppercase Roman script ($\mathcal{R}, \mathcal{S}, \mathcal{T}$). A block column of such matrices is denoted with MATLAB indexing:

$$\mathcal{R}_{1:k-1,k} = \begin{bmatrix} R_{1,k} \\ R_{2,k} \\ \vdots \\ R_{k-1,k} \end{bmatrix}.$$

For simplicity, we also abbreviate standard $ks \times ks$ submatrices as $\mathcal{R}_k := \mathcal{R}_{1:k,1:k}$.

Bold uppercase Roman letters ($\boldsymbol{Q}_k$, $\boldsymbol{X}_k$, $\boldsymbol{U}_k$) denote $m \times s$ block vectors, and bold, uppercase Roman script ($\boldsymbol{\mathcal{Q}}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{U}}$) denotes an indexed concatenation of p such vectors. Standard $m \times ks$ submatrices are abbreviated as

$$\boldsymbol{\mathcal{Q}}_k := \boldsymbol{\mathcal{Q}}_{1:k} = \begin{bmatrix} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 & \cdots & \boldsymbol{Q}_k \end{bmatrix}.$$

Note that when $\boldsymbol{X}$ is used on its own throughout the text, it denotes a generic block vector. When a subscript is added to $\boldsymbol{X}$ as $\boldsymbol{X}_k$, we are referring to the specific kth block vector of the input matrix $\boldsymbol{\mathcal{X}}$. Furthermore, $\boldsymbol{\mathcal{X}}_k$ is the concatenation of the first k block vectors of $\boldsymbol{\mathcal{X}}$, as defined above.

The function $[\boldsymbol{Q}, R] = \texttt{IO}(\boldsymbol{X})$ denotes an *intraorthogonalization* routine, i.e., a method used to orthogonalize vectors within a generic block vector $\boldsymbol{X}$. This can be any number of methods, including Householder, classical Gram-Schmidt, modified Gram-Schmidt, or Cholesky QR.

We use ε to denote the *unit roundoff* of a chosen *working precision*. For example, if we use IEEE double precision arithmetic, then $\varepsilon = 2^{-53} \approx 1.1 \cdot 10^{-16}$, and for IEEE single precision, $\varepsilon = 2^{-24} \approx 6.0 \cdot 10^{-8}$. Throughout the text, we use standard results from [16] (particularly Sections 2.2 and 3.5) for rounding-error analysis.

We also make use of big-O notation like $\mathcal{O}(\varepsilon)$, which is essentially $c(m, ps)\varepsilon$, with $c(m, ps)$ denoting a low-degree polynomial in dimensional constants m and ps. To enhance clarity, we employ $\mathcal{O}(\varepsilon)$ to disregard the dimensional factor $c(m, ps)$, although this factor can become significant when m and ps are large.

2. Stability of reorthogonalized variants of BCGS-PIP

BCGS-PIP (Algorithm 1) is a corrected version of Block Classical Gram-Schmidt (BCGS), where the block diagonal entries of the $\mathcal{R}$ factor are computed via the block Pythagorean theorem [8]. This correction stabilizes the algorithm by keeping the relative Cholesky residual (3) close to the working precision. However, the overall LOO for BCGS-PIP can be quite high, and the bound

$$\left\| I - \bar{\boldsymbol{\mathcal{Q}}}^T \bar{\boldsymbol{\mathcal{Q}}} \right\| \leq \mathcal{O}(\varepsilon)\, \kappa^2(\boldsymbol{\mathcal{X}})$$

only holds when $\mathcal{O}(\varepsilon)\, \kappa^2(\boldsymbol{\mathcal{X}}) \leq 1$.

Thus as $\kappa(\boldsymbol{\mathcal{X}}) \to \varepsilon^{-1/2}$, any guarantees on the orthogonality of $\bar{\boldsymbol{\mathcal{Q}}}$ are lost.

A standard strategy for improving $\bar{\boldsymbol{\mathcal{Q}}}$ is running the Gram-Schmidt procedure twice; see, e.g., [4, 24] for analysis of reorthogonalized block variants of BCGS. We do not consider BCGS further in this manuscript; for a detailed analysis and provable bounds on its LOO, see [7], which also analyzes low-sync reorthogonalized versions of BCGS. In the following two subsections, we propose two reorthogonalized versions of BCGS-PIP and prove bounds for their LOO and residuals.

Algorithm 1 $[\mathcal{Q}, \mathcal{R}] = \texttt{BCGS-PIP}(\mathcal{X}, \texttt{IO})$

1: $[\boldsymbol{Q}_1, R_{11}] = \texttt{IO}(\boldsymbol{X}_1)$
2: **for** $k = 2, \ldots, p$ **do**
3: $\quad \begin{bmatrix} \mathcal{R}_{1:k-1,k} \\ P_k \end{bmatrix} = \begin{bmatrix} \mathcal{Q}_{k-1} & \boldsymbol{X}_k \end{bmatrix}^T \boldsymbol{X}_k$
4: $\quad R_{kk} = \texttt{chol}\left(P_k - \mathcal{R}_{1:k-1,k}^T \mathcal{R}_{1:k-1,k}\right)$
5: $\quad \boldsymbol{V}_k = \boldsymbol{X}_k - \mathcal{Q}_{k-1} \mathcal{R}_{1:k-1,k}$
6: $\quad \boldsymbol{Q}_k = \boldsymbol{V}_k R_{kk}^{-1}$
7: **end for**
8: **return** $\mathcal{Q} = [\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_p]$, $\mathcal{R} = (R_{ij})$

2.1. BCGS-PIP+

We first consider the simple approach of running BCGS-PIP twice in a row. See Algorithm 2 for what we call BCGS-PIP+, where + stands for "reorthogonalization." Note that despite how the pseudocode is written, it is not necessary to store two bases in practice: $\mathcal{U}$ can be stored in place of $\mathcal{X}$ throughout the first step. Similarly, it is possible to build $\mathcal{R}$ by replacing $\mathcal{S}$ gradually in the second step, and there is no need to construct $\mathcal{T}$ explicitly; cf. the second phase of Algorithm 3. The pseudocode is written in such a way as to simplify the mathematical analysis.

Algorithm 2 $[\mathcal{Q}, \mathcal{R}] = \texttt{BCGS-PIP+}(\mathcal{X}, \texttt{IO})$

1: $[\mathcal{U}, \mathcal{S}] = \texttt{BCGS-PIP}(\mathcal{X}, \texttt{IO})$
2: $[\mathcal{Q}, \mathcal{T}] = \texttt{BCGS-PIP}(\mathcal{U}, \texttt{IO})$
3: $\mathcal{R} = \mathcal{T}\mathcal{S}$;
4: **return** $\mathcal{Q} = [\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_p]$, $\mathcal{R} = (R_{ij})$

Under minimal assumptions, proving an $\mathcal{O}(\varepsilon)$ LOO bound for Algorithm 2 follows directly from the results of [8]. In particular, we can obtain the following result by applying [8, Theorem 3.4] twice. Note the condition $\kappa(\boldsymbol{X}) \leq \kappa(\mathcal{X})$ required for $\texttt{IO}(\boldsymbol{X})$: when $\boldsymbol{X}_k$ is a block vector of $\mathcal{X}$, this condition holds by [15, Corollary 8.6.3]. Here and elsewhere, it is meant to ensure that IO does not handle block vectors with worse conditioning than that of $\mathcal{X}$.

Corollary 1. *Let $\mathcal{X} \in \mathbb{R}^{m \times ps}$ with $\mathcal{O}(\varepsilon)\kappa^2(\mathcal{X}) \leq \frac{1}{2}$ and $\bar{\mathcal{Q}}$ and $\bar{\mathcal{R}}$ be computed by Algorithm 2. Assuming that for all $\boldsymbol{X} \in \mathbb{R}^{m \times s}$ with $\kappa(\boldsymbol{X}) \leq \kappa(\mathcal{X})$, $[\bar{\boldsymbol{Q}}, \bar{R}] = \texttt{IO}(\boldsymbol{X})$ satisfy*

$$\begin{aligned} \bar{R}^T \bar{R} &= \boldsymbol{X}^T \boldsymbol{X} + \Delta E, \quad \|\Delta E\| \leq \mathcal{O}(\varepsilon) \|\boldsymbol{X}\|^2, \text{ and} \\ \bar{\boldsymbol{Q}} \bar{R} &= \boldsymbol{X} + \Delta \boldsymbol{D}, \quad \|\Delta \boldsymbol{D}\| \leq \mathcal{O}(\varepsilon) \left(\|\boldsymbol{X}\| + \left\|\bar{\boldsymbol{Q}}\right\| \left\|\bar{R}\right\|\right), \end{aligned}$$

then $\bar{\mathcal{Q}}$ and $\bar{\mathcal{R}}$ satisfy

$$\left\| I - \bar{\mathcal{Q}}^T \bar{\mathcal{Q}} \right\| \leq \mathcal{O}(\varepsilon), \text{ and} \tag{4}$$

$$\bar{\mathcal{Q}} \bar{\mathcal{R}} = \mathcal{X} + \Delta \mathcal{D}, \quad \|\Delta \mathcal{D}\| \leq \mathcal{O}(\varepsilon) \|\mathcal{X}\|. \tag{5}$$

We can obtain a further result on the Cholesky residual of BCGS-PIP+ via the following corollary, which follows by applying [8, Theorem 3.2] twice.

Corollary 2. *Let $\mathcal{X} \in \mathbb{R}^{m \times ps}$ with $\mathcal{O}(\varepsilon)\kappa^2(\mathcal{X}) \leq \frac{1}{2}$ and $\bar{\mathcal{Q}}$ and $\bar{\mathcal{R}}$ be computed by Algorithm 2. Assuming that for all $\boldsymbol{X} \in \mathbb{R}^{m \times s}$ with $\kappa(\boldsymbol{X}) \leq \kappa(\mathcal{X})$, $[\bar{\boldsymbol{Q}}, \bar{R}] = \texttt{IO}(\boldsymbol{X})$ satisfy*

$$\begin{aligned} \bar{R}^T \bar{R} &= \boldsymbol{X}^T \boldsymbol{X} + \Delta E, \quad \|\Delta E\| \leq \mathcal{O}(\varepsilon) \|\boldsymbol{X}\|^2, \text{ and} \\ \bar{\boldsymbol{Q}} \bar{R} &= \boldsymbol{X} + \Delta \boldsymbol{D}, \quad \|\Delta \boldsymbol{D}\| \leq \mathcal{O}(\varepsilon) \left(\|\boldsymbol{X}\| + \left\|\bar{\boldsymbol{Q}}\right\| \left\|\bar{R}\right\|\right), \end{aligned}$$

then $\bar{\mathcal{R}}$ satisfies

$$\bar{\mathcal{R}}^T \bar{\mathcal{R}} = \mathcal{X}^T \mathcal{X} + \Delta \mathcal{E}, \quad \|\Delta \mathcal{E}\| \leq \mathcal{O}(\varepsilon) \|\mathcal{X}\|^2. \tag{6}$$

For our mixed-precision analysis, it will also be useful to have a generalized formulation of these bounds that do not rely on a specific precision and that also reveal all the sources of error. From standard rounding-error principles [16], we can write the following bounds for each step of Algorithm 2 for constants $\delta_{US}, \omega_U, \delta_{QT}, \omega_Q, \delta_{TS} \in (0,1)$:

$$\bar{\boldsymbol{\mathcal{U}}}\bar{\mathcal{S}} = \boldsymbol{\mathcal{X}} + \Delta\boldsymbol{\mathcal{D}}_{US}, \quad \|\Delta\boldsymbol{\mathcal{D}}_{US}\| \le \delta_{US}\|\boldsymbol{\mathcal{X}}\|; \tag{7}$$

$$\left\|I - \bar{\boldsymbol{\mathcal{U}}}^T\bar{\boldsymbol{\mathcal{U}}}\right\| \le \omega_U\kappa^2(\boldsymbol{\mathcal{X}}); \tag{8}$$

$$\bar{\boldsymbol{\mathcal{Q}}}\bar{\mathcal{T}} = \bar{\boldsymbol{\mathcal{U}}} + \Delta\boldsymbol{\mathcal{D}}_{QT}, \quad \|\Delta\boldsymbol{\mathcal{D}}_{QT}\| \le \delta_{QT}\left\|\bar{\boldsymbol{\mathcal{U}}}\right\|; \tag{9}$$

$$\left\|I - \bar{\boldsymbol{\mathcal{Q}}}^T\bar{\boldsymbol{\mathcal{Q}}}\right\| \le \omega_Q\kappa^2(\bar{\boldsymbol{\mathcal{U}}}); \text{ and} \tag{10}$$

$$\bar{\mathcal{R}} = \bar{\mathcal{T}}\bar{\mathcal{S}} + \Delta\mathcal{D}_{TS}, \quad \|\Delta\mathcal{D}_{TS}\| \le \delta_{TS}\left\|\bar{\mathcal{T}}\right\|\left\|\bar{\mathcal{S}}\right\|. \tag{11}$$

The following theorem summarizes how each step of Algorithm 2 influences the LOO and residual per iteration, which will be useful in Section 3 when we consider a two-precision variant. We omit dependencies on k from the δ constants, meaning each can be regarded as a maximum over all constants stemming from the same step of the algorithm. We also generally assume that such constants are much smaller than 1 and drop "quadratic" terms resulting from their products.

Theorem 1. *Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m\times ps}$ and suppose $[\bar{\boldsymbol{\mathcal{Q}}}, \bar{\mathcal{R}}] =$ BCGS-PIP+$(\boldsymbol{\mathcal{X}}, IO)$. Assuming that (7)–(11) are satisfied with $\omega_U\kappa^2(\boldsymbol{\mathcal{X}}) \le \frac{1}{2}$ and $\omega_Q \le \frac{1}{6}$, then $\bar{\boldsymbol{\mathcal{Q}}}$ and $\bar{\mathcal{R}}$ satisfy*

$$\left\|I - \bar{\boldsymbol{\mathcal{Q}}}^T\bar{\boldsymbol{\mathcal{Q}}}\right\| \le 3\cdot\omega_Q, \text{ and} \tag{12}$$

$$\bar{\boldsymbol{\mathcal{Q}}}\bar{\mathcal{R}} = \boldsymbol{\mathcal{X}} + \Delta\boldsymbol{\mathcal{D}}, \quad \|\Delta\boldsymbol{\mathcal{D}}\| \le \left(\delta_{US} + \sqrt{6}\cdot\delta_{QT} + 9\cdot\delta_{TS}\right)\|\boldsymbol{\mathcal{X}}\|. \tag{13}$$

Proof. From (8) and the assumption $\omega_U\kappa^2(\boldsymbol{\mathcal{X}}) \le \frac{1}{2}$, we obtain

$$\|\bar{\boldsymbol{\mathcal{U}}}\|^2 = \left\|\bar{\boldsymbol{\mathcal{U}}}^T\bar{\boldsymbol{\mathcal{U}}}\right\| \le \left\|I - \bar{\boldsymbol{\mathcal{U}}}^T\bar{\boldsymbol{\mathcal{U}}}\right\| + \|I\| \le \frac{3}{2}, \tag{14}$$

and by the perturbation theory of singular values [15, Corollary 8.6.2], we obtain a lower bound on $\sigma^2_{\min}(\bar{\boldsymbol{\mathcal{U}}})$, namely, $\sigma^2_{\min}(\bar{\boldsymbol{\mathcal{U}}}) \ge \frac{1}{2}$. Consequently, we have the following for $\kappa^2(\bar{\boldsymbol{\mathcal{U}}})$:

$$\kappa^2(\bar{\boldsymbol{\mathcal{U}}}) = \frac{\sigma^2_{\max}(\boldsymbol{\mathcal{U}})}{\sigma^2_{\min}(\boldsymbol{\mathcal{U}})} \le 3. \tag{15}$$

Applying (15) to (10) immediately confirms (12).

To prove (13), we combine (7), (9), and (11) to arrive at

$$\bar{\boldsymbol{\mathcal{Q}}}\bar{\mathcal{R}} = \boldsymbol{\mathcal{X}} + \underbrace{\Delta\boldsymbol{\mathcal{D}}_{US} + \Delta\boldsymbol{\mathcal{D}}_{QT}\bar{\mathcal{S}} + \bar{\boldsymbol{\mathcal{Q}}}\Delta\mathcal{D}_{TS}}_{=:\Delta\boldsymbol{\mathcal{D}}}. \tag{16}$$

Note that $\left\|\bar{\boldsymbol{\mathcal{Q}}}\right\| \le \frac{\sqrt{6}}{2}$ is derived similarly to (14), which together yield

$$\begin{aligned}\|\Delta\boldsymbol{\mathcal{D}}\| &\le \|\Delta\boldsymbol{\mathcal{D}}_{US}\| + \|\Delta\boldsymbol{\mathcal{D}}_{QT}\|\left\|\bar{\mathcal{S}}\right\| + \|\boldsymbol{\mathcal{Q}}\|\,\|\Delta\mathcal{D}_{TS}\| \\ &\le \delta_{US}\|\boldsymbol{\mathcal{X}}\| + \delta_{QT}\left\|\bar{\mathcal{S}}\right\| + \frac{\sqrt{6}}{2}\cdot\delta_{TS}\left\|\bar{\mathcal{T}}\right\|\left\|\bar{\mathcal{S}}\right\|.\end{aligned} \tag{17}$$

By (7) and (9), we obtain

$$\begin{aligned}\bar{\mathcal{S}}^T\bar{\mathcal{S}} &= \boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{X}}^T\Delta\boldsymbol{\mathcal{D}}_{US} + \Delta\boldsymbol{\mathcal{D}}_{US}\boldsymbol{\mathcal{X}} + \bar{\mathcal{S}}^T\left(I - \bar{\boldsymbol{\mathcal{U}}}^T\bar{\boldsymbol{\mathcal{U}}}\right)\bar{\mathcal{S}}, \\ \bar{\mathcal{T}}^T\bar{\mathcal{T}} &= \bar{\boldsymbol{\mathcal{U}}}^T\bar{\boldsymbol{\mathcal{U}}} + \bar{\boldsymbol{\mathcal{U}}}^T\Delta\boldsymbol{\mathcal{D}}_{QT} + \Delta\boldsymbol{\mathcal{D}}_{QT}^T\bar{\boldsymbol{\mathcal{U}}} + \bar{\mathcal{T}}^T\left(I - \bar{\boldsymbol{\mathcal{Q}}}^T\bar{\boldsymbol{\mathcal{Q}}}\right)\bar{\mathcal{T}}.\end{aligned}$$

Furthermore, from (14), (15), and the assumptions $\omega_U\kappa^2(\boldsymbol{\mathcal{X}}) \le \frac{1}{2}$ and $\omega_Q \le \frac{1}{6}$, it follows that

$$\left\|\bar{\mathcal{S}}\right\| \le \sqrt{2 + 4\cdot\delta_{US}}\,\|\boldsymbol{\mathcal{X}}\| \le \sqrt{6}\,\|\boldsymbol{\mathcal{X}}\|, \tag{18}$$

$$\left\|\bar{\mathcal{T}}\right\| \le \sqrt{3 + 6\cdot\delta_{QT}} \le 3. \tag{19}$$

Substituting (18) and (19) into (17) proves (13). □

2.2. BCGS-PIPI+

One drawback to Algorithm 2 is that two for-loops are required to reorthogonalize the entire basis. In practice, we can combine the for-loops without introducing additional sync points. The resulting algorithm is denoted BCGS-PIPI+, where I+ stands for "inner reorthogonalization", and is provided as Algorithm 3. Note that, as in Algorithm 2, we construct matrices $\mathcal{S}$, $\mathcal{T}$, and $\boldsymbol{\mathcal{U}}$ throughout the algorithm, although none of these needs to be explicitly formed in practice.

