Stability Analysis of QR factorization in an Oblique Inner Product

Bradley R. Lowery* and Julien Langou*
Mathematical and Statistical Sciences
University of Colorado Denver
Denver, CO, USA
{*Bradley.Lowery, Julien.Langou*}@*ucdenver.edu*

Abstract—Sets of vectors that are mutually orthogonal with respect to an oblique inner product are often used in iterative methods such as the conjugate gradient method. However, these vectors cannot be computed exactly since floating point arithmetic introduces error into the computation. Large error in the computation my slow the convergence of the associated iterative method, a topic we do not address in this paper. Rather, we wish to improve our understanding of the error introduced in computing such vectors. For this study we focus on QR factorization algorithms.

In this paper, we consider the stability of the QR factorization in an oblique inner product. The oblique inner product is defined by a symmetric positive definite matrix A. We analyze two algorithm that are based a factorization of A and converting the problem to the Euclidean case. The two algorithms we consider use the Cholesky decomposition and the eigenvalue decomposition. We also analyze algorithms that are based on computing the Cholesky factor of the normal equation. We present numerical experiments to show the error bounds are tight. Finally we present performance results for these algorithms as well as Gram-Schmidt methods on parallel architecture. The performance experiments demonstrate the benefit of the communication avoiding algorithms.

Keywords-QR factorization; oblique; stability; backward error; Cholesky; normal equation; communication avoiding;

I. INTRODUCTION

We are interested in computing a QR factorization of Z in an oblique inner product space. Given an $m \times m$ symmetric positive definite matrix A, we define the inner product by

$$x^T A y, \quad x, y \in \mathbb{R}^m$$

and the induced norm by $\|x\|_A = \sqrt{x^T A x}$. We seek factors Q and R such that $Z = QR$, R is an $n \times n$ upper triangular matrix, and Q is an $m \times n$ matrix such that $Q^T A Q = I$. Z is assumed to be full rank and $m \geq n$.

Since $R^T R = (A^{1/2} Z)^T (A^{1/2} Z)$, the singular values of R are $\sigma_i(R) = \sigma_i(A^{1/2} Z)$. We can also derive an equality for the singular values of Q. Since $I = Q^T A Q = (A^{1/2} Q)^T (A^{1/2} Q)$, $A^{1/2} Q$ has orthonormal columns. Let $\hat{Z}$ have columns that form an orthonormal basis of Z. $\hat{Z} \hat{Z}^T$ is a projection onto the column space of Z, so $Q = \hat{Z} \hat{Z}^T Q$

and $A^{1/2} Q = A^{1/2} \hat{Z} \hat{Z}^T Q$. Then

$$\begin{aligned}
\sigma_i(A^{1/2} \hat{Z}) &= \sigma_i(A^{1/2} \hat{Z} \hat{Z}^T Q (\hat{Z}^T Q)^{-1}) \\
&= \sigma_i((\hat{Z}^T Q)^{-1}) \\
&= \sigma_{n-i+1}(\hat{Z}^T Q)^{-1} \\
&= \sigma_{n-i+1}(\hat{Z} \hat{Z}^T Q)^{-1} \\
&= \sigma_{n-i+1}(Q)^{-1}.
\end{aligned}$$

We organize these results in the following theorem.

Theorem I.1. *Let Z be full rank. Let $\hat{Z}$ have columns that form an orthonormal basis of Z. Then the Q and R factor of the oblique QR factorization in the A inner product satisfy,*

$$\sigma_i(Q) = \sigma_{n-i+1}(A^{1/2} \hat{Z})^{-1},$$
$$\sigma_i(R) = \sigma_i(A^{1/2} Z),$$

for all $i = 1, \ldots, n$. In particular, $\|Q\|_2 = \sigma_n(A^{1/2} \hat{Z})^{-1}$ and $\|R\|_2 = \|A^{1/2} Z\|_2$.

For the Euclidean cases ($A = I$), we recover the standard QR factorization, a classic problem in linear algebra. The stability of Euclidean QR decomposition algorithms has been studied greatly. For Householder, Givens, and modified Gram-Schmidt we refer to [1, Chapter 19]. For classical Gram Schmidt we refer to [2], [3].

There are two main classes of algorithms for computing the oblique QR factorization: algorithms that use a factorization of A, and those that do not. The first compute a factorization of $A = B^T B$ and convert the problem to an Euclidean QR factorization of $BZ = YR$. The Q factor may be obtained by $Q = B^{-1} Y$. We consider the cases where the computed B is the Cholesky factor of A and where B is obtained by the eigenvalue decomposition. If $A = VDV^T$, where D is diagonal and V is unitary, then we let $B = D^{1/2} V$. Gram-Schmidt schemes are designed to be orthogonalized with respect to any inner product and do not require a factorization of A. Also, since we have the relationship $Z^T A Z = R^T R$, we can compute the R factor as the Cholesky factor of the normal equation $Z^T A Z$. Again no factorization of A is required. Algorithms similar to Householder or Givens do not currently exist. However, a closely related problem of computing the QR factorization, where Q is A-invariant (i.e. $Q^T A Q = A$) has

been considered. Gulliksson [4] presents stability analysis for a Householder-like algorithm when A is diagonal and the result is used to solve the weighted least squares problem.

Rozložník et al. [5] extend the error analysis of the Gram-Schmidt methods to the general case of an oblique inner product. They also analyzed the method based on the eigenvalue decomposition of A. In Section IV we provide stability bounds for the when B is the Cholesky factorization of A. In Section V we analyze the case when B is given from the eigenvalue decomposition and present improved error bounds of those from Rozložník et al. In Section II we analyze the stability of the algorithm based on computing the R factor as the Cholesky factor of the normal equation, $Z^T A Z$. In Section III we analyze the normal equation algorithm when first computing a Euclidean QR factorization.

An application of the oblique QR factorization is computing the solution to the generalized least squares problem

$$\min_x \|Zx - b\|_A. \tag{I.1}$$

The linear least squares estimate computed via an oblique QR factorization is given by $\tilde{x} = R^{-1} Q^T A b$. If the elements of b are taken to be random variables, then the best linear unbiased estimate is the solution to (I.1), where A is the inverse covariance matrix. For an overview of the generalized least squares problem we refer to [6, Chapter 4].

In some applications it is reasonable to assume that initially a factorization of A is known rather than A itself. This is the case when A is an estimate of the covariance matrix computed by $\frac{1}{n-1} B^T B$, where B has n rows corresponding to data samples. [7], [8], [9] study the stability and performance of the modified Gram-Schmidt algorithm when a factorization $A = B^T B$ is know initially. Here, A is never explicitly formed. We don't assume any factorization of A is know initially in this paper.

Another application is the generalized eigenvalue problem

$$Bx = \lambda A x, \tag{I.2}$$

where $A \in \mathbb{R}^{m \times m}$ is symmetric positive definite, and $B \in \mathbb{R}^{m \times m}$ is symmetric. If $Q \in \mathbb{R}^{m \times m}$ is A-orthogonal, then the standard eigenvalue problem

$$Q^T B Q x = \lambda x$$

has the same eigenvalues as (I.2). The same idea can be used for iterative methods where $Q \in \mathbb{R}^{m \times n}$ is a basis of a Krylov subspace of dimension n.

In Section VII we test the bounds obtained in Sections II, III, IV, and V. From these experiments we conclude that our bounds are tight. In Section VIII we provide performance experiments on an Intel cluster.

Notation: We assume standard floating point arithmetic

$$\text{fl}(x \text{ op } y) = (x \text{ op } y)(1 + \delta), \quad |\delta| \le u,$$

where u denotes machine precision and op $= +, -, .*, /$. To avoid clouding the major results we will not track small constants and merely let c represent a small constant that may change from one line to another. We will also make assumptions on the size of m and n so as to simplify the error bounds. We will frequently assume, for example, that $nu < 1/2$, so

$$\frac{nu}{1 - nu} \le 2nu = cnu.$$

Backward error results for major kernels: matrix multiplication, triangle solve, Cholesky factorization, and Householder QR factorization can be found in [1, Chapters 3, 8, 10, 19].

The set of $m \times n$ real matrices is represented by $\mathbb{R}^{m \times n}$. If $A \in \mathbb{R}^{m \times n}$, then A_j or $A_{:,j}$ is the j-th column of A and $A_{i,:}$ is the i-th row of A. Inequalities between matrices are understood to hold entry-wise.

