

# ROUNDING ERROR ANALYSIS OF MIXED-PRECISION HOUSEHOLDER QR ALGORITHMS

L. MINAH YANG, ALYSON FOX, AND GEOFFREY SANDERS

**Abstract.** Although mixed precision arithmetic has recently garnered interest for training dense neural networks, many other applications could benefit from the speed-ups and lower storage if applied appropriately. The growing interest in employing mixed precision computations motivates the need for rounding error analysis that properly handles behavior from mixed precision arithmetic. We present a framework for mixed precision analysis that builds on the foundations of rounding error analysis presented in [13] and demonstrate its practicality by applying the analysis to various Householder QR Algorithms.

**1. Introduction.** The accuracy of a numerical algorithm depends on several factors, including numerical stability and well-conditionedness of the problem, both of which may be sensitive to rounding errors, the difference between exact and finite-precision arithmetic. Low precision floats use fewer bits than high precision floats to represent the real numbers and naturally incur larger rounding errors. Therefore, error attributed to round-off may have a larger influence over the total error when using low precision, and some standard algorithms may yield insufficient accuracy when using low precision storage and arithmetic. However, many applications exist that would benefit from the use of lower precision arithmetic and storage that are less sensitive to floating-point round off error, such as clustering or ranking graph algorithms [21] or training dense neural networks [18], to name a few.

Many computing applications today require solutions quickly and often under low size, weight, and power constraints (low SWaP), e.g., sensor formation, etc. Computing in low-precision arithmetic offers the ability to solve many problems with improvement in all four parameters. Utilizing mixed-precision, one can achieve similar quality of computation as high-precision and still achieve speed, size, weight, and power constraint improvements. There have been several recent demonstrations of computing using half-precision arithmetic (16 bits) achieving around half an order to an order of magnitude improvement of these categories in comparison to double precision (64 bits). Trivially, the size and weight of memory required for a specific problem is  $4\times$ . Additionally, there exist demonstrations that the power consumption improvement is similar [9]. Modern accelerators (e.g., GPUs, Knights Landing, or Xeon Phi) are able to achieve this factor or better speedup improvements. Several examples include: (i)  $2\text{--}4\times$  speedup in solving dense large linear equations [11, 12], (ii)  $12\times$  speedup in training dense neural networks, and (iii)  $1.2\text{--}10\times$  speedup in small batched dense matrix multiplication [1] (up to  $26\times$  for batches of tiny matrices). Training deep artificial neural networks by employing lower precision arithmetic to various tasks such as multiplication [5] and storage [6] can easily be implemented on GPUs and are already a common practice in data science applications.

The low precision computing environments that we consider are *mixed precision* settings, which are designed to imitate those of new GPUs that employ multiple precision types for certain tasks. For example, Tesla V100's Tensor Cores perform matrix-multiply-and-accumulate of half precision input data with exact products and single precision (32 bits) summation accumulate [3]. The existing rounding error analyses are built within what we call a *uniform precision* setting, which is the assumption that all arithmetic operations and storage are performed via the same precision. In

---

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and was supported by the LLNL-LDRD Program under Project No. 17-SI-004, LLNL-JRNL-795525-DRAFT.

42 this work, we develop a framework for deterministic mixed-precision rounding error analysis, and  
 43 explore half-precision Householder QR factorization (HQR) algorithms for data and graph analysis  
 44 applications. QR factorization is known to provide a backward stable solution to the linear least  
 45 squares problem and thus, is ideal for mixed-precision.

46 However, additional analysis is needed as the additional round-off error will effect orthogonality,  
 47 and thus the accuracy of the solution. Here, we focus on analyzing specific algorithms in a specific  
 48 set of types (IEEE754 half (fp16), single (fp32, and double(fp64)), but the framework we develop  
 49 could be used on different algorithms or different floating point types (such as bfloat16 in [20]).

50 This work discusses several aspects of using mixed-precision arithmetic: (i) error analysis that  
 51 can more accurately describe mixed-precision arithmetic than existing analyses, (ii) algorithmic de-  
 52 sign that is more resistant against lower numerical stability associated with lower precision types,  
 53 and (iii) an example where mixed-precision implementation performs as sufficiently as double-  
 54 precision implementations. Our key findings are that the new mixed-precision error analysis pro-  
 55 duces tighter error bounds, that some block QR algorithms by Demmel et al. [8] are able to operate  
 56 in low precision more robustly than non-block techniques, and that some small-scale benchmark  
 57 graph clustering problems can be successfully solved with mixed-precision arithmetic.