Combining the for-loops and introducing a general IO for the first orthogonalization step of Algorithm 3 creates new challenges in proving its stability bounds, primarily because the first block vector is no longer reorthogonalized. We can therefore no longer directly use the results from [8], as a stricter condition will have to be placed on the choice of IO. In the following, we will first develop a generalized approach depending on small constants stemming from particular sources of error, similar to Theorem 1. We then conclude the section with an application to the uniform-precision case by inducting over the generalized results.

Algorithm 3 $[\boldsymbol{\mathcal{Q}}, \mathcal{R}] = \texttt{BCGS-PIPI+}(\boldsymbol{\mathcal{X}}, \texttt{IO})$

$[\boldsymbol{Q}_1, R_{11}] = \texttt{IO}(\boldsymbol{X}_1)$ ▷ $S_{11} = R_{11}$, $\boldsymbol{U}_1 = \boldsymbol{Q}_1$, $T_{11} = I$
for $k = 2, \ldots, p$ **do**
  $\begin{bmatrix} \mathcal{S}_{1:k-1,k} \\ \Omega_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mathcal{Q}}_{k-1} & \boldsymbol{X}_k \end{bmatrix}^T \boldsymbol{X}_k$ ▷ First BCGS-PIP step
  $S_{kk} = \texttt{chol}\left(\Omega_k - \mathcal{S}_{1:k-1,k}^T \mathcal{S}_{1:k-1,k}\right)$
  $\boldsymbol{V}_k = \boldsymbol{X}_k - \boldsymbol{\mathcal{Q}}_{k-1} \mathcal{S}_{1:k-1,k}$
  $\boldsymbol{U}_k = \boldsymbol{V}_k S_{kk}^{-1}$
  $\begin{bmatrix} \mathcal{T}_{1:k-1,k} \\ P_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mathcal{Q}}_{k-1} & \boldsymbol{U}_k \end{bmatrix}^T \boldsymbol{U}_k$ ▷ Second BCGS-PIP step
  $T_{kk} = \texttt{chol}\left(P_k - \mathcal{T}_{1:k-1,k}^T \mathcal{T}_{1:k-1,k}\right)$
  $\boldsymbol{W}_k = \boldsymbol{U}_k - \boldsymbol{\mathcal{Q}}_{k-1} \mathcal{T}_{1:k-1,k}$
  $\boldsymbol{Q}_k = \boldsymbol{W}_k T_{kk}^{-1}$
  $\mathcal{R}_{1:k-1,k} = \mathcal{S}_{1:k-1,k} + \mathcal{T}_{1:k-1,k} S_{kk}$ ▷ Finalize $\mathcal{R}$ entries
  $R_{kk} = T_{kk} S_{kk}$
end for
return $\boldsymbol{\mathcal{Q}} = [\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_p]$, $\mathcal{R} = (R_{ij})$

For the following, we assume all quantities are computed by Algorithm 3 and that at each iteration $k \in \{2, \ldots, p\}$, there exists $\omega_{k-1} \in (0, 1)$ such that

$$\left\| I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \right\| \leq \omega_{k-1}. \tag{20}$$

Then similarly to (14),

$$\left\| \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \right\| \leq \sqrt{1 + \omega_{k-1}} \leq \sqrt{2}. \tag{21}$$

We can write the following for intermediate quantities computed by BCGS-PIPI+ at each iteration

$k \in \{2, \dots, p\}$, where we summarize rounding-error bounds via constants $\delta_* \in (0, 1)$:

$$\bar{\mathcal{S}}_{1:k-1,k} = \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \boldsymbol{X}_k + \Delta \boldsymbol{\mathcal{S}}_k, \quad \|\Delta \boldsymbol{\mathcal{S}}_k\| \le \delta_{Q^T X} \|\boldsymbol{X}_k\|; \tag{22}$$

$$\bar{\Omega}_k = \boldsymbol{X}_k^T \boldsymbol{X}_k + \Delta \Omega_k, \quad \|\Delta \Omega_k\| \le \delta_{X^T X} \|\boldsymbol{X}_k\|^2; \tag{23}$$

$$\bar{S}_{kk}^T \bar{S}_{kk} = \bar{\Omega}_k - \bar{\mathcal{S}}_{1:k-1,k}^T \bar{\mathcal{S}}_{1:k-1,k} + \Delta F_k^{(1)} + \Delta C_k^{(1)}, \tag{24}$$

$$\left\|\Delta F_k^{(1)}\right\| \le \delta_{S^T S} \|\boldsymbol{X}_k\|^2, \quad \left\|\Delta C_k^{(1)}\right\| \le \delta_{\texttt{chol}_1} \|\boldsymbol{X}_k\|^2; \tag{25}$$

$$\bar{\boldsymbol{V}}_k = \boldsymbol{X}_k - \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\mathcal{S}}_{1:k-1,k} + \Delta \boldsymbol{V}_k, \quad \|\Delta \boldsymbol{V}_k\| \le \delta_{QS} \|\boldsymbol{X}_k\|; \tag{26}$$

$$\bar{\boldsymbol{U}}_k \bar{S}_{kk} = \bar{\boldsymbol{V}}_k + \Delta \boldsymbol{G}_k^{(1)}, \quad \left\|\Delta \boldsymbol{G}_k^{(1)}\right\| \le \delta_U \left\|\bar{\boldsymbol{U}}_k\right\| \|\bar{S}_{kk}\|; \tag{27}$$

$$\bar{\mathcal{T}}_{1:k-1,k} = \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \bar{\boldsymbol{U}}_k + \Delta \boldsymbol{\mathcal{T}}_k, \quad \|\Delta \boldsymbol{\mathcal{T}}_k\| \le \delta_{Q^T U} \left\|\bar{\boldsymbol{U}}_k\right\|; \tag{28}$$

$$\bar{P}_k = \bar{\boldsymbol{U}}_k^T \bar{\boldsymbol{U}}_k + \Delta P_k, \quad \|\Delta P_k\| \le \delta_{U^T U} \left\|\bar{\boldsymbol{U}}_k\right\|^2; \tag{29}$$

$$\bar{T}_{kk}^T \bar{T}_{kk} = \bar{P}_k - \bar{\mathcal{T}}_{1:k-1,k}^T \bar{\mathcal{T}}_{1:k-1,k} + \Delta F_k^{(2)} + \Delta C_k^{(2)}, \tag{30}$$

$$\left\|\Delta F_k^{(2)}\right\| \le \delta_{T^T T} \left\|\bar{\boldsymbol{U}}_k\right\|^2 \text{ and } \left\|\Delta C_k^{(2)}\right\| \le \delta_{\texttt{chol}_2} \left\|\bar{\boldsymbol{U}}_k\right\|^2; \tag{31}$$

$$\bar{\boldsymbol{W}}_k = \bar{\boldsymbol{U}}_k - \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\mathcal{T}}_{1:k-1,k} + \Delta \boldsymbol{W}_k, \quad \|\Delta \boldsymbol{W}_k\| \le \delta_{QT} \left\|\bar{\boldsymbol{U}}_k\right\|; \text{ and} \tag{32}$$

$$\bar{\boldsymbol{Q}}_k \bar{T}_{kk} = \bar{\boldsymbol{W}}_k + \Delta \boldsymbol{G}_k^{(2)}, \quad \left\|\Delta \boldsymbol{G}_k^{(2)}\right\| \le \delta_Q \left\|\bar{\boldsymbol{Q}}_k\right\| \|\bar{T}_{kk}\|. \tag{33}$$

We have applied (21) and dropped quadratic error terms throughout the proofs in this section and in (22)–(33) to simplify the expressions; namely, $\|\boldsymbol{\mathcal{U}}_k\|$ and $\|\boldsymbol{\mathcal{Q}}_k\|$ are baked into the constants, where $\boldsymbol{\mathcal{U}}_k = \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 & \cdots & \boldsymbol{U}_k \end{bmatrix}$. We also again omit explicit dependence on k for readability. Note as well that each $\Delta F_k^{(i)}$ denotes the floating-point error from the subtraction of the inner product from the previously computed $\bar{\Omega}_k$ or $\bar{P}_k$, and $\Delta C_k^{(i)}$ denotes the error from the Cholesky factorization of that result.

In order to apply the Cholesky factorization and obtain (24) and (30) in the first place, we need that $\bar{\Omega}_k - \bar{\mathcal{S}}_{1:k-1,k}^T \bar{\mathcal{S}}_{1:k-1,k} + \Delta F_k^{(1)}$ and $\bar{P}_k - \bar{\mathcal{T}}_{1:k-1,k}^T \bar{\mathcal{T}}_{1:k-1,k} + \Delta F_k^{(2)}$ are symmetric positive definite. We show that $\bar{\Omega}_k - \bar{\mathcal{S}}_{1:k-1,k}^T \bar{\mathcal{S}}_{1:k-1,k} + \Delta F_k^{(1)}$ is positive definite using the following lemma; symmetry is already clear.

Lemma 1. *Fix $k \in \{2, \dots, p\}$ and suppose that (20)–(23) are satisfied, along with*

$$\bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\mathcal{R}}_{k-1} = \boldsymbol{\mathcal{X}}_{k-1} + \Delta \boldsymbol{\mathcal{D}}_{k-1}, \quad \|\Delta \boldsymbol{\mathcal{D}}_{k-1}\| \le \delta_X \|\boldsymbol{\mathcal{X}}_{k-1}\|. \tag{34}$$

Assume

$$\left(\delta_X + 2 \cdot \omega_{k-1} + \delta_{S^T S} + \delta_{X^T X} + 2\sqrt{2} \cdot \delta_{Q^T X}\right) \kappa^2(\boldsymbol{X}_k) < 1. \tag{35}$$

Then it holds that

$$\lambda_{\min}\left(\bar{\Omega}_k - \bar{\mathcal{S}}_{1:k-1,k}^T \bar{\mathcal{S}}_{1:k-1,k} + \Delta F_k^{(1)}\right) > 0. \tag{36}$$

Proof. By (22) and (23), and dropping quadratic error terms, we have

$$\begin{aligned} &\bar{\Omega}_k - \bar{\mathcal{S}}_{1:k-1,k}^T \bar{\mathcal{S}}_{1:k-1,k} + \Delta F_k^{(1)} \\ &\quad = \boldsymbol{X}_k^T (I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T)(I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T) \boldsymbol{X}_k + \Delta G_k, \end{aligned} \tag{37}$$

where

$$\begin{aligned} \Delta G_k := \Delta \Omega_k + \Delta F_k^{(1)} &- \boldsymbol{X}_k^T \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \Delta \boldsymbol{\mathcal{S}}_k - \Delta \boldsymbol{\mathcal{S}}_k^T \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \boldsymbol{X}_k \\ &+ \boldsymbol{X}_k^T \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T (I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T) \boldsymbol{X}_k \end{aligned}$$

satisfies the following, thanks to

$$\begin{aligned} \left\|\boldsymbol{X}_k^T \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T (I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T) \boldsymbol{X}_k\right\| &= \left\|\boldsymbol{X}_k^T \bar{\boldsymbol{\mathcal{Q}}}_{k-1} (I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \bar{\boldsymbol{\mathcal{Q}}}_{k-1}) \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \boldsymbol{X}_k\right\| \\ &\le \|\boldsymbol{X}_k\|^2 \left\|\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\right\|^2 \left\|I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\right\| \end{aligned}$$

and (21)–(23):

$$\begin{aligned}\|\Delta G_k\| &\leq \|\boldsymbol{X}_k\|^2 \left\|\bar{\boldsymbol{Q}}_{k-1}\right\|^2 \left\|I - \bar{\boldsymbol{Q}}_{k-1}^T \bar{\boldsymbol{Q}}_{k-1}\right\| + \delta_{S^TS} \|\boldsymbol{X}_k\|^2 \\ &\quad + \delta_{X^TX} \|\boldsymbol{X}_k\|^2 + 2\sqrt{1+\omega_{k-1}}\delta_{Q^TX} \|\boldsymbol{X}_k\|^2 \\ &\leq \left((1+\omega_{k-1})\omega_{k-1} + \delta_{S^TS} + \delta_{X^TX} + 2\sqrt{1+\omega_{k-1}}\delta_{Q^TX}\right) \|\boldsymbol{X}_k\|^2 .\end{aligned} \tag{38}$$

Using (37) and (38), we obtain

$$\begin{aligned}&\lambda_{\min}\left(\bar{\Omega}_k - \bar{\mathcal{S}}_{1:k-1,k}^T \bar{\mathcal{S}}_{1:k-1,k} + \Delta F_k^{(1)}\right) \\ &\quad \geq \lambda_{\min}(\boldsymbol{X}_k^T (I - \bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T)(I - \bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T)\boldsymbol{X}_k) - \|\Delta G_k\| \\ &\quad = \sigma_{\min}^2((I - \bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T)\boldsymbol{X}_k) - \|\Delta G_k\| \\ &\quad \geq \sigma_{\min}^2((I - \bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T)\boldsymbol{X}_k) - \left((1+\omega_{k-1})\omega_{k-1} + \delta_{S^TS}\right. \\ &\qquad \left. + \delta_{X^TX} + 2\sqrt{1+\omega_{k-1}}\delta_{Q^TX}\right) \|\boldsymbol{X}_k\|^2 .\end{aligned} \tag{39}$$

Then we estimate $\sigma_{\min}((I - \bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T)\boldsymbol{X}_k)$. Notice that $(I - \bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T)\boldsymbol{X}_k$ can be written as, by the assumption (34),

$$\begin{aligned}\boldsymbol{X}_k - \bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T\boldsymbol{X}_k &= \boldsymbol{X}_k - \bar{\boldsymbol{Q}}_{k-1}\bar{\mathcal{R}}_{k-1}\bar{\mathcal{R}}_{k-1}^{-1}\bar{\boldsymbol{Q}}_{k-1}^T\boldsymbol{X}_k \\ &= \begin{bmatrix}\boldsymbol{\mathcal{X}}_{k-1} + \Delta\boldsymbol{\mathcal{D}}_{k-1} & \boldsymbol{X}_k\end{bmatrix} \boldsymbol{C}^T\end{aligned} \tag{40}$$

with $\boldsymbol{C} := \begin{bmatrix}-(\bar{\mathcal{R}}_{k-1}^{-1}\bar{\boldsymbol{Q}}_{k-1}^T\boldsymbol{X}_k)^T & I_s\end{bmatrix} \in \mathbb{R}^{s\times sk}$, and I_s denoting the $s \times s$ identity matrix. Then from the perturbation theory of singular values [15, Corollary 8.6.2], we obtain

$$\begin{aligned}&\sigma_{\min}((I - \bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T)\boldsymbol{X}_k) \\ &= \sqrt{\min_{\boldsymbol{v}\in\mathbb{R}^{sk}\setminus\boldsymbol{0}} \left(\frac{\left\|\begin{bmatrix}\boldsymbol{\mathcal{X}}_{k-1} + \Delta\boldsymbol{\mathcal{D}}_{k-1} & \boldsymbol{X}_k\end{bmatrix} \boldsymbol{C}^T\boldsymbol{v}\right\|^2}{\|\boldsymbol{v}\|^2}\right)} \\ &= \sqrt{\min_{\boldsymbol{v}\in\mathbb{R}^{sk}\setminus\boldsymbol{0}} \left(\frac{\left\|\begin{bmatrix}\boldsymbol{\mathcal{X}}_{k-1} + \Delta\boldsymbol{\mathcal{D}}_{k-1} & \boldsymbol{X}_k\end{bmatrix} \boldsymbol{C}^T\boldsymbol{v}\right\|^2}{\|\boldsymbol{C}^T\boldsymbol{v}\|^2} \frac{\|\boldsymbol{C}^T\boldsymbol{v}\|^2}{\|\boldsymbol{v}\|^2}\right)} \\ &\geq \sqrt{\min_{\boldsymbol{v}\in\mathbb{R}^{sk}\setminus\boldsymbol{0}} \left(\frac{\left\|\begin{bmatrix}\boldsymbol{\mathcal{X}}_{k-1} + \Delta\boldsymbol{\mathcal{D}}_{k-1} & \boldsymbol{X}_k\end{bmatrix} (\boldsymbol{C}^T\boldsymbol{v})\right\|^2}{\|\boldsymbol{C}^T\boldsymbol{v}\|^2}\right) \min_{\boldsymbol{v}\in\mathbb{R}^{sk}\setminus\boldsymbol{0}} \left(\frac{\|\boldsymbol{C}^T\boldsymbol{v}\|^2}{\|\boldsymbol{v}\|^2}\right)} \\ &\geq \sigma_{\min}(\begin{bmatrix}\boldsymbol{\mathcal{X}}_{k-1} + \Delta\boldsymbol{\mathcal{D}}_{k-1} & \boldsymbol{X}_k\end{bmatrix}) \\ &\geq \sigma_{\min}(\boldsymbol{\mathcal{X}}_k) - \|\Delta\boldsymbol{\mathcal{D}}_{k-1}\| \\ &\geq \sigma_{\min}(\boldsymbol{\mathcal{X}}_k) - \delta_X \|\boldsymbol{\mathcal{X}}_{k-1}\| .\end{aligned}$$

Together with (39), it holds that

$$\begin{aligned}&\lambda_{\min}\left(\bar{\Omega}_k - \bar{\mathcal{S}}_{1:k-1,k}^T \bar{\mathcal{S}}_{1:k-1,k} + \Delta F_k^{(1)}\right) \\ &\quad \geq \sigma_{\min}^2(\boldsymbol{\mathcal{X}}_k) - \delta_X \|\boldsymbol{\mathcal{X}}_k\|^2 - \left((1+\omega_{k-1})\omega_{k-1} + \delta_{S^TS}\right. \\ &\qquad \left. + \delta_{X^TX} + 2\sqrt{1+\omega_{k-1}}\delta_{Q^TX}\right) \|\boldsymbol{X}_k\|^2 \\ &\quad \geq \sigma_{\min}^2(\boldsymbol{\mathcal{X}}_k)\big(1 - \big(\delta_X + (1+\omega_{k-1})\omega_{k-1} + \delta_{S^TS} \\ &\qquad + \delta_{X^TX} + 2\sqrt{1+\omega_{k-1}}\delta_{Q^TX}\big)\kappa^2(\boldsymbol{X}_k)\big).\end{aligned} \tag{41}$$

By dropping the quadratic error terms and using assumption (35), which guarantees

$$\left(\delta_X + (1+\omega_{k-1})\omega_{k-1} + \delta_{S^TS} + \delta_{X^TX} + 2\sqrt{1+\omega_{k-1}}\delta_{Q^TX}\right)\kappa^2(\boldsymbol{X}_k) < 1,$$

we can therefore conclude that $\lambda_{\min}\left(\bar{\Omega}_k - \bar{\mathcal{S}}_{1:k-1,k}^T \bar{\mathcal{S}}_{1:k-1,k} + \Delta F_k^{(1)}\right) > 0$. □

From Lemma 1, we have shown that $\bar{\Omega}_k - \bar{\mathcal{S}}_{1:k-1,k}^T \bar{\mathcal{S}}_{1:k-1,k} + \Delta F_k^{(1)}$ is symmetric positive definite. Before proving that $\bar{P}_k - \bar{\mathcal{T}}_{1:k-1,k}^T \bar{\mathcal{T}}_{1:k-1,k} + \Delta F_k^{(2)}$ is also symmetric positive definite, we need to first bound $\|\bar{\boldsymbol{U}}_k\|$.