II. STABILITY ANALYSIS OF CHOLQR

The R of an oblique QR factorization is the Cholesky factor of the $Z^T A Z$. This is easily seen from

$$Z^T A Z = R^T Q^T A Q R = R^T R.$$

Once the R factor is found from a Cholesky decomposition, the Q factor is obtained by a triangle solve. Algorithm 1, which we call CHOLQR, summarizes this method.

Algorithm 1: CHOLQR

Input : $Z \in \mathbb{R}^{m \times n}$, $A \in \mathbb{R}^{m \times m}$ - symmetric positive definite
Output: $Q \in \mathbb{R}^{m \times n}$, $R \in \mathbb{R}^{n \times n}$
$B = AZ$;
$C = Z^T B$;
$R = \text{chol}(C)$;
$Q = Z/R$;

There are essentially three kernels: matrix multiplication, Cholesky decomposition, and triangle solve. The error in each of these computations are

$$\tilde{B} = AZ + \Delta B, \quad \text{s.t. } |\Delta B| \le cmu|A||Z| \tag{II.1}$$

$$\tilde{C} = Z^T \tilde{B} + \Delta C_1, \quad \text{s.t. } |\Delta C_1| \le cmu|Z^T||\tilde{B}| \tag{II.2}$$

$$\tilde{C} = \tilde{R}^T \tilde{R} + \Delta C_2, \quad \text{s.t. } |\Delta C_2| \le cnu|\tilde{R}^T||\tilde{R}| \tag{II.3}$$

$$\tilde{Q}_{i,:}(\tilde{R} + \Delta R^i) = Z_{i,:}, \quad \text{s.t. } |\Delta R^i| \le cnu|\tilde{R}| \tag{II.4}$$

To ensure that the Cholesky factorization in (II.4) runs to completion we assume that $cn^{3/2}\kappa(\tilde{C}) < 1$ (see [1, Chapter 10]). Let $\Delta Z = -[\tilde{Q}_{1,:}\Delta R^1; \ldots; \tilde{Q}_{i,:}\Delta R^i; \ldots; \tilde{Q}_{m,:}\Delta R^m]$. Then the final equation becomes

$$\tilde{Q}\tilde{R} = Z + \Delta Z, \quad \text{s.t. } |\Delta Z| \le cnu|\tilde{Q}||\tilde{R}|.$$

This equation gives us the componentwise representativity bound

$$|Z - \tilde{Q}\tilde{R}| \le cnu|\tilde{Q}||\tilde{R}|.$$

And the normwise bound follows:

$$\|Z - \tilde{Q}\tilde{R}\|_2 \le cnu\|\tilde{Q}\|_2\|\tilde{R}\|_2.$$

For the loss of orthogonality we use that $\tilde{Q} = Z\tilde{R}^{-1} + \Delta Z\tilde{R}^{-1}$ and form

$$\tilde{Q}^T A\tilde{Q} = (\tilde{R}^{-T}Z^T + \tilde{R}^{-T}\Delta Z^T)A(Z\tilde{R}^{-1} + \Delta Z\tilde{R}^{-1})$$

(II.5)

$$= \tilde{R}^{-T}Z^T AZ\tilde{R}^{-1} + \tilde{R}^{-T}Z^T A\Delta Z\tilde{R}^{-1}$$
$$+ \tilde{R}^{-T}\Delta Z^T AZ\tilde{R}^{-1} + \mathcal{O}(u^2).$$

From (II.1), (II.2), and (II.4),

$$Z^T AZ = Z^T(\tilde{B} - \Delta B) \qquad (\text{II.6})$$
$$= \tilde{C} - \Delta C_1 - Z^T\Delta B$$
$$= \tilde{R}^T\tilde{R} + \Delta C_2 - \Delta C_1 - Z^T\Delta B.$$

Substituting II.6 into II.5 gives

$$\tilde{Q}^T A\tilde{Q} = I + \tilde{R}^{-T}\Delta C_2\tilde{R}^{-1} + \tilde{R}^{-T}\Delta C_1\tilde{R}^{-1} \qquad (\text{II.7})$$
$$+ \tilde{R}^{-T}Z^T\Delta B\tilde{R}^{-1} + \tilde{R}^{-T}Z^T A\Delta Z\tilde{R}^{-1}$$
$$+ \tilde{R}^{-T}\Delta Z^T AZ\tilde{R}^{-1} + \mathcal{O}(u^2).$$

Since $Z\tilde{R}^{-1} = \tilde{Q} - \Delta Z\tilde{R}^{-1}$, $\|Z\tilde{R}^{-1}\|_2 \le \|\tilde{Q}\|_2 + \mathcal{O}(u)$. Similarly, we have $\|\tilde{C}\|_2 \le \|Z^T\tilde{B}\|_2 + \mathcal{O}(u)$ and $\|\tilde{B}\|_2 \le \|AZ\|_2 + \mathcal{O}(u)$. Using these equalities with (II.7) gives the bound

$$\|\tilde{Q}^T A\tilde{Q} - I\|_2 \le cmnu(\|\tilde{R}^{-1}\|_2^2\|Z\|_2\|AZ\|_2 \qquad (\text{II.8})$$
$$+ \|\tilde{R}^{-1}\|_2\|\tilde{Q}\|_2\|A\|_2\|Z\|_2$$
$$+ \|\tilde{R}^{-1}\|_2\|A\tilde{Q}\|_2\|\tilde{Q}\|_2\|\tilde{R}\|_2) + \mathcal{O}(u^2).$$

We may eliminate the third term from (II.8) by noting that

$$\|A\tilde{Q}\|_2\|\tilde{R}\|_2 \le \|A\|_2\|Z\|_2 + \mathcal{O}(u). \qquad (\text{II.9})$$

Inequality (II.9) can be obtained from the following. Let $\Delta N = Z^T AZ - \tilde{R}^T\tilde{R}$. From (II.6), $\|\Delta N\|_2 = \mathcal{O}(u)$ and $\|\tilde{R}\|_2^2 \le \|A^{1/2}Z\|_2^2 + \|\Delta N\|_2$. Hence,

$$\|\tilde{R}\|_2 \le \|A^{1/2}Z\|_2\sqrt{1 + \frac{\|\Delta N\|_2}{\|A^{1/2}Z\|_2^2}} \qquad (\text{II.10})$$

$$\le \|A^{1/2}Z\|_2\left(1 + \frac{\|\Delta N\|_2}{\|A^{1/2}Z\|_2^2}\right)$$

$$\le \|A^{1/2}Z\|_2 + \mathcal{O}(u).$$

Similarly, since $\tilde{Q}^T A\tilde{Q} - I = \mathcal{O}(u)$,

$$\|A^{1/2}\tilde{Q}\|_2 \le 1 + \mathcal{O}(u). \qquad (\text{II.11})$$

Combining (II.10) and (II.11) we obtain (II.9). Theorem II.1 summarizes the results.

Theorem II.1. *The computed factors, $\tilde{Q}$ and $\tilde{R}$, of algorithm 1 satisfy*

$$\|Z - \tilde{Q}\tilde{R}\|_2 \le cnu\|\tilde{Q}\|_2\|\tilde{R}\|_2, \qquad (\text{II.12})$$

$$|Z - \tilde{Q}\tilde{R}| \le cnu|\tilde{Q}||\tilde{R}|, \qquad (\text{II.13})$$

$$\|\tilde{Q}^T A\tilde{Q} - I\|_2 \le cmnu\|\tilde{R}^{-1}\|_2\|Z\|_2 \qquad (\text{II.14})$$
$$(\|\tilde{R}^{-1}\|_2\|AZ\|_2 + \|\tilde{Q}\|_2\|A\|_2) + \mathcal{O}(u^2)$$

The loss of orthogonality bound can be difficult to understand in its current form. At first glance a worse case bound is

$$\|\tilde{Q}^T A\tilde{Q} - I\|_2 \le cmnu\kappa(\tilde{R})^2\kappa(A) + \mathcal{O}(u^2).$$

However, in many cases the conditioning of R increases when increasing the conditioning of A. This creates a bound that is proportional to $\kappa(A)^2$, a result not seen in numerical experiments (see Section VII). Instead we search for a worst case bound that is proportional to $\kappa(A)$ if $\kappa(Z)$ is constant.