**2. Background: Build up to rounding error analysis for inner products.** In this  
 section, we introduce the basic motivations and tools for mixed-precision rounding error analysis  
 needed for the *QR factorization*. A matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  for  $m \geq n$  can be written as

$$\mathbf{A} = \mathbf{Q}\mathbf{R}, \quad \mathbf{Q} \in \mathbb{R}^{m \times m}, \quad \mathbf{R} \in \mathbb{R}^{m \times n},$$

58 where  $\mathbf{Q}$  is orthogonal,  $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_{m \times m}$ , and  $\mathbf{R}$  is upper trapezoidal. The above formulation is a  
 59 *full* QR factorization, whereas a more efficient *thin* QR factorization results in  $\mathbf{Q}_1 \in \mathbb{R}^{m \times n}$  and  
 60  $\mathbf{R}_1 \in \mathbb{R}^{n \times n}$ , that is

$$61 \quad \mathbf{A} = \mathbf{Q}\mathbf{R} = [\mathbf{Q}_1 \quad \mathbf{Q}_2] \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0}_{m-n \times n} \end{bmatrix} = \mathbf{Q}_1 \mathbf{R}_1.$$

62 If  $\mathbf{A}$  is full rank then the columns of  $\mathbf{Q}_1$  are orthonormal (i.e.  $\mathbf{Q}_1^\top \mathbf{Q}_1 = \mathbf{I}_{n \times n}$ ) and  $\mathbf{R}_1$  is upper  
 63 triangular. In many applications, computing the *thin* decomposition requires less computation and  
 64 is sufficient in performance. While important definitions are stated explicitly in the text, Table 1  
 65 serves to establish basic notation.

Symbol(s)	Definition(s)	Section(s)
$\mathbf{x}, \mathbf{A}$	Vector, matrix	2
$\mathbf{Q}$	Orthogonal factor $\mathbf{A} \in \mathbb{R}^{m \times n}$ : $m$ -by- $m$ (full) or $m$ -by- $n$ (thin)	2
$\mathbf{R}$	Upper triangular or trapezoidal factor of $\mathbf{A} \in \mathbb{R}^{m \times n}$ : $m$ -by- $n$ (full) or $n$ -by- $n$ (thin)	2
$\text{fl}(\mathbf{x}), \hat{\mathbf{x}}$	Quantity $\mathbf{x}$ calculated from floating point operations	2.1
$b, t, \mu, \eta$	Base/precision/mantissa/exponent bits	2.1
$k$	Number of successive FLOPs	2.1
$u^q$	Unit round-off for precision $t_q$ and base $b_q$ : $\frac{1}{2}b_q^{1-t_q}$	2.1
$\delta^q$	Quantity bounded by: $ \delta^q  < u^q$	2.1
$\gamma_k^q, \theta_k^q$	$\frac{ku^q}{1-ku^q}$ , Quantity bounded by: $ \theta_k^q  \leq \gamma_k^q$	2.1

TABLE 1  
Basic definitions

Subsection 2.1 introduces basic concepts for rounding error analysis, and Subsection 2.2 exemplifies the need for mixed-precision rounding error analysis using the inner product.

**2.1. Basic rounding error analysis of floating point operations.** We use and analyze the IEEE 754 Standard floating point number systems. Let  $\mathbb{F} \subset \mathbb{R}$  denote the space of some floating point number system with base  $b \in \mathbb{N}$ , precision  $t \in \mathbb{N}$ , significand  $\mu \in \mathbb{N}$ , and exponent range  $[\eta_{\min}, \eta_{\max}] \subset \mathbb{Z}$ . Then every element  $y$  in  $\mathbb{F}$  can be written as

$$(2.1) \quad y = \pm \mu \times b^{\eta-t},$$

where  $\mu$  is any integer in  $[0, b^t - 1]$  and  $\eta$  is an integer in  $[\eta_{\min}, \eta_{\max}]$ . While base, precision, and exponent range are fixed and define a floating point number, the sign, significand, and exponent identifies a unique number within that system. Although operations we use on  $\mathbb{R}$  cannot be replicated exactly due to the finite cardinality of  $\mathbb{F}$ , we can still approximate the accuracy of analogous floating point operations (FLOPs). We adopt the rounding error analysis tools described in [13], which allow a relatively simple framework for formulating error bounds for complex linear algebra operations. A short analysis of FLOPs (see Theorem 2.2 [13]) shows that the relative error is controlled by the unit round-off,  $u := \frac{1}{2}b^{1-t}$ .

Name	$b$	$t$	# of exponent bits	$\eta_{\min}$	$\eta_{\max}$	unit round-off $u$
fp16 (IEEE754 half)	2	11	5	-15	16	4.883e-04
fp32 (IEEE754 single)	2	24	8	-127	128	5.960e-08
fp64 (IEEE754 double)	2	53	11	-1023	1024	1.110e-16

TABLE 2  
IEEE754 formats and their primary attributes.