Lemma 2. *Fix $k \in \{2, \dots, p\}$ and suppose that (20)–(27) are satisfied. Furthermore suppose that $\bar{\boldsymbol{\mathcal{Q}}}_{k-1}$ and $\bar{\mathcal{R}}_{k-1}$ satisfy the following:*

$$\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{R}}_{k-1} = \boldsymbol{\mathcal{X}}_{k-1} + \Delta\boldsymbol{\mathcal{D}}_{k-1}, \quad \|\Delta\boldsymbol{\mathcal{D}}_{k-1}\| \leq \delta_X \|\boldsymbol{\mathcal{X}}_{k-1}\|, \tag{42}$$

with $\delta_X, \omega_{k-1} \in (0,1)$. Assume

$$8 \cdot \delta_U \kappa^2(\boldsymbol{\mathcal{X}}) \leq 1 \tag{43}$$

and

$$\begin{aligned}&\big(\delta_X + 2\cdot\omega_{k-1} + 2\cdot\delta_{S^TS} + 2\cdot\delta_{X^TX} + 2\cdot\delta_{\texttt{chol}_1} \\ &+ 18\cdot\delta_{Q^TX} + 8\cdot\delta_{QS}\big)\kappa^2(\boldsymbol{\mathcal{X}}_k) \leq \frac{1}{2}.\end{aligned} \tag{44}$$

Then it holds that

$$\bar{S}_{kk}^T\bar{S}_{kk} = \boldsymbol{X}_k^T\boldsymbol{X}_k - \boldsymbol{X}_k^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\boldsymbol{X}_k + \Delta S_{kk}, \text{ and} \tag{45}$$

$$\bar{\boldsymbol{U}}_k\bar{S}_{kk} = \boldsymbol{X}_k - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\boldsymbol{X}_k + \Delta\boldsymbol{U}_k, \text{ where} \tag{46}$$

$$\|\Delta S_{kk}\| \leq \left(\delta_{S^TS} + \delta_{X^TX} + \delta_{\texttt{chol}_1} + 2\sqrt{2}\cdot\delta_{Q^TX}\right)\|\boldsymbol{X}_k\|^2, \tag{47}$$

$$\|\bar{S}_{kk}\| \leq 3\|\boldsymbol{X}_k\|, \tag{48}$$

$$\|\bar{S}_{kk}^{-1}\| \leq \frac{\sqrt{2}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_k)}, \tag{49}$$

$$\|\Delta\boldsymbol{U}_k\| \leq \left(\delta_U\|\bar{\boldsymbol{U}}_k\| + \sqrt{2}\cdot\delta_{Q^TX} + \delta_{QS}\right)\|\boldsymbol{X}_k\|, \text{ and} \tag{50}$$

$$\|\bar{\boldsymbol{U}}_k\| \leq 2. \tag{51}$$

Proof. From (21), (22), and (23), we obtain

$$\begin{aligned}\|\bar{\mathcal{S}}_{1:k-1,k}\| &\leq (\sqrt{2} + \delta_{Q^TX})\|\boldsymbol{X}_k\| \text{ and} \\ \|\bar{\Omega}_k\| &\leq (1 + \delta_{X^TX})\|\boldsymbol{X}_k\|^2.\end{aligned} \tag{52}$$

Substituting (22), (23), and (52) into (24) gives (45) with

$$\Delta S_{kk} := \Delta\Omega_k + \Delta F_k^{(1)} + \Delta C_k^{(1)} - \boldsymbol{X}_k^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\Delta\boldsymbol{S}_k - \Delta\boldsymbol{S}_k^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\boldsymbol{X}_k.$$

From the bounds (21), (22), (23), (24), and (25), the desired bounds (47) and (48) then follow immediately.

To find (49), we rewrite (45) as

$$\bar{S}_{kk}^T\bar{S}_{kk} = \boldsymbol{X}_k^T(I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T)(I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T)\boldsymbol{X}_k + \Delta\widetilde{S}_k, \tag{53}$$

where $\Delta\widetilde{S}_k := \boldsymbol{X}_k^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T(I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T)\boldsymbol{X}_k + \Delta S_{kk}$. Using a similar logic in the proof of Lemma 1 together with (20) and (21) we obtain (49) since $\|\bar{S}_{kk}^{-1}\| = \frac{1}{\sigma_{\min}(\bar{S}_{kk})}$ and

$$\begin{aligned}\sigma_{\min}^2(\bar{S}_{kk}) \geq \sigma_{\min}^2(\boldsymbol{\mathcal{X}}_k) - \delta_X\|\boldsymbol{\mathcal{X}}_k\|^2 - \big(&\omega_{k-1} + \delta_{S^TS} \\ &+ \delta_{X^TX} + \delta_{\texttt{chol}_1} + 2\sqrt{2}\cdot\delta_{Q^TX}\big)\|\boldsymbol{X}_k\|^2.\end{aligned} \tag{54}$$

Note that the above reasoning makes sense only if

$$\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_k) > \left(\delta_X + \omega_{k-1} + \delta_{S^TS} + \delta_{X^TX} + \delta_{\texttt{chol}_1} + 2\sqrt{2}\cdot\delta_{Q^TX}\right)\|\boldsymbol{X}_k\|^2;$$

that is, if

$$\left(\delta_X + \omega_{k-1} + \delta_{S^TS} + \delta_{X^TX} + \delta_{\texttt{chol}_1} + 2\sqrt{2}\cdot\delta_{Q^TX}\right)\kappa^2(\boldsymbol{\mathcal{X}}_k) < 1,$$

which can be guaranteed by the assumption (44).

To prove (50) we first note from (26) and the substitution of (22) that

$$\begin{aligned} \bar{\boldsymbol{V}}_k &= \boldsymbol{X}_k - \bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T\boldsymbol{X}_k + \Delta\boldsymbol{V}_k, \\ \|\Delta\boldsymbol{V}_k\| &\leq \left(\sqrt{2}\cdot\delta_{Q^TX} + \delta_{QS}\right)\|\boldsymbol{X}_k\|. \end{aligned} \tag{55}$$

Substituting (55) into (27) leads to (46), with

$$\Delta\boldsymbol{U}_k := \Delta\boldsymbol{G}_k^{(1)} + \Delta\boldsymbol{V}_k$$

and the desired bound (50) satisfied.

The final bound (51) requires a bit more work. Multiplying $\bar{\boldsymbol{U}}_k\bar{S}_{kk}$ with its transpose gives

$$\begin{aligned} (\bar{\boldsymbol{U}}_k\bar{S}_{kk})^T(\bar{\boldsymbol{U}}_k\bar{S}_{kk}) &= \boldsymbol{X}_k^T\boldsymbol{X}_k - 2\boldsymbol{X}_k^T\bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T\boldsymbol{X}_k \\ &\quad + \boldsymbol{X}_k^T\bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T\bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T\boldsymbol{X}_k + \Delta H_k \\ &= \boldsymbol{X}_k^T\boldsymbol{X}_k - \boldsymbol{X}_k^T\bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T\boldsymbol{X}_k \\ &\quad + \boldsymbol{X}_k^T\bar{\boldsymbol{Q}}_{k-1}(\bar{\boldsymbol{Q}}_{k-1}^T\bar{\boldsymbol{Q}}_{k-1} - I)\bar{\boldsymbol{Q}}_{k-1}^T\boldsymbol{X}_k + \Delta H_k, \end{aligned} \tag{56}$$

where

$$\Delta H_k := \boldsymbol{X}_k^T\Delta\boldsymbol{U}_k + \left(\Delta\boldsymbol{U}_k\right)^T\boldsymbol{X}_k + \boldsymbol{X}_k^T\bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T\Delta\boldsymbol{U}_k + \left(\Delta\boldsymbol{U}_k\right)^T\bar{\boldsymbol{Q}}_{k-1}\bar{\boldsymbol{Q}}_{k-1}^T\boldsymbol{X}_k.$$

Applying (21) and (50) gives

$$\|\Delta H_k\| \leq \left(4 + 2\cdot\omega_{k-1}\right)\left(\delta_U\left\|\bar{\boldsymbol{U}}_k\right\| + \sqrt{2}\cdot\delta_{Q^TX} + \delta_{QS}\right)\|\boldsymbol{X}_k\|^2. \tag{57}$$

Substituting (45) into (56) leads to

$$(\bar{\boldsymbol{U}}_k\bar{S}_{kk})^T(\bar{\boldsymbol{U}}_k\bar{S}_{kk}) = \bar{S}_{kk}^T\bar{S}_{kk} + \boldsymbol{X}_k^T\bar{\boldsymbol{Q}}_{k-1}(\bar{\boldsymbol{Q}}_{k-1}^T\bar{\boldsymbol{Q}}_{k-1} - I)\bar{\boldsymbol{Q}}_{k-1}^T\boldsymbol{X}_k + \Delta H_k - \Delta S_{kk},$$

and then multiplying by S_{kk}^{-T} on the left and S_{kk}^{-1} on the right yields

$$\begin{aligned} \bar{\boldsymbol{U}}_k^T\bar{\boldsymbol{U}}_k &= I + \bar{S}_{kk}^{-T}\boldsymbol{X}_k^T\bar{\boldsymbol{Q}}_{k-1}(\bar{\boldsymbol{Q}}_{k-1}^T\bar{\boldsymbol{Q}}_{k-1} - I)\bar{\boldsymbol{Q}}_{k-1}^T\boldsymbol{X}_k\bar{S}_{kk}^{-1} \\ &\quad + \bar{S}_{kk}^{-T}(\Delta H_k - \Delta S_{kk})\bar{S}_{kk}^{-1}. \end{aligned} \tag{58}$$

Recalling the assumption (43), we note that

$$4\big(2+\omega_{k-1}\big)\delta_U\kappa^2(\boldsymbol{\mathcal{X}}_k) \leq 4\big(2+\omega_{k-1}\big)\delta_U\kappa^2(\boldsymbol{\mathcal{X}}) \leq 1.$$

Applying (20), (21), (57), (47), (43) along with (49), taking norms yields the following:

$$\begin{aligned} \left\|\bar{\boldsymbol{U}}_k\right\|^2 &= \left\|\bar{\boldsymbol{U}}_k^T\bar{\boldsymbol{U}}_k\right\| \\ &\leq 1 + \omega_{k-1}\left\|\bar{S}_{kk}^{-1}\right\|^2\|\boldsymbol{X}_k\|^2 + 4\big(\delta_U\left\|\bar{\boldsymbol{U}}_k\right\| + \sqrt{2}\cdot\delta_{Q^TX} \\ &\quad + \delta_{QS}\big)\left\|\bar{S}_{kk}^{-1}\right\|^2\|\boldsymbol{X}_k\|^2 + \big(\delta_{S^TS} + \delta_{X^TX} + \delta_{\texttt{chol}_1} \\ &\quad + 2\sqrt{2}\cdot\delta_{Q^TX}\big)\left\|\bar{S}_{kk}^{-1}\right\|^2\|\boldsymbol{X}_k\|^2 \\ &\leq 1 + 2\big(\omega_{k-1} + \delta_{S^TS} + \delta_{X^TX} + \delta_{\texttt{chol}_1} + 9\cdot\delta_{Q^TX} + 4\cdot\delta_{QS}\big)\kappa^2(\boldsymbol{\mathcal{X}}_k) \\ &\quad + 8\cdot\delta_U\left\|\bar{\boldsymbol{U}}_k\right\|\kappa^2(\boldsymbol{\mathcal{X}}_k) \\ &\leq \frac{3}{2} + \left\|\bar{\boldsymbol{U}}_k\right\|. \end{aligned} \tag{59}$$

Solving the quadratic inequality (59) gives

$$\left\|\bar{\boldsymbol{U}}_k\right\| \leq 2.$$

□

Now using similar logic as in the proof of Lemma 1, we can combine the above lemma with (28) and (29) to conclude that $\bar{P}_k - \bar{\mathcal{T}}_{1:k-1,k}^T \bar{\mathcal{T}}_{1:k-1,k} + \Delta F_k^{(2)}$ is symmetric positive definite.

The following theorem makes the relationship between each source of error and the bound on the LOO explicit.

Theorem 2. *Fix $k \in \{2, \dots, p\}$ and assume that (20)–(33) are satisfied with*

$$\begin{aligned} &2\big(4 \cdot \omega_{k-1} + \delta_{S^TS} + \delta_{X^TX} + \delta_{\texttt{chol}_1} + 7 \cdot \delta_{QS} + 13 \cdot \delta_{Q^TX} + 14 \cdot \delta_U\big)\kappa^2(\boldsymbol{\mathcal{X}}) \\ &+ 4\big(\omega_{k-1} + \delta_{U^TU} + \delta_{T^TT} + \delta_{\texttt{chol}_2} + 3 \cdot \delta_{Q^TU}\big) \leq \frac{1}{2}. \end{aligned} \tag{60}$$

Further assume that the assumptions of Lemma 2 hold. Then $\bar{\boldsymbol{\mathcal{Q}}}_k$ satisfies

$$\begin{aligned} \left\| I - \bar{\boldsymbol{\mathcal{Q}}}_k^T \bar{\boldsymbol{\mathcal{Q}}}_k \right\| \leq\ & 17 \cdot \omega_{k-1} + 30 \cdot \delta_{Q^TU} + 40 \cdot \delta_{QT} \\ & + 120 \cdot \delta_Q + 8 \cdot \delta_{T^TT} + 8 \cdot \delta_{\texttt{chol}_2}. \end{aligned} \tag{61}$$

Proof. Writing $\bar{\boldsymbol{\mathcal{Q}}}_k = \begin{bmatrix} \bar{\boldsymbol{\mathcal{Q}}}_{k-1} & \bar{\boldsymbol{Q}}_k \end{bmatrix}$, we can look at $I - \bar{\boldsymbol{\mathcal{Q}}}_k^T \bar{\boldsymbol{\mathcal{Q}}}_k$ block-by-block:

$$I - \bar{\boldsymbol{\mathcal{Q}}}_k^T \bar{\boldsymbol{\mathcal{Q}}}_k = \begin{bmatrix} I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \bar{\boldsymbol{\mathcal{Q}}}_{k-1} & \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \bar{\boldsymbol{Q}}_k \\ \bar{\boldsymbol{Q}}_k^T \bar{\boldsymbol{\mathcal{Q}}}_{k-1} & I - \bar{\boldsymbol{Q}}_k^T \bar{\boldsymbol{Q}}_k \end{bmatrix}. \tag{62}$$

The induction hypothesis takes care of bounding the upper left block. For the off-diagonals, by (51), (28), (29), and Lemma 2, we obtain

$$\begin{aligned} \left\| \bar{\mathcal{T}}_{1:k-1,k} \right\| &\leq 2\big(\sqrt{1 + \omega_{k-1}} + \delta_{Q^TU}\big), \\ \left\| \bar{P}_k \right\| &\leq 4\big(1 + \delta_{U^TU}\big). \end{aligned} \tag{63}$$

Using (30), the assumption (60), and dropping the quadratic terms, the following bound holds:

$$\begin{aligned} \left\| \bar{T}_{kk} \right\| &\leq \sqrt{4\big(1 + \delta_{U^TU}\big) + 4\big(\sqrt{1 + \omega_{k-1}} + \delta_{Q^TU}\big)^2 + 4\delta_{T^TT} + 4\delta_{\texttt{chol}_2}} \\ &\leq \sqrt{\frac{17}{2}} \leq 3. \end{aligned} \tag{64}$$

Using (28) and (29), we can rewrite (30) as

$$\begin{aligned} \bar{T}_{kk}^T \bar{T}_{kk} &= \bar{\boldsymbol{U}}_k^T (I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T)(I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T) \bar{\boldsymbol{U}}_k + \Delta \widetilde{T}_k, \\ \left\| \Delta \widetilde{T}_k \right\| &\leq 4\big(\omega_{k-1} + \delta_{U^TU} + \delta_{T^TT} + \delta_{\texttt{chol}_2} + 3 \cdot \delta_{Q^TU}\big). \end{aligned} \tag{65}$$

To bound $\left\| \bar{T}_{kk}^{-1} \right\|$ we can use (65) to write

$$\begin{aligned} \sigma_{\min}^2(\bar{T}_{kk}) &\geq \sigma_{\min}^2\big((I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T) \bar{\boldsymbol{U}}_k\big) - \left\| \Delta \widetilde{T}_k \right\| \\ &\geq \sigma_{\min}^2\big((I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T) \bar{\boldsymbol{U}}_k\big) \\ &\quad - 4\big(\omega_{k-1} + \delta_{U^TU} + \delta_{T^TT} + \delta_{\texttt{chol}_2} + 3 \cdot \delta_{Q^TU}\big). \end{aligned} \tag{66}$$

By (46), we find that

$$\bar{\boldsymbol{U}}_k = \Big((I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T) \boldsymbol{X}_k + \Delta \boldsymbol{U}_k\Big) \bar{S}_{kk}^{-1}. \tag{67}$$

Multiplying (67) by $I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T$ on the left yields

$$\begin{aligned} (I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T) \bar{\boldsymbol{U}}_k = \bar{\boldsymbol{U}}_k &- \bar{\boldsymbol{\mathcal{Q}}}_{k-1} (I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \bar{\boldsymbol{\mathcal{Q}}}_{k-1}) \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \boldsymbol{X}_k \bar{S}_{kk}^{-1} \\ &- \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \Delta \boldsymbol{U}_k \bar{S}_{kk}^{-1}. \end{aligned} \tag{68}$$

Define $\Delta \boldsymbol{E}_k := \bar{\boldsymbol{\mathcal{Q}}}_{k-1} (I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \bar{\boldsymbol{\mathcal{Q}}}_{k-1}) \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \boldsymbol{X}_k \bar{S}_{kk}^{-1} + \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T \Delta \boldsymbol{U}_k \bar{S}_{kk}^{-1}$. Using (49), we then have

$$\|\boldsymbol{X}_k\| \left\| \bar{S}_{kk}^{-1} \right\| \leq \sqrt{2} \cdot \kappa(\boldsymbol{\mathcal{X}}_k).$$

Together with the induction hypothesis and (51), we can bound $\|\Delta \boldsymbol{E}_k\|$ by

$$\|\Delta \boldsymbol{E}_k\| \leq \sqrt{2}\big(\omega_{k-1} + 2\cdot\delta_U + \sqrt{2}\delta_{Q^TX} + \delta_{QS}\big)\kappa(\boldsymbol{\mathcal{X}}_k). \tag{69}$$

Via (58), (59), (68), and (69), we find lower bounds on some key singular values[1]:

$$\begin{aligned}\sigma_{\min}^2(\bar{\boldsymbol{U}}_k) &\geq 1 - \left\|\bar{S}_{kk}^{-T}\boldsymbol{X}_k^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}(I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1})\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\boldsymbol{X}_k\bar{S}_{kk}^{-1}\right\| \\ &\quad - \left\|\bar{S}_{kk}^{-T}(\Delta H_k - \Delta S_{kk})\bar{S}_{kk}^{-1}\right\| \\ &\geq 1 - 2\big(\omega_{k-1} + \delta_{S^TS} + \delta_{X^TX} + \delta_{\texttt{chol}_1} + 6\sqrt{2}\cdot\delta_{Q^TX} + 4\cdot\delta_{QS} \\ &\quad + 8\cdot\delta_U\big)\kappa^2(\boldsymbol{\mathcal{X}}_k),\end{aligned}$$

and

$$\begin{aligned}&\sigma_{\min}^2((I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T)\bar{\boldsymbol{U}}_k) \\ &\qquad \geq \left(\sigma_{\min}(\bar{\boldsymbol{U}}_k) - \|\Delta \boldsymbol{E}_k\|\right)^2 \\ &\qquad \geq \sigma_{\min}^2(\bar{\boldsymbol{U}}_k) - 2\,\|\Delta \boldsymbol{E}_k\|\,\|\bar{\boldsymbol{U}}_k\| \\ &\qquad \geq \sigma_{\min}^2(\bar{\boldsymbol{U}}_k) - 4\sqrt{2}\big(\omega_{k-1} + 2\cdot\delta_U + \sqrt{2}\delta_{Q^TX} + \delta_{QS}\big)\kappa(\boldsymbol{\mathcal{X}}_k) \\ &\qquad \geq 1 - 2\big(4\cdot\omega_{k-1} + \delta_{S^TS} + \delta_{X^TX} + \delta_{\texttt{chol}_1} + 13\cdot\delta_{Q^TX} \\ &\qquad\quad + 7\cdot\delta_{QS} + 14\cdot\delta_U\big)\kappa^2(\boldsymbol{\mathcal{X}}_k).\end{aligned} \tag{70}$$