Let $\Delta Z = Z - \tilde{Q}\tilde{R}$. Assume that $cnu\|\tilde{Q}\|_2\|R\|_2/\|Z\|_2 \le 1/2$. From perturbation theory of singular values

$$\sigma_{min}(\tilde{Q}\tilde{R}) \ge \sigma_{min}(Z) - \|\Delta Z\|_2 > 0$$

and

$$\frac{1}{\sigma_{max}(\tilde{Q})\sigma_{min}(\tilde{R})} \le \frac{1}{\sigma_{min}(\tilde{Q}\tilde{R})}$$
$$\le \frac{1}{\sigma_{min}(Z) - \|\Delta Z\|_2}$$
$$\le \frac{2}{\sigma_{min}(Z)}.$$

Therefore,

$$\|\tilde{R}^{-1}\|_2 \le \frac{2\|\tilde{Q}\|_2}{\sigma_{min}(Z)}.$$

Substituting this into (II.14) along with $\|\tilde{Q}\|_2 \le \|A^{-1/2}\|_2 + \mathcal{O}(u)$ gives the upper bound

$$\|\tilde{Q}^T A\tilde{Q} - I\|_2 \le cmnu\kappa(Z)^2\kappa(A) + \mathcal{O}(u^2).$$

Hence, the loss of orthogonality is at worst proportional to $\kappa(A)$ when $\kappa(Z)$ is constant.

III. STABILITY ANALYSIS OF PRE-CHOLQR

In the CHOLQR algorithm the conditioning of Z plays an important roll in the loss of orthogonality. To remove the dependency on the conditioning of Z one can first use a stable Euclidean QR factorization (such as Householder QR) and apply the CHOLQR algorithm to the computed orthonormal factor. Algorithm 2 summarizes PRE-CHOLQR.

A normwise bound for the loss of orthogonality is obtained by letting $Z = \tilde{Y}$ in Theorem II.1, where $\tilde{Y}$ is close to an orthonormal matrix Y that is a bases for the column space of Z. Then

$$Y\tilde{U}^{-1} = \tilde{Q} + \mathcal{O}(u)$$

Algorithm 2: PRE-CHOLQR	**Algorithm 3:** CHOL-EQR

Algorithm 2: PRE-CHOLQR

Input : $Z \in \mathbb{R}^{m \times n}$, $A \in \mathbb{R}^{m \times m}$ - symmetric positive definite
Output: $Q \in \mathbb{R}^{m \times n}$, $R \in \mathbb{R}^{n \times n}$
$[Y, S] = qr(Z)$;
$[Q, U] = \text{CHOLQR}(A, Y)$;
$R = US$;

Algorithm 3: CHOL-EQR

Input : $Z \in \mathbb{R}^{m \times n}$, $A \in \mathbb{R}^{m \times m}$ - symmetric positive definite
Output: $Q \in \mathbb{R}^{m \times n}$, $R \in \mathbb{R}^{n \times n}$
$C = \text{chol}(A)$;
$W = CZ$;
$[Y, R] = qr(W)$;
$Q = C \backslash Z$;

and

$$\|\tilde{U}^{-1}\|_2 \le \|\tilde{Q}\|_2 + \mathcal{O}(u).$$

This gives the loss of orthogonality bound

$$\|\tilde{Q}^T A \tilde{Q} - I\|_2 \le cmnu\|\tilde{Q}\|_2^2\|A\|_2 + \mathcal{O}(u^2).$$

The representativity error can be obtained by the error bounds

$$Z = \tilde{Y}\tilde{S} + \Delta Z, \quad \|\Delta Z\|_2 \le cmn^2 u\|Z\|_2 \quad \text{(III.1)}$$

$$\tilde{Q}\tilde{U} = \tilde{Y} + \Delta Y, \quad \|\Delta Y\|_2 \le cmu\|\tilde{Q}\|_2\|\tilde{U}\|_2 \quad \text{(III.2)}$$

$$\tilde{R} = \tilde{U}\tilde{S} + \Delta R, \quad \|\Delta R\|_2 \le cmu\|\tilde{U}\|_2\|\tilde{S}\|_2. \quad \text{(III.3)}$$

The second equation corresponds to the error of a triangle solve for multiple columns. Normwise and componentwise bounds follow immediately. Theorem III.1 summarizes these results.

Theorem III.1. *The computed factors, $\tilde{Q}$ and $\tilde{R}$, of algorithm 2 satisfy*

$$\|Z - \tilde{Q}\tilde{R}\|_2 \le cmn^2 u\|\tilde{Q}\|_2\|\tilde{U}\|_2\|\tilde{S}\|_2, \quad \text{(III.4)}$$

$$|Z - \tilde{Q}\tilde{R}| \le cmn^2 u|\tilde{Q}||\tilde{U}||\tilde{S}|, \quad \text{(III.5)}$$

$$\|\tilde{Q}^T A \tilde{Q} - I\|_2 \le cmn^2 u\|\tilde{Q}\|_2^2\|A\|_2 + \mathcal{O}(u^2). \quad \text{(III.6)}$$

IV. Factorization based on Cholesky Factor of A

The factorization based on computing the Cholesky factor of A and converting the problem to the Euclidean case is given in Algorithm 3, which we name CHOL-EQR. Here one computes the Cholesky factor C, then computes the Euclidean QR factorization of CZ. This provides the correct R factor, and one obtains the Q factor by $Q = C^{-1}Y$, where Y is the computed orthonormal factor of the Euclidean QR factorization.

The algorithm consists of four major kernels. We have the following backward error results for each kernel.

$$A = \tilde{C}^T \tilde{C} + \Delta A, \quad s.t. \ \|\Delta A\|_2 = cm^2 u\|A\|_2, \quad \text{(IV.1)}$$

$$\text{and } |\Delta A| = cmu|\tilde{C}^T||\tilde{C}|,$$

$$\tilde{W} + \Delta W^{(1)} = \tilde{C}Z, \quad s.t. \ |\Delta W^{(1)}| \le cmu|\tilde{C}||Z|, \quad \text{(IV.2)}$$

$$\tilde{W} = \tilde{Y}\tilde{R} + \Delta W^{(2)}, \quad s.t. \ \|\Delta W^{(2)}\|_2 \le cmn^2 u\|\tilde{W}\|_2, \quad \text{(IV.3)}$$

$$\text{and } |\Delta W^{(2)}| \le cmn^{3/2} uee^T |\tilde{W}|,$$

$$\tilde{Y} = Y + \Delta Y, \quad s.t. \ \|\Delta Y\|_2 \le cmn^2 u, \quad \text{(IV.4)}$$

$$(\tilde{C} + \Delta C^{(j)})\tilde{Q}_j = \tilde{Y}_j, \quad s.t. \ |\Delta C^{(j)}| \le cmu|\tilde{C}|, \quad \text{(IV.5)}$$

where e is a vector of all ones and Y is an exact orthonormal matrix. Let

$$\Delta C = [\Delta C^{(1)} Q_1, \dots, \Delta C^{(j)} Q_j, \dots, \Delta C^{(n)} Q_n].$$

Then Equation (IV.5) becomes

$$\tilde{C}\tilde{Q} + \Delta C = \tilde{Y}, \quad s.t. \ |\Delta C| \le cmu|\tilde{C}||\tilde{Q}|. \quad \text{(IV.6)}$$

Combining (IV.2), (IV.3), and (IV.6) we have,

$$\tilde{C}Z = \tilde{Y}\tilde{R} + \Delta W^{(1)} + \Delta W^{(2)}$$
$$= (\tilde{C}\tilde{Q} + \Delta C)\tilde{R} + \Delta W^{(1)} + \Delta W^{(2)}. \quad \text{(IV.7)}$$

Solving for $Z - \tilde{Q}\tilde{R}$ gives

$$Z - \tilde{Q}\tilde{R} = \tilde{C}^{-1}\Delta C\tilde{R} + \tilde{C}^{-1}\Delta W^{(1)} + \tilde{C}^{-1}\Delta W^{(2)}.$$

A componentwise bound follows immediately:

$$|Z - \tilde{Q}\tilde{R}| \le cmu|\tilde{C}^{-1}||\tilde{C}||\tilde{Q}||\tilde{R}| + cnu|\tilde{C}^{-1}||\tilde{C}||Z|$$
$$+ cmn^{3/2} u|\tilde{C}^{-1}|ee^T|\tilde{C}Z|$$
$$\le cmu|\tilde{C}^{-1}||\tilde{C}||\tilde{Q}||\tilde{R}|$$
$$+ cmn^{3/2} u|\tilde{C}^{-1}|ee^T|\tilde{C}Z|.$$

A normwise bound also follows:

$$\|Z - \tilde{Q}\tilde{R}\|_2 \le cmu\kappa(\tilde{C})\|\tilde{Q}\|_2\|\tilde{R}\|_2 + cnu\kappa(\tilde{C})\|\tilde{Z}\|_2$$
$$+ cmn^{3/2} u\|\tilde{C}^{-1}\|_2\|\tilde{C}Z\|_2 \quad \text{(IV.8)}$$
$$\le cmn^2 u\kappa(\tilde{C})\|\tilde{Q}\|_2\|\tilde{R}\|_2. \quad \text{(IV.9)}$$

Analyzing the normwise bound, it appears that a worse case upper bound for the loss of representativity is proportional

to $\kappa(A)$. However, in the numerical experiments the dependency on $\kappa(A)$ is not observed (see Section VII). The componentwise bound also fails to provide a descriptive bound on the loss of representativity. This leads us to consider a different measure of representativity error, namely, the A-norm.