Combining (70) with (65), (66), and using the fact that $\left\|\bar{T}_{kk}^{-1}\right\|^2 = \frac{1}{\sigma_{\min}^2(\bar{T}_{kk})}$, we arrive at

$$\left\|\bar{T}_{kk}^{-1}\right\| = \sqrt{\left\|\bar{T}_{kk}^{-1}\right\|^2} \leq \sqrt{2}. \tag{71}$$

With similar logic as in the derivation of (27), (33) gives

$$\bar{\boldsymbol{Q}}_k\bar{T}_{kk} = \bar{\boldsymbol{W}}_k + \Delta\boldsymbol{G}_k^{(2)}, \quad \left\|\Delta\boldsymbol{G}_k^{(2)}\right\| \leq \delta_Q\left\|\bar{\boldsymbol{Q}}_k\right\|\left\|\bar{T}_{kk}\right\| \leq 6\cdot\delta_Q, \tag{72}$$

where we have applied (21) and (64) and simplified the bound. Multiplying (72) by $\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T$ on the left yields

$$\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{Q}}_k\bar{T}_{kk} = \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{W}}_k + \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\Delta\boldsymbol{G}_k^{(2)}. \tag{73}$$

Then we substitute (32) into (73), multiply both sides by $\bar{T}_{kk}^{-1}$, and use (28) to obtain

$$\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{Q}}_k = (I - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1})\mathcal{T}_{1:k-1,k}\bar{T}_{kk}^{-1} + \left(\Delta\boldsymbol{T}_k + \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T(\Delta\boldsymbol{G}_k^{(2)} + \Delta\boldsymbol{W}_k)\right)\bar{T}_{kk}^{-1}.$$

Combining (61), (51), Lemma 2, and bounds (28), (71), (72), and (32) leads to

$$\left\|\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{Q}}_k\right\| \leq 4\cdot\omega_{k-1} + 2\sqrt{2}\cdot\delta_{Q^TU} + 12\cdot\delta_Q + 4\cdot\delta_{QT}. \tag{74}$$

To bound the bottom right entry of (62), we combine (30) and (33) and note that

$$\begin{aligned}\bar{T}_{kk}^T(I - \bar{\boldsymbol{Q}}_k^T\bar{\boldsymbol{Q}}_k)\bar{T}_{kk} &= \bar{T}_{kk}^T\bar{T}_{kk} - \bar{T}_{kk}^T\bar{\boldsymbol{Q}}_k^T\bar{\boldsymbol{Q}}_k\bar{T}_{kk} \\ &= \bar{P}_k - \bar{\mathcal{T}}_{1:k-1,k}^T\bar{\mathcal{T}}_{1:k-1,k} + \Delta F_k^{(2)} + \Delta C_k^{(2)} \\ &\quad - (\bar{\boldsymbol{W}}_k - \Delta\boldsymbol{G}_k^{(2)})^T(\bar{\boldsymbol{W}}_k - \Delta\boldsymbol{G}_k^{(2)}) \\ &= \bar{P}_k - \bar{\mathcal{T}}_{1:k-1,k}^T\bar{\mathcal{T}}_{1:k-1,k} - \bar{\boldsymbol{W}}_k^T\bar{\boldsymbol{W}}_k \\ &\quad \underbrace{- \bar{\boldsymbol{W}}_k^T\Delta\boldsymbol{G}_k^{(2)} - (\Delta\boldsymbol{G}_k^{(2)})^T\bar{\boldsymbol{W}}_k + \Delta F_k^{(2)} + \Delta C_k^{(2)}}_{=:\Delta L_k}.\end{aligned} \tag{75}$$

[1] again by [15, Corollary 8.6.2].

Further substituting (29) and (32) into (75) simplifies to

$$\begin{aligned}\bar{T}_{kk}^T(I-\bar{\boldsymbol{Q}}_k^T\bar{\boldsymbol{Q}}_k)\bar{T}_{kk} = &-\bar{\mathcal{T}}_{1:k-1,k}^T\bar{\mathcal{T}}_{1:k-1,k}+\bar{\boldsymbol{U}}_k^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{T}}_{1:k-1,k}\\ &+\bar{\mathcal{T}}_{1:k-1,k}^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{U}}_k-\bar{\mathcal{T}}_{1:k-1,k}^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{T}}_{1:k-1,k}\\ &-\Delta J_k-\Delta L_k,\end{aligned} \tag{76}$$

where $\Delta J_k := \left(\bar{\boldsymbol{U}}_k-\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{T}}_{1:k-1,k}\right)^T\Delta\boldsymbol{W}_k-\left(\Delta\boldsymbol{W}_k\right)^T\left(\bar{\boldsymbol{U}}_k-\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{T}}_{1:k-1,k}\right)$. One more substitution of (28) into (76) and further simplification yields

$$\begin{aligned}\bar{T}_{kk}^T(I-\bar{\boldsymbol{Q}}_k^T\bar{\boldsymbol{Q}}_k)\bar{T}_{kk} = &\,\bar{\mathcal{T}}_{1:k-1,k}^T(I-\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1})\bar{\mathcal{T}}_{1:k-1,k}\\ &-\left(\Delta\boldsymbol{T}_k\right)^T\bar{\mathcal{T}}_{1:k-1,k}-\bar{\mathcal{T}}_{1:k-1,k}^T\Delta\boldsymbol{T}_k-\Delta J_k-\Delta L_k.\end{aligned} \tag{77}$$

Multiplying (77) by $\bar{T}_{kk}^{-T}$ on the left and $\bar{T}_{kk}^{-1}$ on the right, taking norms, and applying nearly all previous bounds along with (20) leads to

$$\begin{aligned}\left\|I-\bar{\boldsymbol{Q}}_k^T\bar{\boldsymbol{Q}}_k\right\| \le &\,8\cdot\omega_{k-1}+24\cdot\delta_{Q^TU}+32\cdot\delta_{QT}\\ &+96\cdot\delta_Q+8\cdot\delta_{T^TT}+8\cdot\delta_{\texttt{chol}_2}.\end{aligned} \tag{78}$$

Finally, using (62) along with (20) and the bounds (74) and (78), we see that

$$\begin{aligned}\left\|I-\bar{\boldsymbol{\mathcal{Q}}}_k^T\bar{\boldsymbol{\mathcal{Q}}}_k\right\| &= \left\|\begin{bmatrix}I-\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1} & \bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{Q}}_k\\ \bar{\boldsymbol{Q}}_k^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1} & I-\bar{\boldsymbol{Q}}_k^T\bar{\boldsymbol{Q}}_k\end{bmatrix}\right\|\\ &\le \left\|\begin{bmatrix}\left\|I-\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\right\| & \left\|\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{Q}}_k\right\|\\ \left\|\bar{\boldsymbol{Q}}_k^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\right\| & \left\|I-\bar{\boldsymbol{Q}}_k^T\bar{\boldsymbol{Q}}_k\right\|\end{bmatrix}\right\|\\ &\le \left\|\begin{bmatrix}\left\|I-\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\right\| & \left\|\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{Q}}_k\right\|\\ \left\|\bar{\boldsymbol{Q}}_k^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\right\| & \left\|I-\bar{\boldsymbol{Q}}_k^T\bar{\boldsymbol{Q}}_k\right\|\end{bmatrix}\right\|_{\mathrm{F}}\\ &\le \left\|I-\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\right\|+2\left\|\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\bar{\boldsymbol{Q}}_k\right\|+\left\|I-\bar{\boldsymbol{Q}}_k^T\bar{\boldsymbol{Q}}_k\right\|\\ &\le 17\cdot\omega_{k-1}+30\cdot\delta_{Q^TU}+40\cdot\delta_{QT}+120\cdot\delta_Q+8\cdot\delta_{T^TT}+8\cdot\delta_{\texttt{chol}_2},\end{aligned}$$

where we have used [13, P.15.50] as in [8, Theorem 3.1]. □

If we assume a uniform working precision with unit roundoff ε and apply standard floating-point point analysis [16] to Theorem 2, it is not hard to show that

$$\delta_{Q^TX},\delta_{X^TX},\delta_{S^TS},\delta_{QS},\delta_U,\delta_{Q^TU},\delta_{U^TU},\delta_{T^TT},\delta_{QT},\delta_Q \le \mathcal{O}(\varepsilon)$$

and

$$\delta_{\texttt{chol}_1},\delta_{\texttt{chol}_2} \le \mathcal{O}(\varepsilon).$$

In contrast to Corollary 1, we must impose a LOO condition on `IO`; `HouseQR` [16], `TSQR` [20], `CholQR++` [28], or `ShCholQR++` [12] should satisfy the requirement, but only `TSQR` could do so and still maintain a single sync point for the `IO` [2][2]. The following corollaries summarize bounds for `BCGS-PIPI+` in uniform precision.

Corollary 3. *Assume that* $\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}) \le \frac{1}{2}$ *and that for all* $\boldsymbol{X}\in\mathbb{R}^{m\times s}$ *with* $\kappa(\boldsymbol{X}) \le \kappa(\boldsymbol{\mathcal{X}})$, $[\bar{\boldsymbol{Q}},\bar{R}] = \texttt{IO}(\boldsymbol{X})$ *satisfy*

$$\begin{aligned}\bar{R}^T\bar{R} &= \boldsymbol{X}^T\boldsymbol{X}+\Delta E, \quad \|\Delta E\| \le \mathcal{O}(\varepsilon)\|\boldsymbol{X}\|^2\\ \bar{\boldsymbol{Q}}\bar{R} &= \boldsymbol{X}+\Delta\boldsymbol{D}, \quad \|\Delta\boldsymbol{D}\| \le \mathcal{O}(\varepsilon)\|\boldsymbol{X}\| \;\; \textit{and}\\ \left\|I-\bar{\boldsymbol{Q}}^T\bar{\boldsymbol{Q}}\right\| &\le \frac{\mathcal{O}(\varepsilon)}{1-\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{X})}.\end{aligned}$$

[2]Note that TSQR can be performed as a single reduction, but the resulting $\boldsymbol{Q}$ factor is implicitly represented in a tree-based format; if the Householder representation of the $\boldsymbol{Q}$ factor is desired, a second reduction is required.

Assume also that $[\bar{\boldsymbol{\mathcal{Q}}}, \mathcal{R}] =$ BCGS-PIPI+$(\boldsymbol{\mathcal{X}}, \texttt{IO})$ *and for all* $k \in \{3, \dots, p\}$,

$$\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{R}}_{k-1} = \boldsymbol{\mathcal{X}}_{k-1} + \Delta\boldsymbol{\mathcal{D}}_{k-1}, \quad \|\Delta\boldsymbol{\mathcal{D}}_{k-1}\| \leq \mathcal{O}(\varepsilon)\|\boldsymbol{\mathcal{X}}_{k-1}\|.$$

Then for all $k \in \{1, \dots, p\}$, $\bar{\boldsymbol{\mathcal{Q}}}_k$ *satisfies*

$$\left\|I - \bar{\boldsymbol{\mathcal{Q}}}_k^T \bar{\boldsymbol{\mathcal{Q}}}_k\right\| \leq \frac{\mathcal{O}(\varepsilon)}{1 - \mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}_k)} \leq \mathcal{O}(\varepsilon). \tag{79}$$

Proof. First note that $\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}) \leq \frac{1}{2}$ implies that for all $k \in \{1, \dots, p\}$, $\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}_k) \leq \frac{1}{2}$. Then for the base case, the assumptions on IO directly give

$$\left\|I - \bar{\boldsymbol{\mathcal{Q}}}_1^T \bar{\boldsymbol{\mathcal{Q}}}_1\right\| = \|I - \bar{\boldsymbol{Q}}_1^T \bar{\boldsymbol{Q}}_1\| \leq \frac{\mathcal{O}(\varepsilon)}{1 - \mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{X}_1)} = \frac{\mathcal{O}(\varepsilon)}{1 - \mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}_1)},$$

which also means that $\omega_1 \leq \frac{\mathcal{O}(\varepsilon)}{1-\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}_1)}$. Now assume that (79) holds for all $j \in \{1, \dots, k-1\}$, i.e., $\omega_{k-1} \leq \frac{\mathcal{O}(\varepsilon)}{1-\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}_{k-1})}$. Noting that the assumption $\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}) \leq \frac{1}{2}$ can guarantee (60), Theorem 2 proves that (61) holds for k.

Note that it seems that we prove that $\left\|I - \bar{\boldsymbol{\mathcal{Q}}}_k^T \bar{\boldsymbol{\mathcal{Q}}}_k\right\| \leq \frac{\mathcal{O}(\varepsilon)}{1-\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}_{k-1})}$, which is because we omitted $\frac{1}{1-\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}_{k-1})}$ for simplicity using the assumption $\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}) \leq \frac{1}{2}$ in (71). Replacing (71) in the proof of Theorem 2 with

$$\|\bar{T}_{kk}^{-1}\| = \sqrt{\|\bar{T}_{kk}^{-1}\|^2} \leq \sqrt{\frac{1}{1 - \mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}_k)}} \leq \frac{1}{1 - \mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}_k)}, \tag{80}$$

we can draw the conclusion. □

A key takeaway from Corollary 3, in particular the bound (61), is that the LOO of BCGS-PIPI+ depends on the conditioning of $\boldsymbol{\mathcal{X}}$. We have made the rather artificial assumption that $\mathcal{O}(\varepsilon)\kappa(\boldsymbol{\mathcal{X}}) \leq \frac{1}{2}$. Of course a constant closer to 1 could be used instead, and it would become clear that the constant on $\mathcal{O}(\varepsilon)$ can grow arbitrarily large the closer we let $\kappa(\boldsymbol{\mathcal{X}})$ get to $\frac{1}{\sqrt{\varepsilon}}$. In practice, this edge-case behavior can be quite dramatic, which examples in Section 4 demonstrate.

We close this section by bounding both the standard residual (2) and Cholesky residual (3) for Algorithm 3 in uniform precision.

Corollary 4. *Assume that* $\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}) \leq \frac{1}{2}$ *and that for all* $\boldsymbol{X} \in \mathbb{R}^{m\times s}$ *with* $\kappa(\boldsymbol{X}) \leq \kappa(\boldsymbol{\mathcal{X}})$, $[\bar{\boldsymbol{Q}}, \bar{R}] = \texttt{IO}(\boldsymbol{X})$ *satisfy*

$$\bar{\boldsymbol{Q}}\bar{R} = \boldsymbol{X} + \Delta\boldsymbol{D}, \quad \|\Delta\boldsymbol{D}\| \leq \mathcal{O}(\varepsilon)\|\boldsymbol{X}\| \text{ and } \left\|I - \bar{\boldsymbol{Q}}^T\bar{\boldsymbol{Q}}\right\| \leq \frac{\mathcal{O}(\varepsilon)}{1 - \mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{X})}.$$

Then for $[\bar{\boldsymbol{\mathcal{Q}}}, \mathcal{R}] =$ BCGS-PIPI+$(\boldsymbol{\mathcal{X}}, \texttt{IO})$ *and all* $k \in \{1, \dots, p\}$,

$$\bar{\boldsymbol{\mathcal{Q}}}_k\bar{\mathcal{R}}_k = \boldsymbol{\mathcal{X}}_k + \Delta\boldsymbol{\mathcal{D}}_k, \quad \|\Delta\boldsymbol{\mathcal{D}}_k\| \leq \mathcal{O}(\varepsilon)\|\boldsymbol{\mathcal{X}}_k\|.$$

Proof. By the assumption on IO, we have for $k = 1$,

$$\bar{\boldsymbol{\mathcal{Q}}}_1\bar{\mathcal{R}}_1 = \boldsymbol{\mathcal{X}}_1 + \Delta\boldsymbol{\mathcal{D}}_1, \quad \|\Delta\boldsymbol{\mathcal{D}}_1\| \leq \mathcal{O}(\varepsilon)\|\boldsymbol{\mathcal{X}}_1\|.$$

Now we assume that for all $j \in \{2, \dots, k-1\}$, it holds that

$$\bar{\boldsymbol{\mathcal{Q}}}_j\bar{\mathcal{R}}_j = \boldsymbol{\mathcal{X}}_j + \Delta\boldsymbol{\mathcal{D}}_j, \quad \|\Delta\boldsymbol{\mathcal{D}}_j\| \leq \mathcal{O}(\varepsilon)\|\boldsymbol{\mathcal{X}}_j\|. \tag{81}$$

For k, it then follows that

$$\begin{aligned}\Delta\boldsymbol{\mathcal{D}}_k &:= \begin{bmatrix}\Delta\boldsymbol{\mathcal{D}}_{k-1} & \Delta\boldsymbol{X}_k\end{bmatrix} \\ &= \bar{\boldsymbol{\mathcal{Q}}}_k\bar{\mathcal{R}}_k - \boldsymbol{\mathcal{X}}_k \\ &= \begin{bmatrix}\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{R}}_{k-1} - \boldsymbol{\mathcal{X}}_{k-1} & \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{R}}_{1:k-1,k} + \bar{\boldsymbol{Q}}_k\bar{R}_{kk} - \boldsymbol{X}_k\end{bmatrix}.\end{aligned} \tag{82}$$

The first element of (82) is taken care of by the induction hypothesis. As for the second element, we treat $\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{R}}_{1:k-1,k}$ and $\bar{\boldsymbol{Q}}_k\bar{R}_{kk}$ separately. There exists $\Delta\boldsymbol{R}_k$ such that

$$\bar{\mathcal{R}}_{1:k-1,k} = \bar{\mathcal{S}}_{1:k-1,k} + \bar{\mathcal{T}}_{1:k-1,k}\bar{S}_{kk} + \Delta\boldsymbol{R}_k, \tag{83}$$

$$\|\Delta\boldsymbol{R}_k\| \leq \mathcal{O}(\varepsilon)\left(\left\|\bar{\mathcal{S}}_{1:k-1,k}\right\| + \left\|\bar{\mathcal{T}}_{1:k-1,k}\right\|\left\|\bar{S}_{kk}\right\|\right) \leq \mathcal{O}(\varepsilon)\|\boldsymbol{X}_k\|, \tag{84}$$

by (52), (63), and Lemma 2. Combining (83), (84), and (22), it follows that

$$\begin{aligned}
\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{R}}_{1:k-1,k} &= \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\left(\bar{\mathcal{S}}_{1:k-1,k} + \bar{\mathcal{T}}_{1:k-1,k}\bar{S}_{kk} + \Delta\boldsymbol{R}_k\right) \\
&= \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\left(\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\boldsymbol{X}_k + \Delta\boldsymbol{S}_k + \bar{\mathcal{T}}_{1:k-1,k}\bar{S}_{kk} + \Delta\boldsymbol{R}_k\right) \\
&= \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\boldsymbol{X}_k + \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{T}}_{1:k-1,k}\bar{S}_{kk} + \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\left(\Delta\boldsymbol{S}_k + \Delta\boldsymbol{R}_k\right),
\end{aligned} \tag{85}$$

with

$$\left\|\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\left(\Delta\boldsymbol{S}_k + \Delta\boldsymbol{R}_k\right)\right\| \leq \mathcal{O}(\varepsilon)\|\boldsymbol{X}_k\|.$$

From line 12 of Algorithm 3, (48), and (64), we can write

$$\bar{R}_{kk} = \bar{T}_{kk}\bar{S}_{kk} + \Delta R_{kk}, \quad \|\Delta R_{kk}\| \leq \mathcal{O}(\varepsilon)\|\boldsymbol{X}_k\|. \tag{86}$$

Plugging (26), (27), (32), and (33) into (86), the term $\bar{\boldsymbol{Q}}_k\bar{R}_{kk}$ can be rewritten as

$$\begin{aligned}
\bar{\boldsymbol{Q}}_k\bar{R}_{kk} &= \bar{\boldsymbol{Q}}_k\left(\bar{T}_{kk}\bar{S}_{kk} + \Delta R_{kk}\right) \\
&= \bar{\boldsymbol{Q}}_k\bar{T}_{kk}\bar{S}_{kk} + \bar{\boldsymbol{Q}}_k\Delta R_{kk} \\
&= \left(\bar{\boldsymbol{W}}_k + \Delta\boldsymbol{G}_k^{(2)}\right)\bar{S}_{kk} + \bar{\boldsymbol{Q}}_k\Delta R_{kk} \\
&= \left(\bar{\boldsymbol{U}}_k - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{T}}_{1:k-1,k} + \Delta\boldsymbol{W}_k + \Delta\boldsymbol{G}_k^{(2)}\right)\bar{S}_{kk} + \bar{\boldsymbol{Q}}_k\Delta R_{kk} \\
&= \bar{\boldsymbol{U}}_k\bar{S}_{kk} - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{T}}_{1:k-1,k}\bar{S}_{kk} + \left(\Delta\boldsymbol{W}_k + \Delta\boldsymbol{G}_k^{(2)}\right)\bar{S}_{kk} + \bar{\boldsymbol{Q}}_k\Delta R_{kk} \\
&= \boldsymbol{X}_k - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\boldsymbol{X}_k + \Delta\boldsymbol{V}_k + \Delta\boldsymbol{G}_k^{(1)} - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\Delta S_k - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\bar{\mathcal{T}}_{1:k-1,k}\bar{S}_{kk} \\
&\quad + \left(\Delta\boldsymbol{W}_k + \Delta\boldsymbol{G}_k^{(2)}\right)\bar{S}_{kk} + \bar{\boldsymbol{Q}}_k\Delta R_{kk},
\end{aligned} \tag{87}$$

with

$$\left\|\Delta\boldsymbol{V}_k + \Delta\boldsymbol{G}_k^{(1)} - \bar{\boldsymbol{\mathcal{Q}}}_{k-1}\Delta S_k + \left(\Delta\boldsymbol{W}_k + \Delta\boldsymbol{G}_k^{(2)}\right)\bar{S}_{kk} + \bar{\boldsymbol{Q}}_k\Delta R_{kk}\right\| \leq \mathcal{O}(\varepsilon)\|\boldsymbol{X}_k\|.$$

Combining (85), (87), and the induction hypothesis, we can conclude the proof. □

Remark 1. Invoking Lemma 2 in the proof above is not circular logic, as it is valid to use with the induction hypothesis (residual in the $k-1$st step), and (63), (32), and (33) follow from standard rounding-error bounds together with Lemma 2.

The bound on the Cholesky residual follows directly from Theorems 3 and 4.

Corollary 5. *Assume that* $\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}) \leq \frac{1}{2}$ *and for all* $k \in \{1,\ldots,p\}$, $\bar{\boldsymbol{\mathcal{Q}}}_k$ *computed by Algorithm 3 satisfies*

$$\left\|I - \bar{\boldsymbol{\mathcal{Q}}}_k^T\bar{\boldsymbol{\mathcal{Q}}}_k\right\| \leq \frac{\mathcal{O}(\varepsilon)}{1 - \mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}_k)} \text{ and} \tag{88}$$

$$\boldsymbol{\mathcal{X}}_k + \Delta\boldsymbol{\mathcal{D}}_k = \bar{\boldsymbol{\mathcal{Q}}}_k\bar{\mathcal{R}}_k, \quad \|\Delta\boldsymbol{\mathcal{D}}_k\| \leq \mathcal{O}(\varepsilon)\|\boldsymbol{\mathcal{X}}_k\|. \tag{89}$$

Then for all $k \in \{1,\ldots,p\}$,

$$\bar{\mathcal{R}}_k^T\bar{\mathcal{R}}_k = \boldsymbol{\mathcal{X}}_k^T\boldsymbol{\mathcal{X}}_k + \Delta\mathcal{E}_k, \quad \|\Delta\mathcal{E}_k\| \leq \mathcal{O}(\varepsilon)\|\boldsymbol{\mathcal{X}}_k\|^2.$$

Proof. From (89) we can directly write

$$\bar{\mathcal{R}}_k^T \bar{\boldsymbol{\mathcal{Q}}}_k^T \bar{\boldsymbol{\mathcal{Q}}}_k \bar{\mathcal{R}}_k = \boldsymbol{\mathcal{X}}_k^T \boldsymbol{\mathcal{X}}_k + \Delta M_k, \quad \|\Delta M_k\| \leq \mathcal{O}(\varepsilon) \|\boldsymbol{\mathcal{X}}_k\|^2,$$

where $\Delta M_k = \boldsymbol{\mathcal{X}}_k^T \Delta \boldsymbol{\mathcal{D}}_k + (\Delta \boldsymbol{\mathcal{D}}_k)^T \boldsymbol{\mathcal{X}}_k + (\Delta \boldsymbol{\mathcal{D}}_k)^2$. Rearranging terms and applying the assumption (88) yields

$$\begin{aligned} \bar{\mathcal{R}}_k^T \bar{\mathcal{R}}_k &= \boldsymbol{\mathcal{X}}_k^T \boldsymbol{\mathcal{X}}_k + \underbrace{\Delta M_k + \bar{\mathcal{R}}_k^T (I - \bar{\boldsymbol{\mathcal{Q}}}_k^T \bar{\boldsymbol{\mathcal{Q}}}_k) \bar{\mathcal{R}}_k}_{=:\Delta \mathcal{E}_k}, \\ \|\Delta \mathcal{E}_k\| &\leq \mathcal{O}(\varepsilon) \left(\|\boldsymbol{\mathcal{X}}_k\|^2 + \left\|\bar{\mathcal{R}}_k\right\|^2 \right). \end{aligned} \tag{90}$$

Bounding $\left\|\bar{\mathcal{R}}_k\right\|$ follows by multiplying (89) by $\bar{\boldsymbol{\mathcal{Q}}}_k^T$ on the left and rearranging terms to arrive at

$$\bar{\mathcal{R}}_k = (I - \bar{\boldsymbol{\mathcal{Q}}}_k^T \bar{\boldsymbol{\mathcal{Q}}}_k) \bar{\mathcal{R}}_k + \bar{\boldsymbol{\mathcal{Q}}}_k^T \boldsymbol{\mathcal{X}}_k + \bar{\boldsymbol{\mathcal{Q}}}_k^T \Delta \boldsymbol{\mathcal{D}}_k.$$

Under the assumptions of this lemma and by (21), we find

$$\left\|\bar{\mathcal{R}}_k\right\| \leq \frac{1 + \mathcal{O}(\varepsilon)}{1 - \mathcal{O}(\varepsilon)} \|\boldsymbol{\mathcal{X}}_k\| \leq \mathcal{O}(1) \|\boldsymbol{\mathcal{X}}_k\|,$$

which, substituted back into (90), completes the proof. □

Remark 2. A version of Algorithm 3 has been independently developed and its performance studied in [29, Figure 4(b)]; the authors there refer to it as `BCGS-PIP2`. We became aware of [29] as we were finishing this manuscript. The authors provide a high-level discussion of the stability analysis of `BCGS-PIP2`, but they erroneously conflate `BCGS-PIPI+` (Algorithm 3) with `BCGS-PIP+` (Algorithm 2); furthermore, it is unclear what `IO` initializes their `BCGS-PIP2`. Indeed, as our detailed analysis demonstrates, interchanging the for-loops has a nontrivial effect on attainable guarantees for LOO, as well as on conditions for the first `IO`. In particular, `BCGS-PIPI+` requires a stronger `IO` than `BCGS-PIP+` to maintain LOO, because the first block vector is not reorthogonalized. Furthermore, neither their analysis nor ours extends to the "two-stage" algorithm they present, which is essentially a hybrid of `BCGS-PIP+` and `BCGS-PIPI+` and allows for a larger block size in the reorthogonalization step. We leave the stability analysis of this hybrid algorithm to future work.

Remark 3. An immediate consequence of the analysis in this section is that the proven bounds hold trivially for block size $s = 1$, and unfortunately the restriction $\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}) \leq \frac{1}{2}$ cannot be alleviated; indeed, the size of s only affects the (hidden) constants in $\mathcal{O}(\varepsilon)$. As long as $s \ll m$ (i.e., several orders of magnitude smaller than m), we do not expect it to affect the bounds. Indeed, in Section 4, we look at examples with $s = 2$ and $s = 10$ and observe no dependence on block size. In high-performance implementations, practical choices for s depend on the application and hardware but typically remain small for a large number of unknowns.

Remark 4. Communication-avoiding Krylov subspace methods like s-step GMRES [17] typically use a block Gram-Schmidt orthogonalization scheme and in each outer iteration generate a block vector of the form $\begin{bmatrix} p_0(A)\boldsymbol{v} & p_1(A)\boldsymbol{v} & \cdots & p_{s-1}(A)\boldsymbol{v} \end{bmatrix}$, where A is a linear operator, $\boldsymbol{v}$ is some starting vector, and p_i, $i \in \{0, \ldots, s-1\}$, are polynomials of degree i, respectively. The methods considered in this manuscript can be used as block skeletons in s-step GMRES, but the analysis of its backward stability is more complicated than simply applying our results, due to the dual role that s plays. In the present manuscript, s denotes a block partitioning of a *fixed* matrix $\boldsymbol{\mathcal{X}}$ and thus does not affect $\kappa(\boldsymbol{\mathcal{X}})$. Conversely, in s-step GMRES, s determines not only the size of block vectors but also the conditioning of each block (and therefore the entire basis), as each is computed from powers of A. Consequently the backward stability of s-step GMRES *is* sensitive to the choice of s; for a complete analysis, see [10], especially Figure 1 therein. In particular, note that employing `BCGS-PIPI+` for orthogonalization in s-step GMRES may result in a limited backward error for certain examples, as demonstrated in Figures 8 and 9 of [10].

3. Mixed-precision variants

It is possible to use multiple precisions in the implementations of BCGS-PIP+ and BCGS-PIPI+ without affecting the validity of general results from Section 2. We provide pseudocode for two-precision versions of each as Algorithms 5 and Algorithm 6, respectively. In the pseudocode, ε_ℓ denotes computing and storing a quantity in the low precision with the associated unit roundoff, and ε_h likewise for high precision. In particular, $\varepsilon_\ell \geq \varepsilon_h$.

One motivation for using multiple precisions is to attempt to eliminate the restriction on $\kappa(\boldsymbol{\mathcal{X}})$ for the stability bounds and thereby extend stability guarantees for higher condition numbers; see related work in, e.g., [21, 22]. In a uniform working precision, both BCGS-PIP+ and BCGS-PIPI+ require $\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}) \leq 1$, which practically translates into $\kappa(\boldsymbol{\mathcal{X}}) \leq \mathcal{O}\left(10^4\right)$ or $\kappa(\boldsymbol{\mathcal{X}}) \leq \mathcal{O}\left(10^8\right)$, for single or double precisions, respectively. It is natural to consider whether using double the working precision in some parts of the algorithm might alleviate this restriction. At the same time, we do not want to increase communication cost by transmitting high-precision data. Therefore Algorithms 5-6 are formulated so that the data and solutions $\boldsymbol{\mathcal{Q}}$ and $\mathcal{R}$ are stored in low precision, while high precision is used for the local (i.e., on-node) computation of operations like $\boldsymbol{V}^T\boldsymbol{V}$, Cholesky factorization, and inverting Cholesky factors, with the motivation being that the Pythagorean step, based on Cholesky factorization, is ultimately responsible for the condition number restriction. Note that the inner products in line 3 of BCGS-PIP and lines 3 and 7 of BCGS-PIPI+ are now split across two steps to handle different precisions. However, we still regard these a single sync point, as the synchronization itself just involves the movement of memory, which can of course be handled in multiple precisions.

At the same time, doubling the precision implies doubling the computational cost and (potentially) the amount of data moved. This overhead is highly dependent on the problem size and it may be negligible in particular cases, such as latency-bound regimes[3] and when both precisions are implemented in hardware. When high precision computations are performed locally, such as in line 6 in Algorithm 3, the extra overhead may very well be insignificant.

Algorithm 4 $[\boldsymbol{\mathcal{Q}}, \mathcal{R}] = \text{BCGS-PIP}^{\text{MP}}(\boldsymbol{\mathcal{X}}, \texttt{IO})$

1: $[\boldsymbol{Q}_1, R_{11}] = \texttt{IO}(\boldsymbol{X}_1)$ ▷ compute and return in ε_ℓ
2: **for** $k = 2, \ldots, p$ **do**
3: $\quad \mathcal{R}_{1:k-1,k} = \boldsymbol{\mathcal{Q}}_{k-1}^T \boldsymbol{X}_k$ ▷ compute and return in ε_ℓ
4: $\quad P_k = \boldsymbol{X}_k^T \boldsymbol{X}_k$ ▷ compute and return in ε_h
5: $\quad R_{kk} = \texttt{chol}\left(P_k - \mathcal{R}_{1:k-1,k}^T \mathcal{R}_{1:k-1,k}\right)$ ▷ compute and return in ε_h; cast to ε_ℓ after line 7
6: $\quad \boldsymbol{V}_k = \boldsymbol{X}_k - \boldsymbol{\mathcal{Q}}_{k-1}\mathcal{R}_{1:k-1,k}$ ▷ compute and return in ε_ℓ
7: $\quad \boldsymbol{Q}_k = \boldsymbol{V}_k R_{kk}^{-1}$ ▷ compute each $s \times s$ block locally in ε_h; return in ε_ℓ
8: **end for**
9: **return** $\boldsymbol{\mathcal{Q}} = [\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_p]$, $\mathcal{R} = (R_{ij})$

Algorithm 5 $[\boldsymbol{\mathcal{Q}}, \mathcal{R}] = \text{BCGS-PIP+}^{\text{MP}}(\boldsymbol{\mathcal{X}}, \texttt{IO})$

1: $[\boldsymbol{\mathcal{U}}, \mathcal{S}] = \text{BCGS-PIP}^{\text{MP}}(\boldsymbol{\mathcal{X}}, \texttt{IO})$ ▷ computed in mixed; returned in ε_ℓ
2: $[\boldsymbol{\mathcal{Q}}, \mathcal{T}] = \text{BCGS-PIP}^{\text{MP}}(\boldsymbol{\mathcal{U}}, \texttt{IO})$ ▷ computed in mixed; returned in ε_ℓ
3: $\mathcal{R} = \mathcal{T}\mathcal{S}$; ▷ compute and return in ε_ℓ
4: **return** $\boldsymbol{\mathcal{Q}} = [\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_p]$, $\mathcal{R} = (R_{ij})$

Unfortunately, our intuitive proposals for mixed-precision variants do not achieve the desired stability for either BCGS-PIP+$^{\text{MP}}$ or BCGS-PIPI+$^{\text{MP}}$, which we demonstrate in the following sections.

3.1. Two-precision BCGS-PIP and BCGS-PIP+

With Theorem 1, we already have generalized bounds on the LOO that will also hold for BCGS-PIP+$^{\text{MP}}$. To determine what the constants $\delta_{US}, \omega_U, \delta_{QT}, \omega_Q$ and δ_{TS} look like in the two-precision case, we

[3] That is, where *latency*, or the time it takes for memory to travel from one node to another across a network or between levels of cache, dominates the runtime of an algorithm.

Algorithm 6 $[\mathcal{Q}, \mathcal{R}] = \texttt{BCGS-PIPI+}^{\texttt{MP}}(\mathcal{X}, \texttt{IO})$

1: $[\boldsymbol{Q}_1, R_{11}] = \texttt{IO}(\boldsymbol{X}_1)$ ▷ compute and return in ε_ℓ
2: **for** $k = 2, \ldots, p$ **do**
3: $\mathcal{S}_{1:k-1,k} = \mathcal{Q}_{k-1}^T \boldsymbol{X}_k$ ▷ compute and return in ε_ℓ
4: $\Omega_k = \boldsymbol{X}_k^T \boldsymbol{X}_k$ ▷ compute and return in ε_h
5: $S_{kk} = \texttt{chol}\left(\Omega_k - \mathcal{S}_{1:k-1,k}^T \mathcal{S}_{1:k-1,k}\right)$ ▷ compute and return in ε_h
6: $\boldsymbol{V}_k = \boldsymbol{X}_k - \mathcal{Q}_{k-1} \mathcal{S}_{1:k-1,k}$ ▷ compute and return in ε_ℓ
7: $\boldsymbol{U}_k = \boldsymbol{V}_k S_{kk}^{-1}$ ▷ compute each $s \times s$ block locally in ε_h; return in ε_ℓ
8: $\mathcal{T}_{1:k-1,k} = \mathcal{Q}_{k-1}^T \boldsymbol{U}_k$ ▷ compute and return in ε_ℓ
9: $P_k = \boldsymbol{U}_k^T \boldsymbol{U}_k$ ▷ compute and return in ε_h
10: $T_{kk} = \texttt{chol}\left(P_k - \mathcal{T}_{1:k-1,k}^T \mathcal{T}_{1:k-1,k}\right)$ ▷ compute and return in ε_h
11: $\boldsymbol{W}_k = \boldsymbol{U}_k - \mathcal{Q}_{k-1} \mathcal{T}_{1:k-1,k}$ ▷ compute and return in ε_ℓ
12: $\boldsymbol{Q}_k = \boldsymbol{W}_k T_{kk}^{-1}$ ▷ compute each $s \times s$ block locally in ε_h; return in ε_ℓ
13: $\mathcal{R}_{1:k-1,k} = \mathcal{S}_{1:k-1,k} + \mathcal{T}_{1:k-1,k} S_{kk}$ ▷ compute and return in ε_ℓ
14: $R_{kk} = T_{kk} S_{kk}$ ▷ compute in ε_h; return in ε_ℓ
15: **end for**
16: **return** $\mathcal{Q} = [\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_p]$, $\mathcal{R} = (R_{ij})$

need to obtain generalized bounds for BCGS-PIP that do not explicitly rely on a uniform precision but rather express bounds in terms of different constants with subscripts corresponding to the source of error. The following two lemmas take care of this; they generalize [8, Theorems 3.1 and 3.2], respectively. Proofs of Lemma 3 and Lemma 4 are available in Appendix A.1 and A.2, respectively.