Let $\Delta Z = Z - \tilde{Q}\tilde{R}$. From (IV.1), $\Delta Z^T A \Delta Z = \Delta Z^T \tilde{C}^T \tilde{C} \Delta Z + \Delta Z^T \Delta A \Delta Z$. Therefore,

$$\|\Delta Z\|_A^2 \leq \|\tilde{C}\Delta Z\|_2^2 + \|\Delta Z^T \Delta A \Delta Z\|_2.$$

From (IV.7), $\|\tilde{C}\Delta Z\|_2 \leq cmn^2 u\|\tilde{C}\|_2\| \ |\tilde{Q}| \ |\tilde{R}| \ \|_2$. So,

$$\|\Delta Z\|_A^2 \leq (cmn^2 u\|\tilde{C}\|_2\| \ |\tilde{Q}| \ |\tilde{R}| \ \|_2)^2 + \|\Delta Z^T \Delta A \Delta Z\|_2$$

and

$$\|\Delta Z\|_A \leq cmn^2 u\|\tilde{C}\|_2\| \ |\tilde{Q}| \ |\tilde{R}| \ \|_2$$
$$+ \|\Delta Z^T \Delta A \Delta Z\|_2/(cmn^2 u\|\tilde{C}\|_2\| \ |\tilde{Q}| \ |\tilde{R}| \ \|_2)$$

From (IV.1) and (IV.9), we know that $\|\Delta Z^T \Delta Z \Delta Z\|_2 = \mathcal{O}(u^3)$. Which gives us a final error bound of

$$\|\Delta Z\|_A \leq cmn^2 u\|\tilde{C}\|_2\| \ |Q| \ |R| \ \|_2 + \mathcal{O}(u^2).$$

To derive a bound for the loss of orthogonality we begin with (IV.6) and compute

$$\tilde{Q}^T \tilde{C}^T \tilde{C} \tilde{Q} = (\tilde{Y} - \Delta C)^T (\tilde{Y} - \Delta C).$$

Substituting (IV.1) on the left and expanding on the right we have,

$$\tilde{Q}^T (A + \Delta A)\tilde{Q} = \tilde{Y}^T \tilde{Y} - \tilde{Y}^T \Delta C - (\tilde{Y}^T \Delta C)^T + \mathcal{O}(u^2).$$

Substituting (IV.4) for $\tilde{Y}$ and rearranging gives,

$$\tilde{Q}^T A \tilde{Q} - I = Y^T \Delta Y + (Y^T \Delta Y)^T - \tilde{Y}^T \Delta C - (\tilde{Y}^T \Delta C)^T$$
$$- \tilde{Q}^T \Delta A \tilde{Q} + \mathcal{O}(u^2).$$

A normwise bound follows:

$$\|\tilde{Q}^T A \tilde{Q} - I\|_2 \leq cmnu + cmu\|\tilde{Y}^T\|_2\|\tilde{C}\|_2\|\tilde{Q}\|_2$$
$$+ cmu\|\tilde{Q}^T\|_2\|\tilde{C}^T\|_2\|\tilde{Y}^T\|_2$$
$$+ cmu\|\tilde{Q}^T\|_2\|\tilde{C}^T\|_2\|\tilde{C}\|_2\|\tilde{Q}\|_2 + \mathcal{O}(u^2)$$
$$\leq cmu\|\tilde{Q}^T\|_2\|A\|_2\|\tilde{Q}\|_2 + \mathcal{O}(u^2)$$
$$\leq cmu\|A\|_2\|\tilde{Q}\|_2^2 + \mathcal{O}(u^2).$$

We summarize the bounds in the following theorem.

Theorem IV.1. *QR factorization based on the Cholesky factor of A satisfies the following error bounds.*

$$|Z - \tilde{Q}\tilde{R}| \leq cmu|\tilde{C}^{-1}||\tilde{C}||\tilde{Q}||\tilde{R}|$$
$$+ cmn^{3/2}u|\tilde{C}^{-1}|ee^T|\tilde{C}Z|, \quad \text{(IV.10)}$$

$$\|Z - \tilde{Q}\tilde{R}\|_2 \leq cmn^2 u\kappa(\tilde{C})\|\tilde{Q}\|_2\|\tilde{R}\|_2, \quad \text{(IV.11)}$$

$$\|Z - \tilde{Q}\tilde{R}\|_A \leq cmn^2 u\|A\|_2^{1/2}\| \ |\tilde{Q}| \ |\tilde{R}| \ \|_2 + \mathcal{O}(u^2), \quad \text{(IV.12)}$$

$$\|\tilde{Q}^T A \tilde{Q} - I\|_2 \leq cmnu\|A\|_2\|\tilde{Q}\|_2^2 + \mathcal{O}(u^2). \quad \text{(IV.13)}$$

V. FACTORIZATION BASED ON EIGENVALUE DECOMPOSITION OF A

The factorization based on computing the eigenvalue decomposition of A and converting the problem to the Euclidean case is given in Algorithm 4, which we call EQR-SYEV. Here one computes the eigenvalue decomposition $A = VDV^T$, then computes the Euclidean QR factorization of $D^{1/2}V^T Z$. This provides the correct R factor, and we obtain the Q factor by $Q = VD^{-1/2}Y$, where Y has orthonormal columns and is obtained from the Euclidean QR factorization. As explained in the introduction, this is the same procedure as for the Cholesky factorization of A, where the Cholesky factor is replaced by $D^{1/2}V^T$.

Algorithm 4: EQR-SYEV

Input : $Z \in \mathbb{R}^{m \times n}$, $A \in \mathbb{R}^{m \times m}$ - symmetric positive definite

Output: $Q \in \mathbb{R}^{m \times n}$, $R \in \mathbb{R}^{n \times n}$

$[V, D] = \text{eig}(A)$;
$X = V^T Z$;
$W = D^{1/2} X$;
$[Y, R] = \text{qr}(W)$;
$U = D^{-1/2} Y \quad Q = VU$;

The algorithm consists of six major kernels. We have the following backward error results for each kernel.

$$A + \Delta A = \tilde{V}\tilde{D}\tilde{V}^T, \quad s.t. \ \|\Delta A\|_2 \leq cm^{5/2}u\|A\|_2, \quad \text{(V.1)}$$

$$\tilde{V} = V + \Delta E, \quad s.t. \ \|\Delta E\|_2 \leq cm^{5/2}u, \quad \text{(V.2)}$$

$$\tilde{X} = \tilde{V}^T Z + \Delta X, \quad s.t. \ \|\Delta X\|_2 \leq cmu\|Z\|_2, \quad \text{(V.3)}$$

$$\tilde{W} = \tilde{D}^{1/2}(\tilde{X} + \Delta S), \quad s.t. \ \|\Delta S\|_2 \leq cu\|\tilde{X}\|_2, \quad \text{(V.4)}$$

$$\tilde{W} = \tilde{Y}\tilde{R} + \Delta W, \quad s.t. \ \|\Delta W\|_2 \leq cmn^{3/2}u\|\tilde{W}\|_2, \quad \text{(V.5)}$$

$$\tilde{U} = \tilde{D}^{-1/2}(\tilde{Y} + \Delta U), \quad s.t. \ \|\Delta U\|_2 \leq cu\|\tilde{Y}\|_2, \quad \text{(V.6)}$$

$$\tilde{Q} = \tilde{V}\tilde{U} + \Delta Q, \quad s.t. \ \|\Delta Q\|_2 \leq cmu\|\tilde{U}\|_2. \quad \text{(V.7)}$$