Lemma 3. *Let* $\mathcal{X} \in \mathbb{R}^{m \times ps}$ *and* $[\bar{\mathcal{Q}}, \bar{\mathcal{R}}] = \textit{BCGS-PIP}(\mathcal{X}, \texttt{IO})$*, for some* IO*. Suppose* $\xi\kappa^2(\mathcal{X}) \leq \frac{1}{2}$ *and* $\rho\kappa^2(\mathcal{X}) \leq \frac{1}{4}$*, for constants* $\xi, \rho \in (0,1)$*. Furthermore, assume that*

$$\bar{\mathcal{R}}^T \bar{\mathcal{R}} = \mathcal{X}^T \mathcal{X} + \Delta\mathcal{E}, \quad \|\Delta\mathcal{E}\| \leq \xi \|\mathcal{X}\|^2, \text{ and} \tag{91}$$

$$\bar{\mathcal{Q}} \bar{\mathcal{R}} = \mathcal{X} + \Delta\mathcal{D}, \quad \|\Delta\mathcal{D}\| \leq \rho(\|\mathcal{X}\| + \|\bar{\mathcal{Q}}\| \, \|\bar{\mathcal{R}}\|). \tag{92}$$

Then

$$\left\|I - \bar{\mathcal{Q}}^T \bar{\mathcal{Q}}\right\| \leq \frac{(\xi + 11 \cdot \rho)\kappa^2(\mathcal{X})}{1 - \xi\kappa^2(\mathcal{X})}; \tag{93}$$

$$\|\bar{\mathcal{Q}}\| \leq 3; \text{ and} \tag{94}$$

$$\|\Delta\mathcal{D}\| \leq 6 \cdot \rho \|\mathcal{X}\|. \tag{95}$$

For the following, we assume that for each $k \in \{2, \ldots, p\}$, $\bar{\mathcal{R}}_{k-1}$ and $\bar{\mathcal{Q}}_{k-1}$ are computed by BCGS-PIP and that there exist $\xi_{k-1}, \rho_{k-1} \in (0,1)$ such that

$$\begin{aligned} \bar{\mathcal{R}}_{k-1}^T \bar{\mathcal{R}}_{k-1} &= \mathcal{X}_{k-1}^T \mathcal{X}_{k-1} + \Delta\mathcal{E}_{k-1}, \\ \|\Delta\mathcal{E}_{k-1}\| &\leq \xi_{k-1} \|\mathcal{X}_{k-1}\|^2; \end{aligned} \tag{96}$$

and

$$\begin{aligned} \bar{\mathcal{Q}}_{k-1} \bar{\mathcal{R}}_{k-1} &= \mathcal{X}_{k-1} + \Delta\mathcal{D}_{k-1}, \\ \|\Delta\mathcal{D}_{k-1}\| &\leq \rho_{k-1}(\|\mathcal{X}_{k-1}\| + \|\bar{\mathcal{Q}}_{k-1}\| \, \|\bar{\mathcal{R}}_{k-1}\|). \end{aligned} \tag{97}$$

We can furthermore write the following for intermediate quantities computed by BCGS-PIP, where we omit the explicit dependence on k for readability and each $\delta_* \in (0,1)$:

$$\bar{\mathcal{R}}_{1:k-1,k} = \bar{\mathcal{Q}}_{k-1}^T \boldsymbol{X}_k + \Delta\boldsymbol{R}_k, \quad \|\Delta\boldsymbol{R}_k\| \leq \delta_{Q^T X} \|\boldsymbol{X}_k\|; \tag{98}$$

$$\bar{P}_k = \boldsymbol{X}_k^T \boldsymbol{X}_k + \Delta P_k, \quad \|\Delta P_k\| \leq \delta_{X^T X} \|\boldsymbol{X}_k\|^2; \tag{99}$$

$$\bar{R}_{kk}^T \bar{R}_{kk} = \bar{P}_k - \bar{\mathcal{R}}_{1:k-1,k}^T \bar{\mathcal{R}}_{1:k-1,k} + \Delta F_k + \Delta C_k, \tag{100}$$

$$\|\Delta F_k\| \leq \delta_{R^T R} \|\boldsymbol{X}_k\|^2, \quad \|\Delta C_k\| \leq \delta_{\texttt{chol}} \|\boldsymbol{X}_k\|^2; \tag{101}$$

$$\bar{\boldsymbol{V}}_k = \boldsymbol{X}_k - \bar{\mathcal{Q}}_{k-1} \bar{\mathcal{R}}_{1:k-1,k} + \Delta\boldsymbol{V}_k, \quad \|\Delta\boldsymbol{V}_k\| \leq \delta_{QR} \|\boldsymbol{X}_k\|; \text{ and} \tag{102}$$

$$\bar{\boldsymbol{Q}}_k \bar{R}_k = \bar{\boldsymbol{V}}_k + \Delta\boldsymbol{G}_k, \quad \|\Delta\boldsymbol{G}_k\| \leq \delta_Q \|\bar{\boldsymbol{Q}}_k\| \, \|\bar{R}_k\|. \tag{103}$$

Throughout (98)–(103), we have dropped quadratic error terms and applied Lemma 3 to simplify constants (similarly to what we have done in (22)–(33); the contribution of $\left\|\bar{\mathcal{Q}}_{k-1}\right\|$ is absorbed into the constant). Furthermore, ΔF_k denotes the floating-point error from the subtraction of the product $\bar{\mathcal{R}}_{1:k-1,k}^T \bar{\mathcal{R}}_{1:k-1,k}$ from $\bar{P}_k$, and ΔC_k denotes the error from the Cholesky factorization of that result; note that, similar to the argument in the proof of [8, Theorem 3.2], $\bar{P}_k - \bar{\mathcal{R}}_{1:k-1,k}^T \bar{\mathcal{R}}_{1:k-1,k}$ should be symmetric positive definite.

Lemma 4. *Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m \times ps}$ and fix $k \in \{2, \ldots, p\}$. Assume that (96)–(103) are satisfied. Then the following hold:*

$$\bar{\mathcal{R}}_k^T \bar{\mathcal{R}}_k = \boldsymbol{\mathcal{X}}_k^T \boldsymbol{\mathcal{X}}_k + \Delta \mathcal{E}_k, \quad \|\Delta \mathcal{E}_k\| \leq \xi_k \|\boldsymbol{\mathcal{X}}_k\|^2, \tag{104}$$

and

$$\bar{\mathcal{Q}}_k \bar{\mathcal{R}}_k = \boldsymbol{\mathcal{X}}_k + \Delta \boldsymbol{\mathcal{D}}_k, \quad \|\Delta \boldsymbol{\mathcal{D}}_k\| \leq \rho_k (\|\boldsymbol{\mathcal{X}}_k\| + \left\|\bar{\mathcal{Q}}_k\right\| \left\|\bar{\mathcal{R}}_k\right\|), \tag{105}$$

where

$$\xi_k = \xi_{k-1} + 12 \cdot \rho_{k-1} + 2\sqrt{2} \cdot \delta_{Q^T X} + \delta_{X^T X} + \delta_{chol} + \delta_{R^T R}$$

and $\rho_k = 6 \cdot \rho_{k-1} + \delta_{QR} + \delta_Q$.

On their own, Lemmas 3 and 4 do not complete the analysis; they describe the relationship between bounds from one iteration to the next without specifying the precision. Set

$$\rho_{\max} := \max_{k \in \{1, \ldots, p-1\}} \rho_k \quad \text{and} \quad \xi_{\max} := \max_{k \in \{1, \ldots, p-1\}} \xi_k. \tag{106}$$

In the next theorem, we combine these lemmas to obtain bounds on BCGS-PIP$^{\text{MP}}$. A proof of the theorem is available in Appendix A.3.

Theorem 3. *Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m \times ps}$ such that*

$$(\rho_{\max} + \rho_p)\kappa^2(\boldsymbol{\mathcal{X}}) \leq \frac{1}{4} \quad \text{and} \quad (\xi_{\max} + \xi_p)\kappa^2(\boldsymbol{\mathcal{X}}) \leq \frac{1}{2}, \tag{107}$$

for ρ_p and ξ_p defined as in Lemma 4. Suppose $[\bar{\mathcal{Q}}, \bar{\mathcal{R}}] =$ BCGS-PIP$^{\text{MP}}(\boldsymbol{\mathcal{X}}, I\!O)$, where for all $\boldsymbol{X} \in \mathbb{R}^{m \times s}$ with $\kappa(\boldsymbol{X}) \leq \kappa(\boldsymbol{\mathcal{X}})$, $[\bar{\boldsymbol{Q}}, \bar{R}] = I\!O(\boldsymbol{X})$ satisfy

$$\bar{R}^T \bar{R} = \boldsymbol{X}^T \boldsymbol{X} + \Delta E, \quad \|\Delta E\| \leq \mathcal{O}(\varepsilon_\ell) \|\boldsymbol{X}\|^2 \text{ and} \tag{108}$$

$$\bar{\boldsymbol{Q}} \bar{R} = \boldsymbol{X} + \Delta \boldsymbol{D}, \quad \|\Delta \boldsymbol{D}\| \leq \mathcal{O}(\varepsilon_\ell) \|\boldsymbol{X}\|. \tag{109}$$

Then for all $k \in \{1, \ldots, p\}$,

$$\bar{\mathcal{R}}_k^T \bar{\mathcal{R}}_k = \boldsymbol{\mathcal{X}}_k^T \boldsymbol{\mathcal{X}}_k + \Delta \mathcal{E}_k, \quad \|\Delta \mathcal{E}_k\| \leq \mathcal{O}(\varepsilon_\ell) \|\boldsymbol{\mathcal{X}}_k\|^2; \tag{110}$$

$$\bar{\mathcal{Q}}_k \bar{\mathcal{R}}_k = \boldsymbol{\mathcal{X}}_k + \boldsymbol{\mathcal{D}}_k, \quad \|\boldsymbol{\mathcal{D}}_k\| \leq \mathcal{O}(\varepsilon_\ell) \|\boldsymbol{\mathcal{X}}_k\|; \text{ and} \tag{111}$$

$$\left\| I - \bar{\mathcal{Q}}_k^T \bar{\mathcal{Q}}_k \right\| \leq \mathcal{O}(\varepsilon_\ell) \kappa^2(\boldsymbol{\mathcal{X}}_k). \tag{112}$$

Applying Theorem 3 twice to the dual-precision BCGS-PIP+$^{\text{MP}}$ shows that

$$\delta_{US}, \omega_U, \delta_{QT}, \omega_Q, \delta_{TS} = \mathcal{O}(\varepsilon_\ell), \tag{113}$$

where the constants stem from (8)–(11). The following corollary is a direct consequence of (113) and Theorem 1 and demonstrates that BCGS-PIP+$^{\text{MP}}$ is dominated by $\mathcal{O}(\varepsilon_\ell)$ error.

Corollary 6. *Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m \times ps}$ such that $\delta_{\max}\kappa^2(\boldsymbol{\mathcal{X}}) \leq \frac{1}{4}$, where $\delta_{\max} := \max\{\delta_{US}, \omega_U, \delta_{QT}, \omega_Q\}$. Suppose $[\bar{\mathcal{Q}}, \bar{\mathcal{R}}] =$ BCGS-PIP$^{\text{MP}}(\boldsymbol{\mathcal{X}}, I\!O)$, where for all $\boldsymbol{X} \in \mathbb{R}^{m \times s}$ with $\kappa(\boldsymbol{X}) \leq \kappa(\boldsymbol{\mathcal{X}})$, $[\bar{\boldsymbol{Q}}, \bar{R}] = I\!O(\boldsymbol{X})$ satisfy*

$$\bar{R}^T \bar{R} = \boldsymbol{X}^T \boldsymbol{X} + \Delta E, \quad \|\Delta E\| \leq \mathcal{O}(\varepsilon_\ell) \|\boldsymbol{X}\|^2 \text{ and}$$

$$\bar{\boldsymbol{Q}} \bar{R} = \boldsymbol{X} + \Delta \boldsymbol{D}, \quad \|\Delta \boldsymbol{D}\| \leq \mathcal{O}(\varepsilon_\ell) \|\boldsymbol{X}\|.$$

Then for all $k \in \{1, \ldots, p\}$,

$$\bar{\mathcal{R}}_k^T \bar{\mathcal{R}}_k = \boldsymbol{\mathcal{X}}_k^T \boldsymbol{\mathcal{X}}_k + \Delta \mathcal{E}_k, \quad \|\Delta \mathcal{E}_k\| \leq \mathcal{O}(\varepsilon_\ell) \|\boldsymbol{\mathcal{X}}_k\|^2;$$

$$\bar{\mathcal{Q}}_k \bar{\mathcal{R}}_k = \boldsymbol{\mathcal{X}}_k + \boldsymbol{\mathcal{D}}_k, \quad \|\boldsymbol{\mathcal{D}}_k\| \leq \mathcal{O}(\varepsilon_\ell) \|\boldsymbol{\mathcal{X}}_k\|; \text{ and}$$

$$\left\| I - \bar{\mathcal{Q}}_k^T \bar{\mathcal{Q}}_k \right\| \leq \mathcal{O}(\varepsilon_\ell).$$

3.2. Two-precision BCGS-PIPI+

By standard rounding-error analysis we can show that for (22)–(33) in BCGS-PIPI+[MP],

$$\delta_{Q^TX}, \delta_{QS}, \delta_{Q^TU}, \delta_{QT} = \mathcal{O}(\varepsilon_\ell)$$

and

$$\delta_{X^TX}, \delta_{S^TS}, \delta_{\texttt{chol}_1}, \delta_U, \delta_{U^TU}, \delta_{T^TT}, \delta_{\texttt{chol}_2}, \delta_Q = \mathcal{O}(\varepsilon_h).$$

The following corollary follows directly from these constants and the development in Section 2.2, in a similar manner as Corollaries 3, 4, and 5. Consequently, we must conclude that BCGS-PIPI+[MP] is also dominated by $\mathcal{O}(\varepsilon_\ell)$ error.

Corollary 7. *Assume that* $\mathcal{O}(\varepsilon_\ell)\kappa^2(\boldsymbol{\mathcal{X}}) \leq \frac{1}{2}$ *and that for all* $\boldsymbol{X} \in \mathbb{R}^{m\times s}$ *with* $\kappa(\boldsymbol{X}) \leq \kappa(\boldsymbol{\mathcal{X}})$, $[\bar{\boldsymbol{Q}}, \bar{R}] = \texttt{IO}(\boldsymbol{X})$ *satisfy*

$$\begin{aligned} \bar{R}^T\bar{R} &= \boldsymbol{X}^T\boldsymbol{X} + \Delta E, \quad \|\Delta E\| \leq \mathcal{O}(\varepsilon_\ell)\|\boldsymbol{X}\|^2 \\ \bar{\boldsymbol{Q}}\bar{R} &= \boldsymbol{X} + \Delta\boldsymbol{D}, \quad \|\Delta\boldsymbol{D}\| \leq \mathcal{O}(\varepsilon_\ell)\|\boldsymbol{X}\| \quad \textit{and} \\ \|I - \bar{\boldsymbol{Q}}^T\bar{\boldsymbol{Q}}\| &\leq \frac{\mathcal{O}(\varepsilon_\ell)}{1 - \mathcal{O}(\varepsilon_\ell)\kappa^2(\boldsymbol{X})}. \end{aligned}$$

Then for $[\bar{\boldsymbol{\mathcal{Q}}}, \mathcal{R}] =$ BCGS-PIPI+[MP]$(\boldsymbol{\mathcal{X}}, \texttt{IO})$, *the following hold for all* $k \in \{1, \ldots, p\}$:

$$\begin{aligned} \bar{\mathcal{R}}_k^T\bar{\mathcal{R}}_k &= \boldsymbol{\mathcal{X}}_k^T\boldsymbol{\mathcal{X}}_k + \Delta\mathcal{E}_k, \quad \|\Delta\mathcal{E}_k\| \leq \mathcal{O}(\varepsilon_\ell)\|\boldsymbol{\mathcal{X}}_k\|^2; \\ \bar{\boldsymbol{\mathcal{Q}}}_k\bar{\mathcal{R}}_k &= \boldsymbol{\mathcal{X}}_k + \Delta\boldsymbol{\mathcal{D}}_k, \quad \|\Delta\boldsymbol{\mathcal{D}}_k\| \leq \mathcal{O}(\varepsilon_\ell)\|\boldsymbol{\mathcal{X}}_k\| \\ \left\|I - \bar{\boldsymbol{\mathcal{Q}}}_k^T\bar{\boldsymbol{\mathcal{Q}}}_k\right\| &\leq \frac{\mathcal{O}(\varepsilon_\ell)}{1 - \mathcal{O}(\varepsilon_\ell)\kappa^2(\boldsymbol{\mathcal{X}}_k)} \leq \mathcal{O}(\varepsilon_\ell). \end{aligned}$$

Remark 5. Our analytical framework is more general than for an arbitrary uniform precision or even two precisions, which we have focused on. Indeed, it is not hard to see that regardless of the number of precisions used in these algorithms, the lowest precision will dominate the LOO and residual bounds. At the same time, numerical experiments in the next section demonstrate that our bounds are rather pessimistic for practical scenarios. More nuanced bounds remain a topic for future work.

4. Numerical experiments

4.1. BlockStab: a workflow for BGS comparisons

We make use of an updated version of the BlockStab code suite [19][4]. In this version, multiprecision implementations have been added for both the Advanpix Multiprecision Computing Toolbox[5] and the Symbolic Math Toolbox[6], along with additional low-sync versions of various block Gram-Schmidt routines, which are analyzed further in [7]. The procedure for setting up algorithm comparisons is now streamlined to avoid redundant runs and to allow for different choices of Cholesky factorization implementations. Finally, a TeX report is auto-generated with a timestamp, which facilitates sharing experiments with collaborators.

4.2. Comparisons among newly proposed variants

We perform several numerical experiments to study the stability of BCGS-PIP, BCGS-PIP+, and BCGS-PIPI+, using four classes of matrices available in BlockStab, namely `default`, `glued`, `monomial`, and `piled`. Each matrix class is created using dimensional inputs m, p, s, where m denotes the number of rows, p denotes the number of block vectors, and s denotes the number of columns in each block vector. Descriptions of the matrix classes are as follows:

[4] https://github.com/katlund/BlockStab/releases/tag/v2.1.2024
[5] Version 4.8.3.14460. https://www.advanpix.com/
[6] Version depends on MATLAB version. https://mathworks.com/products/symbolic.html

- `default`: built as $\boldsymbol{\mathcal{X}}_t = \boldsymbol{\mathcal{U}}\Sigma_t\boldsymbol{\mathcal{V}}^T$, where $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{m\times ps}$ is orthonormal, $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{ps\times ps}$ is unitary, and $\Sigma_t \in \mathbb{R}^{ps\times ps}$ is diagonal with entries drawn from the logarithmic interval $10^{[-t,0]}$.
- `glued`: first introduced in [23] and constructed to cause classical Gram-Schmidt to break down. The matrix is initialized with a `default` matrix $\boldsymbol{\mathcal{X}}_t$; then each $m \times s$ block vector is multiplied by $\Sigma_r\widetilde{V}$, where $\Sigma_r, \widetilde{V} \in \mathbb{R}^{s\times s}$, and $\widetilde{V}$ is unitary.
- `monomial`: consists of r block vectors $\boldsymbol{X}_k = \begin{bmatrix}\boldsymbol{v}_k & A\boldsymbol{v}_k & \cdots & A^{t-1}\boldsymbol{v}_k\end{bmatrix}$, $k \in \{1,\dots,r\}$, where each $\boldsymbol{v}_k$ is randomly generated from the uniform distribution and normalized, and A is an $m \times m$ diagonal operator having evenly distributed eigenvalues in $(0.1, 10)$. (Note that r is not necessarily equal to p, and likewise $t \neq s$; however $rt = ps$. Varying r and t allows for generating matrices with different condition numbers.)
- `piled`: formed as $\boldsymbol{\mathcal{X}} = \begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2 & \cdots & \boldsymbol{X}_p\end{bmatrix}$, where $\boldsymbol{X}_1$ is a `default` matrix with a small condition number and for $k \in \{2,\dots,p\}$, $\boldsymbol{X}_k = \boldsymbol{X}_{k-1} + \boldsymbol{Z}_k$, where each $\boldsymbol{Z}_k$ is also a `default` matrix with the same condition number for all k. Toggling the condition numbers of $\boldsymbol{X}_1$ and $\{\boldsymbol{Z}_k\}_{k=2}^p$ controls the overall conditioning of the test matrix.

We set $m = 100$, $p = 10$, and $s = 2$ for `glued` and `default` matrices; $m = 2000$, $p = 120$, $s = 10$ for `monomial` matrices; and $m = 100$, $p = 10$, $s = 5$ for `piled` matrices.