Beginning with the backward error result for the Euclidean QR factorization of $\tilde{W}$ (V.5) as well as the backward error result for computing $\tilde{W}$ from the multiplication by a diagonal matrix (V.4) we get

$$\tilde{D}^{1/2}(\tilde{X} + \Delta S) = \tilde{Y}\tilde{R} + \Delta W.$$

Multiply by $\tilde{D}^{-1/2}$ and substituting (V.3) for $\tilde{X}$ and (V.6) for $\tilde{D}^{-1/2}\tilde{Y}$ we get

$$\tilde{V}^T Z + \Delta X + \Delta S = \tilde{U}\tilde{R} - \tilde{D}^{-1/2}\Delta U \tilde{R} + \tilde{D}^{-1/2}\Delta W.$$

Multiplying by $\tilde{V}$ and substituting (V.7) for $\tilde{V}\tilde{U}$ we get

$$(I + \Delta E)Z + \tilde{V}\Delta X + \tilde{V}\Delta S =$$
$$(\tilde{Q} - \Delta Q)\tilde{R} - \tilde{V}\tilde{D}^{-1/2}\Delta U \tilde{R} + \tilde{V}\tilde{D}^{-1/2}\Delta W. \quad \text{(V.8)}$$

Note that $\|\Delta W\|_2 \leq \|\tilde{D}^{1/2}\tilde{V}^T Z\|_2 + \mathcal{O}(u) \leq \|A^{1/2}Z\|_2 + \mathcal{O}(u)$ and $\|A^{1/2}Z\|_2 \leq \|\tilde{R}\|_2 + \mathcal{O}(u)$. Solving for $Z - \tilde{Q}\tilde{R}$ gives the loss representativity error bound

$$\|Z - \tilde{Q}\tilde{R}\|_2 \leq cm^{5/2}u\|A^{-1/2}\|_2\|\tilde{R}\|_2.$$

To calculate the loss of orthogonality we begin with combining (V.6) and (V.7) to get $\tilde{Q} = \tilde{V}D^{-1/2}\tilde{Y} + \tilde{V}\tilde{D}^{-1/2}\Delta U + \Delta Q$. This along with substituting (V.1) for A gives

$$\tilde{Q}^T A \tilde{Q} = (\tilde{V}D^{-1/2}\tilde{Y} + \tilde{V}\tilde{D}^{-1/2}\Delta U + \Delta Q)^T$$
$$\cdot (\tilde{V}\tilde{D}\tilde{V}^T)(\tilde{V}D^{-1/2}\tilde{Y} + \tilde{V}\tilde{D}^{-1/2}\Delta U + \Delta Q).$$

Expanding and simplifying using (V.2) we obtain the bound for loss of orthogonality. The results are summarized in Theorem V.1.

Theorem V.1. *The oblique QR factorization based on the symmetric eigenvalue decomposition of A satisfies the following error bounds.*

$$\|Z - \tilde{Q}\tilde{R}\|_2 \leq cm^{5/2}u\|A^{-1/2}\|_2\|\tilde{R}\|_2, \quad \text{(V.9)}$$
$$\|\tilde{Q}^T A \tilde{Q} - I\|_2 \leq cm^{5/2}u\|A\|_2\|\tilde{Q}\|_2^2 + \mathcal{O}(u^2). \quad \text{(V.10)}$$

The representativity bound is an improvement on the bound presented in [5]. There they give a bound which is proportional to $\kappa(A)^{1/2}\|Z\|_2$, which is clearly an upper bound of (V.9). However, this bound can be an overestimate in some case as we show in Section VII. For completeness we state the result of Rozložník et. al. in Theorem V.2. Note that there is a typographical error in [5] for the representativity bound. The result is correctly displayed in [5, Table 7.1] and in the theorem below.

Theorem V.2. *[5, Thm 2.1] The oblique QR factorization based on the symmetric eigenvalue decomposition of A satisfies the following error bounds.*

$$\|Z - \tilde{Q}\tilde{R}\|_2 \leq cm^{5/2}u\kappa(A)^{1/2}\|Z\|_2, \quad \text{(V.11)}$$
$$\|\tilde{Q}^T A \tilde{Q} - I\|_2 \leq cm^{5/2}u\|A\|_2\|\tilde{Q}\|_2^2 + \mathcal{O}(u^2). \quad \text{(V.12)}$$

VI. Stability of Gram-Schmidt

The Gram-Schmidt algorithms are also possible algorithms for an oblique QR factorization. An updated of the Gram-Schmidt algorithms compute an orthogonal projection. If these projections are computed to be A-orthogonal, then the computed Q is A-orthogonal. The stability of the Gram-Schmidt algorithms was analyzed in detail in [5]. We do not provide any new analysis for these algorithms. For completeness we will state the results from Rozložník et. al. in this section. In the theorems below we state the results as they are stated in [5]. The bounds represent the worst case upper bound. In three cases we provided a slight variation of the bounds. 1. The representativity error in all the Gram-Schmidt algorithms can be expressed as a componentwise bound. 2. The bound for the loss of orthogonality of CGS

can be much less than the worst case bound. A tight bound can be achieved trivially from the previous analysis. 3. The bound for the loss of orthogonality of MGS includes a factor for the maximum growth of the ratio between the 2-norm and A-norm of the updated columns. This factor can be bounded by $\sigma_{min}(A^{1/2}\hat{Z})^{-1}$, where $\hat{Z}$ is an orthonormal bases of Z. Each of these improvements are discussed following the corresponding Theorem.

Classical (CGS) and modified (MGS) Gram-Schmidt both satisfy the same representativity bound and is stated in Theorem VI.1.

Theorem VI.1. *[5, Thm 3.1] The factors $\tilde{Q}$ and $\tilde{R}$ computed by either classical or modified Gram-Schmidt satisfy*

$$Z = \tilde{Q}\tilde{R} + \Delta E, \quad \|\Delta E\|_2 \leq cn^{3/2}u\left(\|Z\|_2 + \|\tilde{Q}\|_2\|\tilde{R}\|_2\right).$$
$$\text{(VI.1)}$$

This is the best normwise bound one can obtain. However, the bound can also be stated as a componentwise bound, which at times can be represent a much smaller error then the corresponding normwise bound. In Theorem VI.1, ΔE satisfies

$$|\Delta E| \leq cn^{3/2}u\left(|Z| + |\tilde{Q}||\tilde{R}|\right). \quad \text{(VI.2)}$$

In Section VII we show that (VI.2) is tight for all of our test cases.

Theorem VI.2. *[5, Thm 4.1] If $cm^{3/2}nu\kappa(A)\kappa(A^{1/2}Z)\kappa(Z) < 1$ then the loss of orthogonality of the computed $\tilde{Q}$ by classical Gram-Schmidt is bounded by*

$$\|\tilde{Q}^T A \tilde{Q} - I\|_2 \leq$$
$$\frac{cm^{3/2}nu\|A\|_2^{1/2}\|\tilde{Q}\|_2\kappa(A^{1/2}Z)\kappa(A)^{1/2}\kappa(Z)}{1 - cm^{3/2}nu\|A\|_2^{1/2}\|\tilde{Q}\|_2\kappa(A^{1/2}Z)\kappa(A)^{1/2}\kappa(Z)}. \quad \text{(VI.3)}$$

The bound in Theorem VI.2 is not tight in all cases. It is easy to adjust the analysis of Rozložník et. al. to provided a tight bound. Also, to simplify the equation we can change the assumption in Theorem VI.2 to $cm^{3/2}nu\kappa(A)\kappa(A^{1/2}Z)\kappa(Z) < 1/2$. With this assumption the denominator is bounded below by $1/2$ and this constant can be absorbed into the arbitrary constant c in the numerator. With this adjustment and an improvement of the numerator, the loss of orthogonality of the computed $\tilde{Q}$ by classical Gram-Schmidt is bounded by

$$\|\tilde{Q}^T A \tilde{Q} - I\|_2 \leq$$
$$cm^{3/2}nu\|A\|_2\|Z\|_2\|\tilde{Q}\|_2\|R^{-1}\|_2\kappa(A^{1/2}Z). \quad \text{(VI.4)}$$

The difference between (VI.4) and (VI.3) is the factor of $\|R^{-1}\|_2$ in (VI.4), which satisfies

$$\|R^{-1}\|_2 \leq \sigma_{min}(A)^{-1/2}\sigma_{min}(Z)^{-1}.$$

This inequality is what gives rise to the factor $\kappa(A^{1/2}Z)\kappa(Z)$ that appears in (VI.3). In Section VII we show that (VI.4) is tight for all of our test cases.

The last two theorems state the results for MGS and classical Gram-Schmidt with reorthogonalization (CGS2). In either case a better bound is not obtained.