To illustrate the numerical behavior of the algorithms from Sections 2 and 3, we plot the LOO (1) and relative Cholesky residual (i.e., (3) divided by $\|\boldsymbol{\mathcal{X}}\|^2$) of each algorithm versus the condition number of the matrix; we refer to these plots as κ-plots, as they are relative to the changing condition number $\kappa(\boldsymbol{\mathcal{X}})$. To observe the effects of the choice of `IO`, we use `HouseQR` and `CholQR`, where a variant of Cholesky factorization is used to bypass MATLAB's `chol` protocol for halting the computation when a matrix loses numerical positive definiteness. One can regard `HouseQR` as a placeholder for `TSQR`, as their numerical behavior is similar, even though the communication properties would differ in practical distributed computing settings.

Double precision ($\varepsilon = 2^{-53} \approx 10^{-16}$) is used for uniform-precision methods. Advanpix is used to simulate quadruple precision ($\varepsilon_h = 2^{-113} \approx 10^{-32}$) in multiprecision algorithms, while the low precision is set to double ($\varepsilon_\ell = \varepsilon$).[7]

All numerical tests are run in MATLAB 2022a. Every test is run on a Lenovo ThinkPad E15 Gen 2 with 8GB memory and AMD Ryzen 5 4500U CPU with Radeon Graphics. The CPU has 6 cores with 384KiB L1 cache, and 3MiB L2 cache at a clockrate of 1 GHz, as well as 8MiB of shared L3 cache. The script
`test_bcgs_pip_reortho.m` can be used for regenerating all plots in this section.

Figure 1 illustrates the stability of reorthogonalized variants compared to BCGS-PIP in uniform precision. We see that BCGS-PIP follows a $\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}})$ LOO trend until $\kappa(\boldsymbol{\mathcal{X}}) \approx \frac{1}{\sqrt{\varepsilon}} \approx 10^8$, as expected. Meanwhile the reorthogonalized variants reach nearly 10^{-16} LOO, regardless of the choice of `IO`, until $\kappa(\boldsymbol{\mathcal{X}}) \approx 10^8$. After this point, we observe breakdowns and an increasing loss of orthogonality for all methods. Missing points for large condition numbers are due to `NaN` being computed during the Cholesky factorization, resulting from operations like $\frac{\texttt{Inf}}{\texttt{Inf}}$ or $\frac{\texttt{Inf}}{0}$.

The effect of the choice of `IO` for uniform-precision methods can be seen in Figure 2. As there is no assumption on the LOO of the `IO` in Corollary 1, we observe that `CholQR` works well for BCGS-PIP+. On the other hand, Theorem 2 places a LOO restriction on the `IO` for BCGS-PIPI+. Indeed, `CholQR` does not satisfy the assumption, so the proven LOO and residual bounds are not guaranteed to hold, and we can see BCGS-PIPI+ $\circ$ `CholQR` failing to reach around 10^{-16} LOO even for small condition numbers. Again, both BCGS-PIP+ and BCGS-PIPI+ exhibit an increasing LOO and residual after $\kappa(\boldsymbol{\mathcal{X}}) > 10^8$.

Figure 3 compares two-precision with uniform-precision algorithms. For `default` matrices with or without mixed precision, reorthogonalization keeps the LOO near 10^{-16} as long as $\kappa(\boldsymbol{\mathcal{X}}) \leq 10^8$ and below $\mathcal{O}(\varepsilon)\kappa(\boldsymbol{\mathcal{X}})$ otherwise. Although BCGS-PIPI+$^{\text{MP}}$ appears to behave well even for high condition numbers, the `default` matrices are known to be "easy" and may not capture all potential numerical behaviors.

[7] Users who don't have access to additional toolboxes can still reproduce the trends in our multiprecision results by replacing `"mp_pair":["double", "quad"]` in https://github.com/katlund/BlockStab/blob/master/tests/configs/bcgs_pip_reortho.json with `"mp_pair":["single", "double"]`.

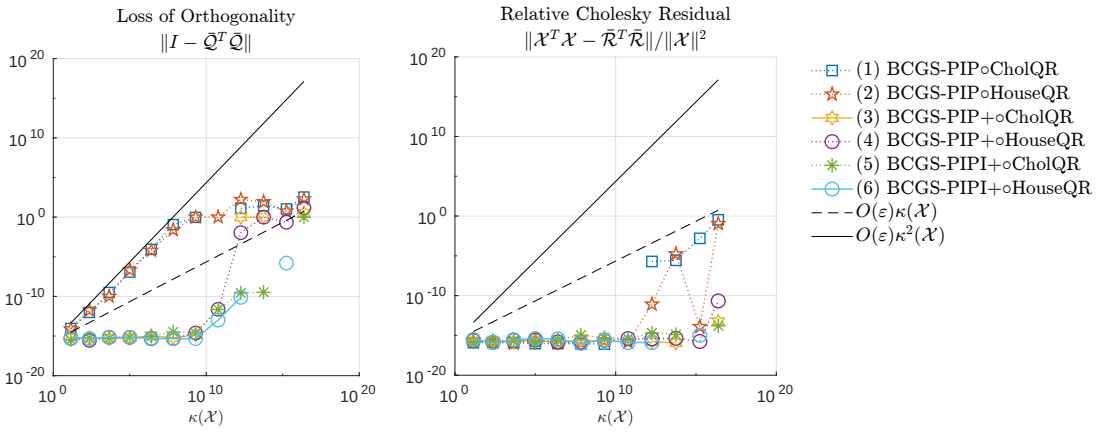

Figure 1: κ-plots for `glued` matrices.

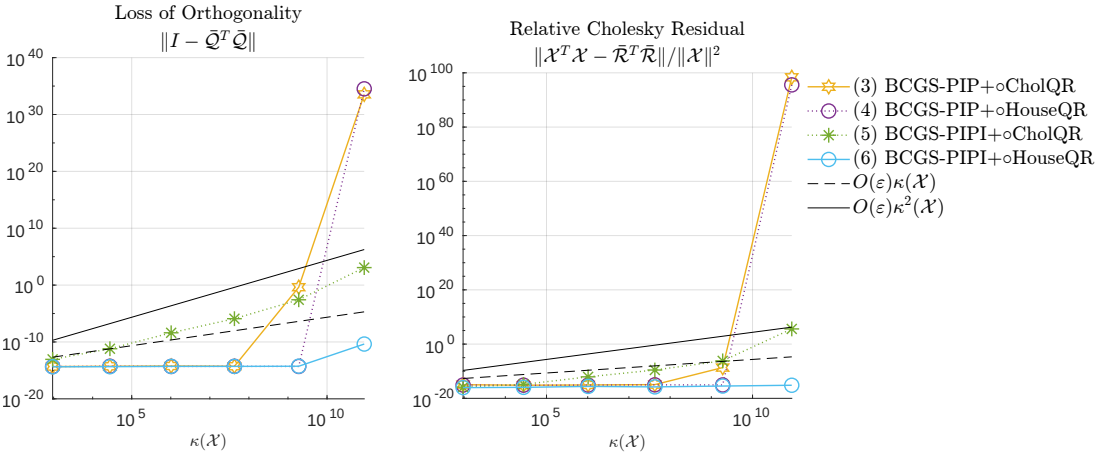

Figure 2: κ-plots for `monomial` matrices.

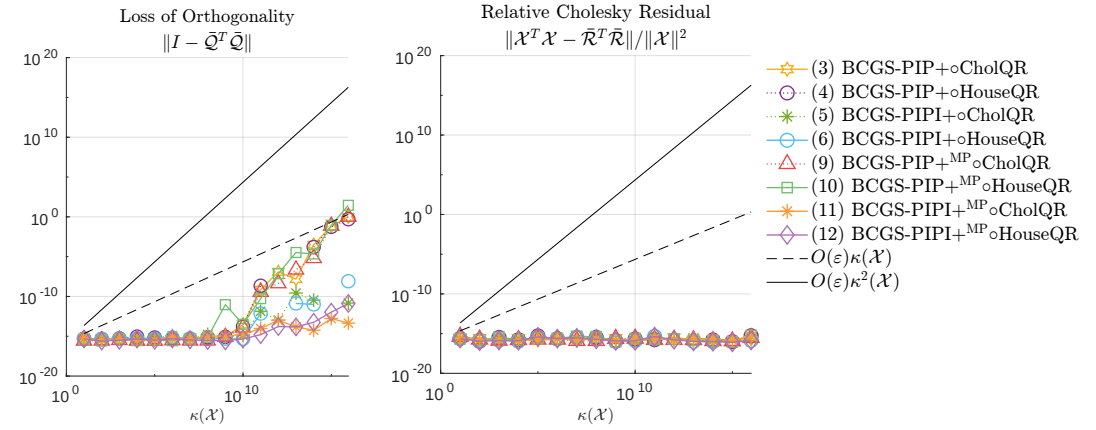

Figure 3: κ-plots for `default` matrices.

The `glued` matrices are more challenging and reveal worse behavior in Figure 4. After $\kappa(\boldsymbol{\mathcal{X}}) \approx 10^8$, BCGS-PIP+ has stability problems and the relative Cholesky residuals of BCGS-PIPI+ ∘ `CholQR` and BCGS-PIP+ ∘ `HouseQR` become `NaN`. Mixed precision overcomes this problem for both methods. The LOO of BCGS-PIPI+$^{\text{MP}}$ remains below $\mathcal{O}\left(10^{-16}\right)\kappa(\boldsymbol{\mathcal{X}})$ whereas BCGS-PIP+$^{\text{MP}}$ begins to exceed this bound. Notably, the behavior between BCGS-PIPI+$^{\text{MP}}$ ∘ `CholQR` and BCGS-PIP+$^{\text{MP}}$ ∘ `HouseQR` is very similar, despite the lack of guarantees for `CholQR`.

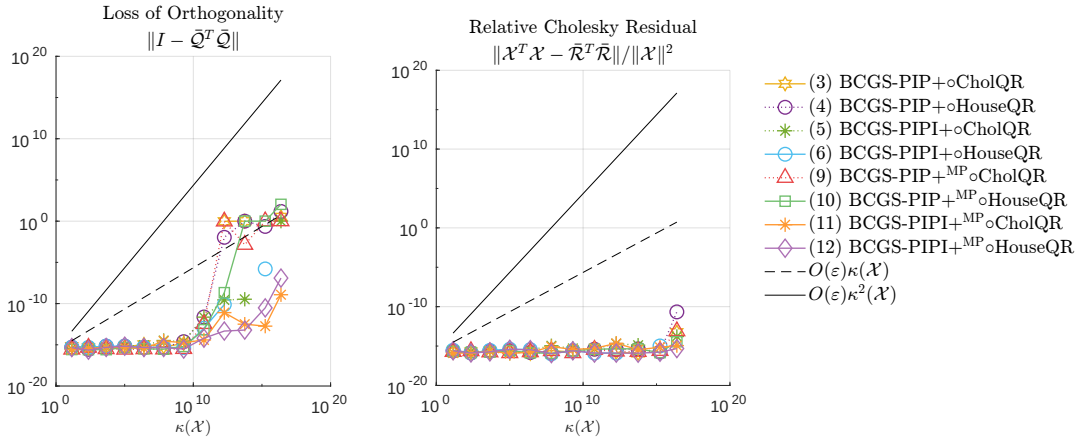

Figure 4: κ-plots for `glued` matrices.

Figures 3 and 4 might trick the reader into concluding that BCGS-PIPI+$^{\text{MP}}$ ∘ `CholQR` is rather reliable, even without theoretical bounds. The `piled` matrices in Figure 5 should dispel this notion. Neither BCGS-PIPI+∘`CholQR` nor BCGS-PIPI+$^{\text{MP}}$∘`CholQR` can attain 10^{-16} LOO for even the smallest condition numbers. And even BCGS-PIPI+$^{\text{MP}}$ ∘ `HouseQR` finally manages to lose all orthogonality for $\kappa(\boldsymbol{\mathcal{X}}) < 10^{16}$, which is still numerically nonsingular in double precision. Moreover, we observe rather erratic behavior in the LOO of both variants of BCGS-PIP+$^{\text{MP}}$ once $\kappa(\boldsymbol{\mathcal{X}}) \geq 10^8$, which emphasizes the lack of predictability outside of the bounds proven in Section 2.

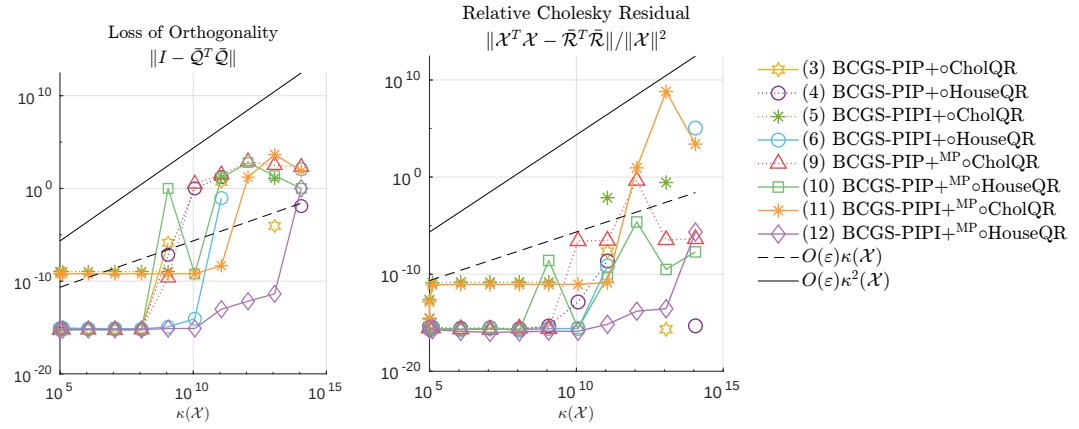

Figure 5: κ-plots for `piled` matrices.

5. Conclusions

Reorthogonalization is a simple technique for regaining stability in a Gram-Schmidt procedure. We have introduced and examined two reorthogonalized variants of BCGS-PIP, BCGS-PIP+ and BCGS-PIPI+, and demonstrated that both can achieve $\mathcal{O}(\varepsilon)$ loss of orthogonality under transparent

conditions on their intraorthogonalization routines and on the condition number of $\boldsymbol{\mathcal{X}}$, namely that $\mathcal{O}(\varepsilon)\kappa^2(\boldsymbol{\mathcal{X}}) \leq 1/2$. We have carried out the analysis in a general enough fashion so that results can be easily extended to multiprecision paradigms, and we have proposed two-precision variants BCGS-PIP+$^{\text{MP}}$ and BCGS-PIPI+$^{\text{MP}}$. Numerical experiments verify our findings and demonstrate that despite the lack of theoretical bounds, BCGS-PIPI+$^{\text{MP}}$ behaves well for several classes of test matrices and nearly overcomes the restriction on $\kappa(\boldsymbol{\mathcal{X}})$.

At the same time, the restriction on $\kappa(\boldsymbol{\mathcal{X}})$ may not be so problematic when BCGS-PIP+ or BCGS-PIPI+ forms the backbone of a block Arnoldi or GMRES algorithm and can be restarted; see, e.g., [18, 27, 29]. In fact, with the recent modular framework developed for the backward stability analysis of GMRES [6], determining reliable, adaptive restarting heuristics should be quite straightforward. In such scenarios, one usually also has access to preconditioning, which can a priori reduce the conditioning of the basis to be orthogonalized and further improve overall stability. As BCGS-PIP+ or BCGS-PIPI+ both require the same number of sync points as BCGS, but with provably better loss of orthogonality, they are promising, stable algorithms for a wide variety of applications.

Acknowledgments

The second author would like to thank the Computational Methods in Systems and Control Theory group at the Max Planck Institute for Dynamics of Complex Technical Systems for funding the fourth author's visit in March 2023. The fourth author would like to thank the Chemnitz University of Technology for funding the second author's visit in July 2023. The first, third, and fourth authors are supported by the European Union (ERC, inEXASCALE, 101075632). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. The first and the fourth authors acknowledge support from the Charles University GAUK project No. 202722 and the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. The first author additionally acknowledges support from the Charles University Research Centre program No. UNCE/24/SCI/005.

References

[1] G. Ballard, E. Carson, J. Demmel, M. Hoemmen, N. Knight, and O. Schwartz. Communication lower bounds and optimal algorithms for numerical linear algebra. *Acta Numerica 2011, Vol 20*, 23(2014):1–155, 2014. doi:10.1017/S0962492914000038.

[2] G. Ballard, J. Demmel, L. Grigori, M. Jacquelin, N. Knight, and H. Nguyen. Reconstructing Householder vectors from Tall-Skinny QR. *J. Parallel Distr. Com.*, 85:3–31, 2015. doi:10.1016/j.jpdc.2015.06.003.

[3] J. L. Barlow. Reorthogonalized block classical Gram-Schmidt using two Cholesky-based TSQR algorithms. *SIAM J. Matrix Anal. Appl.*, 45(3):1487–1517, 2024. doi:10.1137/23M1605387.

[4] J. L. Barlow and A. Smoktunowicz. Reorthogonalized block classical Gram-Schmidt. *Numerische Mathematik*, 123:395–423, 2013. doi:10.1007/s00211-012-0496-2.

[5] D. Bielich, J. Langou, S. Thomas, K. Świrydowicz, I. Yamazaki, and E. G. Boman. Low-synch Gram–Schmidt with delayed reorthogonalization for Krylov solvers. *Parallel Computing*, 112:102940, 2022. doi:10.1016/j.parco.2022.102940.

[6] A. Buttari, N. J. Higham, T. Mary, and B. Vieublé. A modular framework for the backward error analysis of GMRES. Technical Report hal-04525918, HAL science ouverte, 2024. URL: https://hal.science/hal-04525918.

[7] E. Carson, K. Lund, Y. Ma, and E. Oktay. On the loss of orthogonality in low-synchronization variants of reorthogonalized block classical Gram-Schmidt. E-Print arXiv:2408.10109, arXiv, 2024. doi:10.48550/arXiv.2408.10109.

[8] E. Carson, K. Lund, and M. Rozložník. The stability of block variants of classical Gram-Schmidt. *SIAM J. Matrix Anal. Appl.*, 42(3):1365–1380, 2021. doi:10.1137/21M1394424.

[9] E. Carson, K. Lund, M. Rozložník, and S. Thomas. Block Gram-Schmidt algorithms and their stability properties. *Linear Algebra Appl.*, 638(20):150–195, 2022. doi:10.1016/j.laa.2021.12.017.

[10] E. Carson and Y. Ma. On the backward stability of s-step GMRES. E-Print arXiv.2409.03079, arXiv, 2024. doi:10.48550/arXiv.2409.03079.

[11] E. C. Carson. *Communication-Avoiding Krylov Subspace Methods in Theory and Practice*. PhD thesis, Department of Computer Science, University of California, Berkeley, 2015. URL: http://escholarship.org/uc/item/6r91c407.

[12] T. Fukaya, R. Kannan, Y. Nakatsukasa, Y. Yamamoto, and Y. Yanagisawa. Shifted Cholesky QR for computing the QR factorization of ill-conditioned matrices. *SIAM Journal on Scientific Computing*, 42(1):A477–A503, 2020. doi:10.1137/18M1218212.

[13] S. R. Garcia and R. A. Horn. *A Second Course in Linear Algebra*. Cambridge University Press, Cambridge, 2017. doi:10.1017/9781316218419.

[14] L. Giraud, J. Langou, M. Rozložník, and J. Van Den Eshof. Rounding error analysis of the classical Gram-Schmidt orthogonalization process. *Numerische Mathematik*, 101:87–100, 2005. doi:10.1007/s00211-005-0615-4.

[15] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, 4 edition, 2013.

[16] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 2nd ed edition, 2002.

[17] M. Hoemmen. *Communication-Avoiding Krylov Subspace Methods*. PhD thesis, Department of Computer Science, University of California at Berkeley, 2010. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-37.pdf.

[18] K. Lund. Adaptively restarted block Krylov subspace methods with low-synchronization skeletons. *Numerical Algorithms*, 93(2):731–764, 2023. doi:10.1007/s11075-022-01437-1.

[19] K. Lund, E. Oktay, E. Carson, and Y. Ma. BlockStab, 2024. URL: https://github.com/katlund/BlockStab.

[20] D. Mori, Y. Yamamoto, and S. L. Zhang. Backward error analysis of the AllReduce algorithm for householder QR decomposition. *Japan Journal of Industrial and Applied Mathematics*, 29(1):111–130, 2012. doi:10.1007/s13160-011-0053-x.

[21] E. Oktay. *Mixed-Precision Computations in Numerical Linear Algebra*. PhD thesis, Faculty of Mathematics and Physics, Charles University, Prague, 2024. URL: https://dspace.cuni.cz/bitstream/handle/20.500.11956/191480/140119625.pdf?sequence=1.

[22] E. Oktay and E. Carson. Using Mixed Precision in Low-Synchronization Reorthogonalized Block Classical Gram-Schmidt. *PAMM*, 23(1):e202200060, 2023. doi:10.1002/pamm.202200060.

[23] A. Smoktunowicz, J. L. Barlow, and J. Langou. A note on the error analysis of classical Gram-Schmidt. *Numerische Mathematik*, 105(2):299–313, 2006. doi:10.1007/s00211-006-0042-1.

[24] G. W. Stewart. Block Gram-Schmidt orthogonalization. *SIAM Journal on Scientific Computing*, 31(1):761–775, 2008. doi:10.1137/070682563.

[25] S. Thomas, E. Carson, M. Rozložník, A. Carr, and K. Świrydowicz. Iterated Gauss–Seidel GMRES. *SIAM Journal on Scientific Computing*, pages S254–S279, 2023. URL: https://epubs.siam.org/doi/10.1137/22M1491241, doi:10.1137/22M1491241.

[26] L. N. Trefethen and D. I. Bau. *Numerical Linear Algebra*. SIAM, Philadelphia, 1997.

[27] Z. Xu, J. J. Alonso, and E. Darve. A numerically stable communication-avoiding s-step GMRES algorithm. Technical Report arXiv:2303.08953, arXiv, 2023. doi:10.48550/arXiv.2303.08953.

[28] Y. Yamamoto, Y. Nakatsukasa, Y. Yanagisawa, and T. Fukaya. Roundoff error analysis of the Cholesky QR2 algorithm. *Electronic Transactions on Numerical Analysis*, 44:306–326, 2015. URL: http://www.emis.de/journals/ETNA/vol.44.2015/pp306-326.dir/pp306-326.pdf.

[29] I. Yamazaki, A. J. Higgins, E. G. Boman, and D. B. Szyld. Two-Stage Block Orthogonalization to Improve Performance of s-step GMRES. In *2024 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 26–37, San Francisco, CA, USA, 2024. doi:10.1109/IPDPS57955.2024.00012.

[30] I. Yamazaki, S. Thomas, M. Hoemmen, E. G. Boman, K. Świrydowicz, and J. J. Eilliot. Low-synchronization orthogonalization schemes for s-step and pipelined Krylov solvers in Trilinos. In *Proceedings of the 2020 SIAM Conference on Parallel Processing for Scientific Computing (PP)*, pages 118–128, 2020. doi:10.1137/1.9781611976137.11.

[31] Q. Zou. A flexible block classical Gram–Schmidt skeleton with reorthogonalization. *Numerical Linear Algebra with Applications*, 30(5):e2491, 2023. doi:10.1002/nla.2491.

A. Proofs of the theorems in Section 3

A.1. Proof of Lemma 3

Proof. Assumption (91) implies

$$\left\|\bar{\mathcal{R}}\right\|^2 \leq (1+\xi)\left\|\boldsymbol{\mathcal{X}}\right\|^2, \tag{114}$$

and together with $\xi\kappa^2(\boldsymbol{\mathcal{X}}) \leq \frac{1}{2}$ and the perturbation theory of singular values [15, Corollary 8.6.2], we can derive a bound on $\left\|\bar{\mathcal{R}}^{-1}\right\|^2$:

$$\left\|\bar{\mathcal{R}}^{-1}\right\|^2 = \frac{1}{\sigma_{\min}^2(\bar{\mathcal{R}})} \leq \frac{1}{\sigma_{\min}^2(\boldsymbol{\mathcal{X}})(1-\xi\kappa^2(\boldsymbol{\mathcal{X}}))}. \tag{115}$$

Multiplying (92) on the left by its transpose, then on the left by $\bar{\mathcal{R}}^{-T}$ and on the right by $\bar{\mathcal{R}}^{-1}$ and substituting (91) leads to

$$\bar{\boldsymbol{\mathcal{Q}}}^T\bar{\boldsymbol{\mathcal{Q}}} = I - \bar{\mathcal{R}}^{-T}(\Delta\mathcal{E} - \boldsymbol{\mathcal{X}}^T\Delta\boldsymbol{\mathcal{D}} - \Delta\boldsymbol{\mathcal{D}}^T\boldsymbol{\mathcal{X}})\bar{\mathcal{R}}^{-1}, \tag{116}$$

where we have dropped the quadratic term. Taking the norm and applying bounds (91), (92), (114), and (115) to (116) leads to a quadratic inequality:

$$\begin{aligned}\left\|\bar{\boldsymbol{\mathcal{Q}}}\right\|^2 &\leq 1 + \left\|\bar{\mathcal{R}}^{-1}\right\|^2\left(\xi\left\|\boldsymbol{\mathcal{X}}\right\|^2 + 2\rho\left\|\boldsymbol{\mathcal{X}}\right\|\left(\left\|\boldsymbol{\mathcal{X}}\right\| + \left\|\bar{\boldsymbol{\mathcal{Q}}}\right\|\left\|\bar{\mathcal{R}}\right\|\right)\right)\\ &\leq 1 + \frac{(\xi+2\rho)\left\|\boldsymbol{\mathcal{X}}\right\|^2 + 2\rho\sqrt{1+\xi}\left\|\boldsymbol{\mathcal{X}}\right\|^2\left\|\bar{\boldsymbol{\mathcal{Q}}}\right\|}{\sigma_{\min}^2(\boldsymbol{\mathcal{X}})(1-\xi\kappa^2(\boldsymbol{\mathcal{X}}))}\\ &\leq \frac{1+2\rho\kappa^2(\boldsymbol{\mathcal{X}})}{1-\xi\kappa^2(\boldsymbol{\mathcal{X}})} + \frac{2\rho\sqrt{1+\xi}\kappa^2(\boldsymbol{\mathcal{X}})}{1-\xi\kappa^2(\boldsymbol{\mathcal{X}})}\left\|\bar{\boldsymbol{\mathcal{Q}}}\right\|.\end{aligned} \tag{117}$$

Applying the assumptions $\xi\kappa^2(\boldsymbol{\mathcal{X}}) \leq \frac{1}{2}$ and $\rho\kappa^2(\boldsymbol{\mathcal{X}}) \leq \frac{1}{4}$, we see that

$$\frac{1+2\rho\kappa^2(\boldsymbol{\mathcal{X}})}{1-\xi\kappa^2(\boldsymbol{\mathcal{X}})} \leq 3 \text{ and } \frac{2\rho\kappa^2(\boldsymbol{\mathcal{X}})}{1-\xi\kappa^2(\boldsymbol{\mathcal{X}})} \leq 1,$$

so (117) simplifies to

$$\left\|\bar{\boldsymbol{\mathcal{Q}}}\right\|^2 - \sqrt{1+\xi}\left\|\bar{\boldsymbol{\mathcal{Q}}}\right\| - 3 \leq 0, \tag{118}$$

which can be easily solved with the quadratic formula to reveal

$$\left\|\bar{\boldsymbol{\mathcal{Q}}}\right\| \leq \frac{1}{2}\left(\sqrt{1+\xi}+\sqrt{13+\xi}\right) \leq \frac{1}{2}\left(\sqrt{2}+\sqrt{14}\right) \leq 3,$$

thus proving (94). Revisiting (116), we take the norm of $\left\|I-\bar{\boldsymbol{\mathcal{Q}}}^T\bar{\boldsymbol{\mathcal{Q}}}\right\|$ and apply (94) along with bounds (91), (92), (114), (115) again to arrive at (93). Finally, (95) follows from (114) and (94). □

A.2. Proof of Lemma 4

Proof. Writing $\bar{\mathcal{R}}_k$ as

$$\bar{\mathcal{R}}_k = \begin{bmatrix} \bar{\mathcal{R}}_{k-1} & \bar{\mathcal{R}}_{1:k-1,k} \\ 0 & \bar{R}_k \end{bmatrix}$$

implies

$$\bar{\mathcal{R}}_k^T\bar{\mathcal{R}}_k = \begin{bmatrix} \bar{\mathcal{R}}_{k-1}^T\bar{\mathcal{R}}_{k-1} & \bar{\mathcal{R}}_{k-1}^T\bar{\mathcal{R}}_{1:k-1,k} \\ \bar{\mathcal{R}}_{1:k-1,k}^T\bar{\mathcal{R}}_{k-1} & \bar{\mathcal{R}}_{1:k-1,k}^T\bar{\mathcal{R}}_{1:k-1,k}+\bar{R}_k^T\bar{R}_k \end{bmatrix}. \tag{119}$$

The upper diagonal block of (119) can be handled directly from (96). Lemma 3 holds for $k-1$ via the assumptions (96) and (97). Then for the off-diagonals, from (94) and (98), we can write

$$\left\|\bar{\mathcal{R}}_{1:k-1,k}\right\| \leq \left\|\bar{\boldsymbol{\mathcal{Q}}}_{k-1}\right\| \|\boldsymbol{X}_k\| + \|\Delta\boldsymbol{R}_k\| \leq (3+\delta_{Q^TX})\|\boldsymbol{X}_k\| \leq 4\|\boldsymbol{X}_k\|. \tag{120}$$

Multiplying (98) by $\bar{\mathcal{R}}_{k-1}^T$ from left together with (97) leads to

$$\begin{aligned} \bar{\mathcal{R}}_{k-1}^T\bar{\mathcal{R}}_{1:k-1,k} &= \bar{\mathcal{R}}_{k-1}^T(\bar{\boldsymbol{\mathcal{Q}}}_{k-1}^T\boldsymbol{X}_k+\Delta\boldsymbol{R}_k) \\ &= \boldsymbol{\mathcal{X}}_{k-1}^T\boldsymbol{X}_k+\Delta\mathcal{E}_{1:k-1,k}, \end{aligned} \tag{121}$$

where $\Delta\mathcal{E}_{1:k-1,k} := \Delta\boldsymbol{\mathcal{D}}_{k-1}^T\boldsymbol{X}_k + \bar{\mathcal{R}}_{k-1}^T\Delta\boldsymbol{R}_k$. From (96), (97), Lemma 3, and (98) it follows that

$$\begin{aligned} \|\Delta\mathcal{E}_{1:k-1,k}\| &\leq 6\cdot\rho_{k-1}\|\boldsymbol{\mathcal{X}}_{k-1}\|\|\boldsymbol{X}_k\| + \delta_{Q^TX}\sqrt{1+\xi_{k-1}}\|\boldsymbol{\mathcal{X}}_{k-1}\|\|\boldsymbol{X}_k\| \\ &\leq \left(6\cdot\rho_{k-1}+\sqrt{2}\cdot\delta_{Q^TX}\right)\|\boldsymbol{\mathcal{X}}_{k-1}\|\|\boldsymbol{X}_k\|. \end{aligned} \tag{122}$$

As for the bottom diagonal block of (119), we can combine (99)–(101) and rearrange terms to find

$$\begin{aligned} \bar{\mathcal{R}}_{1:k-1,k}^T\bar{\mathcal{R}}_{1:k-1,k}+\bar{R}_k^T\bar{R}_k &= \bar{P}_k+\Delta F_k+\Delta C_k \\ &= \boldsymbol{X}_k^T\boldsymbol{X}_k+\underbrace{\Delta P_k+\Delta F_k+\Delta C_k}_{=:\Delta E_k}, \end{aligned} \tag{123}$$

where

$$\|\Delta E_k\| \leq \left(\delta_{X^TX}+\delta_{R^TR}+\delta_{\texttt{chol}}\right)\|\boldsymbol{X}_k\|^2. \tag{124}$$

Combining (96), (121), (122), and (124) we can write

$$\bar{\mathcal{R}}_k^T\bar{\mathcal{R}}_k = \underbrace{\begin{bmatrix} \boldsymbol{\mathcal{X}}_{k-1}^T\boldsymbol{\mathcal{X}}_{k-1} & \boldsymbol{\mathcal{X}}_{k-1}^T\boldsymbol{X}_k \\ \boldsymbol{X}_k^T\boldsymbol{\mathcal{X}}_{k-1} & \boldsymbol{X}_k^T\boldsymbol{X}_k \end{bmatrix}}_{=\boldsymbol{\mathcal{X}}_k^T\boldsymbol{\mathcal{X}}_k} + \underbrace{\begin{bmatrix} \Delta\mathcal{E}_{k-1} & \Delta\mathcal{E}_{1:k-1,k} \\ \Delta\mathcal{E}_{1:k-1,k}^T & \Delta E_k \end{bmatrix}}_{=:\Delta\mathcal{E}_k},$$

where applying [13, P.15.50] leads to

$$\begin{aligned} \|\Delta\mathcal{E}_k\| &\leq \|\Delta\mathcal{E}_{k-1}\|+2\|\Delta\mathcal{E}_{1:k-1,k}\|+\|\Delta E_k\| \\ &\leq \xi_{k-1}\|\boldsymbol{\mathcal{X}}_{k-1}\|^2+2\left(6\cdot\rho_{k-1}+\sqrt{2}\cdot\delta_{Q^TX}\right)\|\boldsymbol{\mathcal{X}}_{k-1}\|\|\boldsymbol{X}_k\| \\ &\quad+\left(\delta_{X^TX}+\delta_{\texttt{chol}}+\delta_{R^TR}\right)\|\boldsymbol{X}_k\|^2 \\ &\leq \left(\xi_{k-1}+12\cdot\rho_{k-1}+2\sqrt{2}\cdot\delta_{Q^TX}+\delta_{X^TX}+\delta_{\texttt{chol}}+\delta_{R^TR}\right)\|\boldsymbol{\mathcal{X}}_k\|^2, \end{aligned}$$

which proves (104).

To prove (105), combining (102) and (103) yields

$$\bar{\boldsymbol{Q}}_k \bar{R}_k = \boldsymbol{X}_k - \bar{\boldsymbol{Q}}_{k-1} \bar{\mathcal{R}}_{1:k-1,k} + \Delta \boldsymbol{D}_k,$$

where $\Delta \boldsymbol{D}_k = \Delta \boldsymbol{V}_k + \Delta \boldsymbol{G}_k$ and

$$\begin{aligned} \|\Delta \boldsymbol{D}_k\| &\leq \delta_{QR} \|\boldsymbol{X}_k\| + \delta_Q \left\|\bar{\boldsymbol{Q}}_k\right\| \left\|\bar{R}_k\right\| \\ &\leq (\delta_{QR} + \delta_Q)\left(\|\boldsymbol{X}_k\| + \left\|\bar{\boldsymbol{Q}}_k\right\| \left\|\bar{R}_k\right\|\right). \end{aligned} \tag{125}$$

Writing

$$\begin{aligned} \bar{\boldsymbol{\mathcal{Q}}}_k \bar{\mathcal{R}}_k &= \begin{bmatrix} \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\mathcal{R}}_{k-1} & \bar{\boldsymbol{\mathcal{Q}}}_{k-1} \bar{\mathcal{R}}_{1:k-1,k} + \bar{\boldsymbol{Q}}_k \bar{R}_k \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} \boldsymbol{\mathcal{X}}_{k-1} & \boldsymbol{X}_k \end{bmatrix}}_{=\boldsymbol{\mathcal{X}}_k} + \underbrace{\begin{bmatrix} \Delta \boldsymbol{\mathcal{D}}_{k-1} & \Delta \boldsymbol{D}_k \end{bmatrix}}_{=:\Delta \boldsymbol{\mathcal{D}}_k}, \end{aligned}$$

together with (97), Lemma 3, and (125) gives

$$\begin{aligned} \|\Delta \boldsymbol{\mathcal{D}}_k\| &\leq \|\Delta \boldsymbol{\mathcal{D}}_{k-1}\| + \|\Delta \boldsymbol{D}_k\| \\ &\leq 6 \cdot \rho_{k-1} \|\boldsymbol{\mathcal{X}}_{k-1}\| + (\delta_{QR} + \delta_Q)(\|\boldsymbol{X}_k\| + \left\|\bar{\boldsymbol{Q}}_k\right\| \left\|\bar{R}_k\right\|) \\ &\leq \left(6 \cdot \rho_{k-1} + \delta_{QR} + \delta_Q\right)\left(\|\boldsymbol{\mathcal{X}}_k\| + \left\|\bar{\boldsymbol{\mathcal{Q}}}_k\right\| \left\|\bar{\mathcal{R}}_k\right\|\right), \end{aligned}$$

which completes the proof. □

A.3. Proof of theorem 3

Proof. We start with the base case, $k = 1$. By the assumptions (108) and (109), (110) and (111) follow trivially. Consequently, by setting $\xi_1 = \mathcal{O}(\varepsilon_\ell)$, $\rho_1 = \mathcal{O}(\varepsilon_\ell)$ in (101) and (97), respectively, and by applying (106) and (107), it holds that $\xi_1 \kappa^2(\boldsymbol{X}_1) \leq \frac{1}{2}$ and $\rho_1 \kappa^2(\boldsymbol{X}_1) \leq \frac{1}{4}$. Then we can apply Lemma 3 to conclude (112) for $k = 1$.

Now, by following standard rounding-error analysis from [16] (particularly [16, Lemma 6.6, Theorem 8.5, & Theorem 10.3]), we find that for all $k \in \{2, \ldots, p\}$, there exist constants $\delta_{Q^T X}, \delta_{QR}, \delta_{X^T X}, \delta_{R^T R}, \delta_{\texttt{chol}}, \delta_Q$ such that

$$\delta_{Q^T X}, \delta_{QR} = \mathcal{O}(\varepsilon_\ell) \text{ and } \delta_{X^T X}, \delta_{R^T R}, \delta_{\texttt{chol}}, \delta_Q = \mathcal{O}(\varepsilon_h). \tag{126}$$

and (98)–(103) hold.

Consider just $k = 2$ for a moment. Then Lemma 4 can be applied (because we already have the bound (112) for $k = 1$) to conclude

$$\begin{aligned} &\bar{\mathcal{R}}_2^T \bar{\mathcal{R}}_2 = \boldsymbol{\mathcal{X}}_2^T \boldsymbol{\mathcal{X}}_2 + \Delta \mathcal{E}_2, \quad \|\Delta \mathcal{E}_2\| \leq \xi_2 \|\boldsymbol{\mathcal{X}}_2\|^2; \text{ and} \\ &\bar{\boldsymbol{\mathcal{Q}}}_2 \bar{\mathcal{R}}_2 = \boldsymbol{\mathcal{X}}_2 + \boldsymbol{\mathcal{D}}_2, \quad \|\boldsymbol{\mathcal{D}}_2\| \leq \rho_2(\|\boldsymbol{\mathcal{X}}_2\| + \left\|\bar{\boldsymbol{\mathcal{Q}}}_2\right\| \left\|\bar{\mathcal{R}}_2\right\|), \end{aligned}$$

where $\xi_2 = \mathcal{O}(\varepsilon_\ell)$ and $\rho_2 = \mathcal{O}(\varepsilon_\ell)$ because $\xi_1 = \mathcal{O}(\varepsilon_\ell)$ and the bounds (126). Thus (110) and (111) hold for $k = 2$, and Lemma 3 can subsequently be applied to prove (112) for $k = 2$.

Proceeding by strong induction and identical logic as for $k = 2$, we can then conclude (110)–(112) hold for all $k \in \{2, \ldots, p\}$. □