Theorem VI.3. *[5, Thm 3.2] If $cm^{3/2}nu\kappa(A)\kappa(A^{1/2}Z) < 1$, then the loss of orthogonality of the computed $\tilde{Q}$ by modified Gram-Schmidt is bounded by*

$$\|\tilde{Q}^T A\tilde{Q} - I\|_2 \leq$$
$$\frac{cm^{3/2}nu\|A\|_2\|\tilde{Q}\|_2 \max_{j\leq i} \frac{\|\tilde{z}_i^{(j-1)}\|_2}{\|\tilde{z}_i^{(j-1)}\|_A}\kappa(A^{1/2}Z)}{1 - cm^{3/2}nu\|A\|_2\|\tilde{Q}\|_2 \max_{j\leq i} \frac{\|\tilde{z}_i^{(j-1)}\|_2}{\|\tilde{z}_i^{(j-1)}\|_A}\kappa(A^{1/2}Z)}.$$
$$\text{(VI.5)}$$

Theorem VI.4. *[5, Thm 5.1,5.2] The factors $\tilde{Q}$ and $\tilde{R}$ computed by classical Gram-Schmidt with reorthogonalization satisfy*

$$Z = \tilde{Q}\tilde{R} + \Delta E, \quad \|A^{1/2}\Delta E\|_2 \leq cn^{3/2}u\|A\|_2^{1/2}\|\tilde{Q}\|_2\|\tilde{R}\|_2,$$
$$\text{(VI.6)}$$

$$\|\tilde{Q}^T A\tilde{Q} - I\|_2 \leq cm^{3/2}nu\|A\|_2\|\tilde{Q}\|_2^2. \qquad \text{(VI.7)}$$

VII. NUMERICAL EXPERIMENTS

In this section we would like to demonstrate that the error bounds presented are descriptive of the true error. All experiments were performed using MATLAB. It will be helpful for this section to recall identities in Theorem I.1. The loss of orthogonality for the most stable algorithms depends on $\|A\|_2\|Q\|_2^2$ and the loss of representativity depends on $\|Q\|_2\|R\|_2$. The test case we develop represent the extreme situations of these bounds. Both of these bounds depend greatly on the column space of Z. The bound for the loss of orthogonality has the following inequality:

$$\frac{\sigma_1(A)}{\sigma_n(A)} \leq \|A\|_2\|Q\|_2^2 \leq \kappa(A).$$

And the bound for the loss of representativity satisfies

$$\|Z\|_2 \leq \|Q\|_2\|R\|_2 \leq \kappa(A^{1/2})\|Z\|_2.$$

If the eigenvector associated with smallest eigenvalue of A is in the column space of Z, then the upper bound on the loss of orthogonality is met (i.e. $\|A\|_2\|Q\|_2^2 = \kappa(A)$). On the other hand if the eigenvectors associated with the n largest eigenvalues are a basis of the column space of Z, then the lower bound is met (i.e. $\frac{\sigma_1(A)}{\sigma_n(A)} = \|A\|_2\|Q\|_2^2$).

To control the behavior of the representativity bound we must be a little more particular on the choice of Z. If the left singular vectors of Z are also eigenvectors of A we can pair eigenvalues of A and singular values of Z as we wish to form singular values of $A^{1/2}Z$, and equivalently R. If we

take the left singular vector of $\sigma_1(Z)$ be the eigenvector associated with $\sigma_1(A)$, then $\|R\|_2 = \|A\|^{1/2}\|Z\|_2$. To obtain the dependency on $\kappa(A)^{1/2}$ we must also have the eigenvector associate with the smallest eigenvalue of A in the column space of Z. With these conditions we have $\|Q\|_2\|R\|_2 = \kappa(A)^{1/2}\|Z\|_2$.

If instead the left singular vectors are the eigenvectors associated with the n largest eigenvalues of A, then $\|Q\|_2 = \sigma_n(A)^{-1/2}$ and $\|Q\|_2\|R\|_2 = \left(\frac{\sigma_1(A)}{\sigma_n(A)}\right)^{1/2}\|Z\|_2$. If the largest eigenvalues are clustered together, the bound in approximately $\|Z\|_2$.

We may also attain a bound approximately $\|Z\|_2$ if we take the left singular vectors of Z to be the eigenvectors of A associated with n smallest eigenvalues. If we also specify that the left singular vector of Z associated with $\sigma_1(Z)$ is the eigenvector of associated with $\sigma_{m-n+1}(A)$, then $\|Q\|_2\|R\|_2 = \left(\frac{\sigma_{m-n+1}(A)}{\sigma_m(A)}\right)^{1/2}\|Z\|_2$. If the smallest eigenvalues are clustered together, the bound is approximately $\|Z\|_2$.

Our test cases exhibit the different situations described. Case 1 we take left singular vectors of Z to be eigenvectors associated with the n smallest eigenvalues of A. This provides a worst case bound for the loss of orthogonality and a best case bound on the representativity error. Case 2 we take left singular vectors of Z to be eigenvectors associated with the n largest eigenvalues of A. This provides a best case bound for both errors. Case 3 we take left singular vectors of Z to be eigenvectors associated with the $\lceil n/2 \rceil$ smallest and $\lfloor n/2 \rfloor$ largest eigenvalues of A. This provides a worst case bound for both errors. Case 4 we take Z to be random. This example indicates what to be expected when Z and A are not correlated.

The cases described thus far allow us to vary $\kappa(A)$ and $\kappa(Z)$ independently. For completeness we also consider the case when Z is A-orthogonal, case 5. This was a test case presented in [5]. To construct such a Z, we again take the left singular vectors of Z to be eigenvectors of A. The associated singular values are then the inverse square root of eigenvalues of A. We chose to use the eigenvectors associated with $\lceil n/2 \rceil$ smallest and $\lfloor n/2 \rfloor$ largest eigenvalues of A. Other choices could have been considered, but we feel this is adequate to demonstrate the bounds.

A is constructed from its eigenvalue decomposition, $A = VDV^T$. $V \in \mathbb{R}^{m \times m}$ is a random orthonormal matrix computed in MATLAB as $V = \text{orth}(\text{randn}(m))$. V is fixed for each problem. The eigenvalues of A are computed such that $\log d_{ii}$ are evenly spaced and $\kappa(A)$ is the largest eigenvalue. That is, if $\alpha = \log(\kappa(A))/(m-1)$, then $d_{ii} = 10^{(\alpha(i-1))}$. Z is constructed from its singular value decomposition, $Z = U\Sigma W^T$. $U \in \mathbb{R}^{m \times n}$ is either a random orthonormal matrix (case 4) or the appropriate columns of V depending on the case. The singular values are also constructed so $\log \sigma_{ii}$ are evenly spaced and $\kappa(Z)$ is the largest singular

value (excluding case 5). In all cases, $W \in \mathbb{R}^{n \times n}$ is taken to be a random orthonormal matrix. Other distributions on the singular values and eigenvalues where considered, but the case presented here provides a good demonstration of the test cases. In all experiments $m = 80$ and $n = 10$. Finally, we let $\kappa(A)$ vary from 10 to 10^{15} and $\kappa(Z) = \kappa(A)^{1/2}$. This allows us to include case 5 and view the correlated dependency.

To display the loss of orthogonality we plot $\|I - Q^T A Q\|_2$. For the loss of representativity we use $\|Z - QR\|_2 / \|Z\|_2$ or if you are using the A-norm, $\|Z - QR\|_A / \|Z\|_A$. In the stability sections the only algorithm which we explicitly state the loss of representativity in terms of th A-norm is $CHOL - EQR$. However, an A-norm bound for the other algorithms can be readily attained from the 2-norm bound. Since $\|\Delta Z\|_A = \|A^{1/2} \Delta Z\|_2 \leq \|A\|_2^{1/2} \|\Delta Z\|_2$.

Figure 1 shows the componentwise versus normwise representativity error for CHOLQR. The solid lines are the true error in each computation and the dotted lines are the corresponding approximate bound (we omit the constants that are dependent on m and n for plotting purposes). This example shows that the normwise bound can be non-descriptive of the true error. The bound for case 3 is proportional to $\kappa(A)^{1/2}$, however the true error shows almost no dependency on $\kappa(A)$.

Case 3 introduces scaling into the columns of Q and the rows of R. Z is numerically in the span of the eigenvectors associated with the $n/2$ largest eigenvalues. Hence, the first $n/2$ columns of Q are also in the span of these eigenvectors and have a 2-norm of about $\|A\|^{-1/2}$. The remaining columns will be in the span of the eigenvectors associated with the smallest eigenvalues and have a 2-norm of about $1/\sigma_m(A)^{1/2}$. The opposite is true of the rows of R. The first $n/2$ rows of R have a 2-norm of about $\|A\|_2^{1/2} \|Z\|_2$ and the last rows are small, relative to the first $n/2$ rows. Therefore, $\| \, |Q| \, |R| \, \|_2 \approx \|Z\|_2$.

Figure 2 shows the representativity error (solid lines) and bounds (dashed lines) for each case and for algorithms CHOL-EQR and SYEV-EQR. We omitted the graphs for CHOLQR, PRE-CHOLQR, CGS, MGS, and CGS2 since they all showed a constant relative error of approximately $10^{-}15$ for both the actual error and the error bounds.

In Figure 2a we see that the 2-norm bounds for CHOL-EQR are non-descriptive for some cases. The bounds for cases 2,3,4 show a dependency on $\kappa(A^{1/2})$ which is not observed in the experiments. If we instead measure the representativity error in the A-norm, then the bound (IV.12) in Theorem IV.1 is descriptive (see Figure 2c). The dependency of $\kappa(A)^{1/2}$ is seen in Case 1 since $\| \, |Q| \, |R| \, \|_2 \approx \|Q\|_2 \|R\|_2$, therefore $\|A\|_2^{1/2} \| \, |Q| \, |R| \, \|_2 / \|Z\|_A \approx \kappa(A)^{1/2}$. For this example we do not vary $\kappa(Z)$ ($\kappa(Z) = 10$).
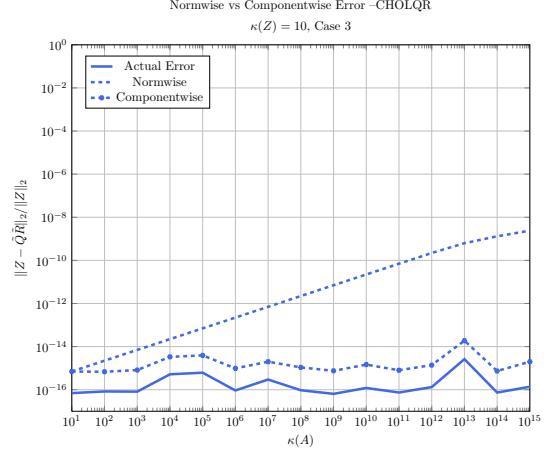

Figure 1. Componentwise versus normwise representativity bound demonstrated with case 3,$\kappa(Z) = 10$, and CHOLQR ($m = 80$ and $n = 10$).

In Figure 2b we see that SYEV-EQR is the only algorithm in which the representativity error is not constant. For cases 2, 3, and 4 the error is proportional to $\kappa(A)^{1/2}$. For cases 2 and 3, this is due to the fact that the eigenvector associated with the largest eigenvalue of A is in the column space of Z. Again, the bounds are descriptive for all cases.

Figure 3 shows the loss of orthogonality for each algorithm and each case. The most stable algorithms are PRE-CHOLQR, CHOL-EQR, SYEV-EQR, and CGS2 (i.e. those with a loss of orthogonality proportional to $\|A\|_2 \|Q\|_2^2$). For these algorithms the loss of orthogonality is independent of the conditioning of Z. Case 1, 3, and 5 all demonstrate the worst case bound of $\kappa(A)$ is obtainable. Case 2 shows that the best case bound of $\sigma_1(A)/\sigma_n(A)$ is obtainable. For our experiments this value is small since we are using a log-linear distribution of the eigenvalues of A. The value could be much larger if $\sigma_1(A)$ was much larger than the other singular values. We also see that for the case when Z is random (case 4) the error tends to follow the case when Z is in the span of the eigenvectors associated with the largest eigenvalues (i.e. the best case).

CHOLQR and CGS are very similar. The best case obtained for these algorithms is worse than any case of the most stable algorithms. Case 2 shows that loss of orthogonality for CHOLQR and CGS are proportional to $\kappa(Z)^2$, since we know that it does not depend on the conditioning of A. Similarly, MGS is proportional to $\kappa(Z)$ in the best case. These observations recover what is known for the Euclidean case ($A = I$).
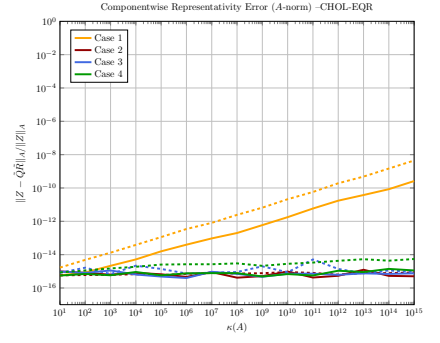
VIII. PERFORMANCE EXPERIMENTS

In this section we consider the performance of the algorithms for tall and skinny matrices ($m \gg n$). Our experiments were performed on an Intel cluster. This machine has 24 compute nodes with each node consisting of two
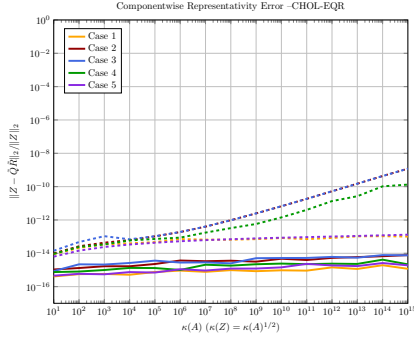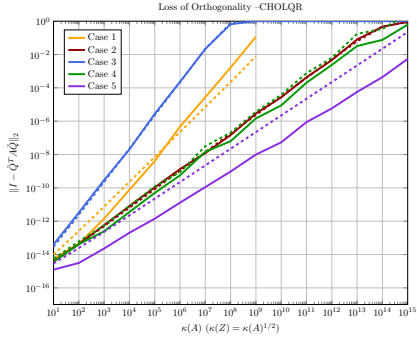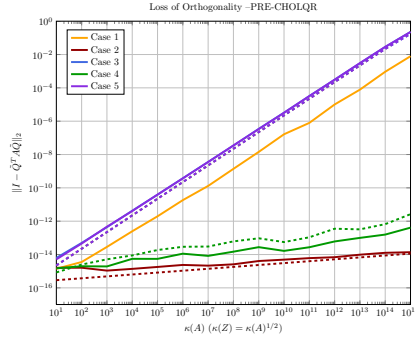
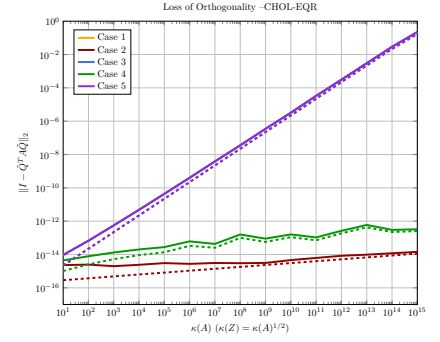(a) CHOL-EQR
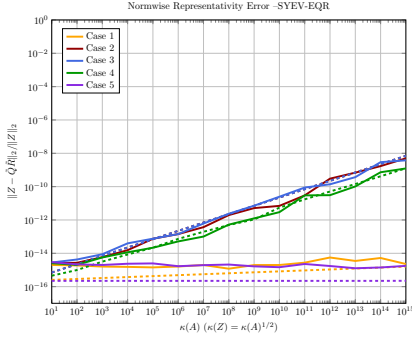
(b) SYEV-EQR

(c) CHOL-EQR (A-norm)

Figure 2. Representativity Error. CHOLQR, PRE-CHOLQR, CGS, MGS, and CGS2 are omitted as they all show constant relative error of approximately 10^{-15}.
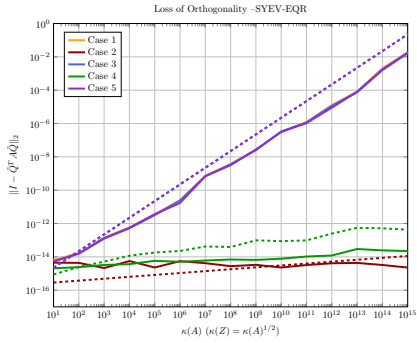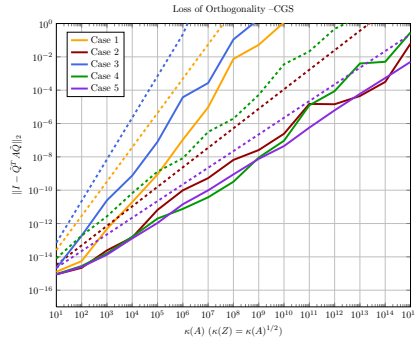

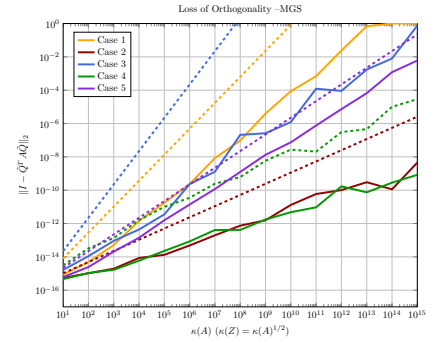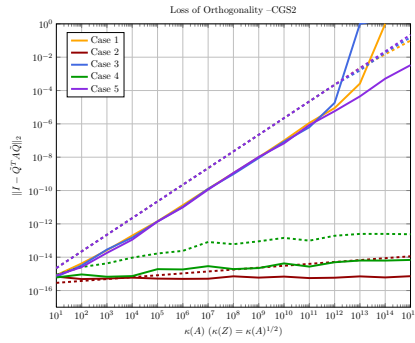(a) CHOLQR

(b) PRE-CHOLQR

(c) CHOL-EQR

(d) SYEV-EQR

(e) CGS

(f) MGS

(g) CGS2

Figure 3. Orthogonality Error

Intel Xeon E5-2670 Sandy Bridge processors running at 3.3 GHz. Thus, each node is a symmetric multiprocessor with 16 cores and 64 GB memory. The theoretical peak is 26.4 GFlops/s per core or 422.4 GFlops/s per node. Each node also has two NVIDIA Tesla M2090 GPUs, however these are not used in the experiments. The nodes are connected with a QDR Infiniband interconnect at 40 Gbit/s. For multi-threaded experiments we link with MKL (11.0) optimized BLAS and LAPACK libraries.

The performance of each algorithm depends on the density of A. To compare the extreme cases we consider the case where A is dense and the case where A is tridiagonal. In both cases we use a normalized FLOP count to plot the FLOP rate. When A is a dense matrix we use $2m^2n + 2mn^2$, which is the number of FLOPs for CHOLQR. When A is tridiagonal matrix we use $2mn^2$, which is again the number of FLOPs for CHOLQR (since the multiplication with A is $\mathcal{O}(mn)$). The time for each algorithm was taken to be the minimum of 10 experiments.

In Figure 4, we show the performance for a for a single node with 16 threads. In Figure 4a, A is a dense matrix and m is fixed at 10000. In this experiment, the both CHOLQR algorithms greatly outperform the other algorithms. CHOLQR reaches a peak performance of 270 GFLOPs/s and PRE-CHOLQR peaks out at about 200 GFLOPs/s. None of the Gram-Schmidt algorithms where above 10 GFLOPs/s. We did not test the CHOL-EQR algorithm since the $\mathcal{O}(m^3)$ cost of the Cholesky factorization of a dense matrix would have been much to costly compared to the other algorithms.

In Figure 4b, A is a tridiagonal matrix and m is fixed at 100000. Again, CHOLQR is outperforms all other algorithms. We do not show the peak performance of CHOLQR (which is about 85 GFLOPs/s) in this figure so we can compare PRE-CHOLQR with the other algorithms. For small n, CGS is able to out perform PRE-CHOLQR, reaching a peak performance of about 7 GFLOPs/s. However, as n increase PRE-CHOLQR continues to improve and eventually out performs all the Gram-Schmidt algorithms. The Cholesky factorization of a tridiagonal matrix is inexpensive and we see for this example that the CHOL-EQR algorithm is a viable option. The performance is fairly similar to PRE-CHOLQR. This is because both algorithms (for small n) are dominated by the Euclidean QR factorization kernel. As n increase the cost of the Cholesky factorization of A becomes a larger overhead than solving the normal equation.

A. Comments

In both plots, there are two implementations of MGS. The difference between the two is the order in which the entries of R are computed. In MGS-col, the entries of R are computed one column at a time. To do this the computation of the inner product (which includes a multiplication with A) is insided the inner most loop. The cost of MGS-col

is $\mathcal{O}(m^2n^2)$ for dense matrices. However, we can also compute the entries of R a row at a time. By doing this, the multiplication of with A can be brought outside the inner most loop and the cost of this algorithm is the same as CGS. The two implementations are equivalent when $A = I$.

IX. CONCLUSION

In this paper, we present stability analysis of many well-known algorithms for performing an oblique QR factorization. We also introduce a new algorithm (PRE-CHOLQR). We find that the most stable algorithms have a loss of orthogonality that is proportional to $\|A\|\|Q\|^2$, which is equivalent to the multiplication error in forming the check, hence this is a best case bound. We also provide a set of test case to asses the tightness of the bounds obtained. From these tests we see that the bounds are tight in all cases.

The new algorithm, PRE-CHOLQR, is stable and can significantly outperform existing stable algorithms. The existing stable algorithm are CGS2, CHOL-EQR, and SYEV-EQR. CGS2 uses level 2 BLAS kernels, which do not perform as well as the level 3 BLAS kernels used in the PRE-CHOLQR algorithm. Both CHOL-EQR and SYEV-EQR require the inner product matrix to be factored. For dense matrices the factorization is too expensive for either algorithm to be efficient. However, for sparse matrices, factoring A may be a viable option.

REFERENCES

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

[2] L. Giraud, J. Langou, M. Rozložník, and J. Eshof, "Rounding error analysis of the classical Gram-Schmidt orthogonalization process," *Numerische Mathematik*, vol. 101, no. 1, pp. 87–100, 2005.

[3] L. Giraud, J. Langou, and M. Rozložník, "The loss of orthogonality in the Gram-Schmidt orthogonalization process," *Computers & Mathematics with Applications*, vol. 50, no. 7, pp. 1069–1075, 2005.

[4] M. Gulliksson, "Backward error analysis for the constrained and weighted linear least squares problem when using the weighted qr factorization," *SIAM journal on matrix analysis and applications*, vol. 16, no. 2, pp. 675–687, 1995.

[5] M. Rozložníık, M. Tůma, A. Smoktunowicz, and J. Kopal, "Numerical stability of orthogonalization methods with a non-standard inner product," *BIT Numerical Mathematics*, pp. 1–24, 2012.

[6] A. Björck, *Numerical methods for least squares problems*. Society for Industrial Mathematics, 1996, no. 51.

[7] S. Thomas and R. Zahar, "Efficient orthogonalization in the M-norm," *Congressus Numerantium*, vol. 80, pp. 23–23, 1991.

[8] ——, "An analysis of orthogonalization in elliptic norms," *Congressus numerantium*, vol. 86, pp. 193–193, 1992.

[9] S. Thomas, "A block algorithm for orthogonalization in elliptic norms," *Lecture Notes in Computer Science*, vol. 634, pp. 379–385, 1992.
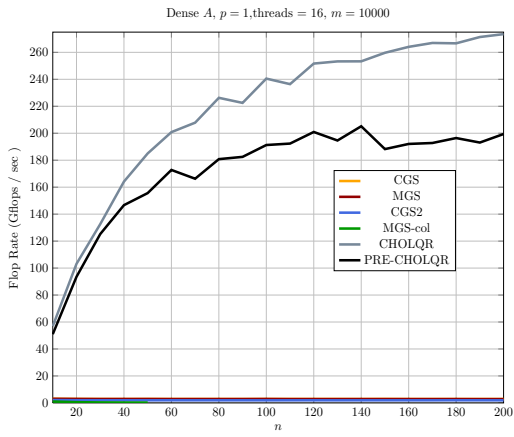
(a) A is a dense matrix with $m = 10000$.
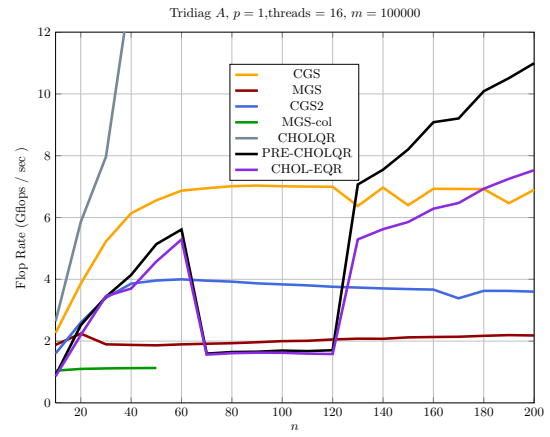
(b) A is a tridiagonal matrix with $m = 100000$.

Figure 4. Performance results on 1 node with 16 threads.