Let ‘op’ be any basic operation from the set  $\text{OP} = \{+, -, \times, \div\}$  and let  $x, y \in \mathbb{R}$ . The true value  $(x \text{ op } y)$  lies in  $\mathbb{R}$ , and it is rounded using some conversion to a floating point number,  $\text{fl}(x \text{ op } y)$ , admitting a rounding error. The IEEE 754 Standard requires *correct rounding*, which rounds the exact solution  $(x \text{ op } y)$  to the closest floating point number and, in case of a tie, to the floating point number that has a mantissa ending in an even number. *Correct rounding* gives us an assumption for the error model where a single basic floating point operation yields a relative error,  $\delta$ , bounded in the following sense:

$$(2.2) \quad \text{fl}(x \text{ op } y) = (1 + \delta)(x \text{ op } y), \quad |\delta| \leq u, \quad \text{op} \in \{+, -, \times, \div\}.$$

We use (2.2) as a building block in accumulating errors from successive FLOPs. For example, consider computing  $x + y + z$ , where  $x, y, z \in \mathbb{R}$  with a machine that can only compute one operation at a time. Then, there is a rounding error in computing  $\hat{s}_1 := \text{fl}(x + y) = (1 + \delta)(x + y)$ , and another rounding error in computing  $\hat{s}_2 := \text{fl}(\hat{s}_1 + z) = (1 + \tilde{\delta})(\hat{s}_1 + z)$ , where  $|\delta|, |\tilde{\delta}| < u$ . Then,

$$(2.3) \quad \text{fl}(x + y + z) = (1 + \tilde{\delta})(1 + \delta)(x + y) + (1 + \tilde{\delta})z.$$

Multiple successive operations introduce multiple rounding error terms, and keeping track of all errors is challenging. Lemma 2.1 introduces a convenient and elegant bound that simplifies accumulation of rounding error.

97 LEMMA 2.1 (Lemma 3.1 [13]). Let  $|\delta_i| < u$  and  $\rho_i \in \{-1, +1\}$ , for  $i = 1, \dots, k$  and  $ku < 1$ .  
 98 Then,

$$99 \quad (2.4) \quad \prod_{i=1}^k (1 + \delta_i)^{\rho_i} = 1 + \theta_k, \quad \text{where} \quad |\theta_k| \leq \frac{ku}{1 - ku} =: \gamma_k.$$

100 We also use

$$101 \quad \tilde{\gamma}_k = \frac{cku}{1 - cku},$$

102 where  $c > 0$  is a small integer and further extend this to  $\theta$  so that  $|\tilde{\theta}_k| \leq \tilde{\gamma}_k$ .

103 In other words,  $\theta_k$  represents the accumulation of rounding errors from  $k$  successive operations, and  
 104 it is bounded by  $\gamma_k$ . Allowing  $\theta_k$ 's to be any arbitrary value within the corresponding  $\gamma_k$  bounds  
 105 further aids in keeping a clear, simple error analysis. Applying this lemma to our example of adding  
 106 three numbers results in

$$107 \quad (2.5) \quad \text{fl}(x + y + z) = (1 + \tilde{\delta})(1 + \delta)(x + y) + (1 + \tilde{\delta})z = (1 + \theta_2)(x + y) + (1 + \theta_1)z.$$

108 Since  $|\theta_1| \leq \gamma_1 < \gamma_2$ , we can further simplify (2.5) to

$$109 \quad (2.6) \quad \text{fl}(x + y + z) = (1 + \tilde{\theta}_2)(x + y + z), \quad \text{where} \quad |\tilde{\theta}_2| \leq \gamma_2,$$

110 at the cost of a slightly larger upper bound. Typically, error bounds formed in the fashion of (2.6)  
 111 are converted to relative errors in order to put the error magnitudes in perspective. The relative  
 112 error bound for our example is

$$113 \quad \frac{|(x + y + z) - \text{fl}(x + y + z)|}{|x + y + z|} \leq \gamma_2$$

114 when we assume  $x + y + z \neq 0$ .

115 Although Lemma 2.1 requires  $ku < 1$ , we actually need  $ku < \frac{1}{2}$  to maintain a meaningful  
 116 relative error bound as this assumption implies  $\gamma_k < 1$  and guarantees a relative error below 100%.  
 117 Since higher precision floating points have smaller unit round-off values, they can tolerate more  
 118 successive FLOPs than lower precision floating points before reaching  $\gamma_m = 1$ . Table 3 shows the  
 119 maximum number of successive floating point operations that still guarantees a relative error below  
 100% for various floating point types. Thus, accumulated rounding errors in lower precision types

precision	$\tilde{k} = \arg \max_k (\gamma_k \leq 1)$
FP16	512
FP32	$\approx 4.194\text{e}06$
FP64	$\approx 2.252\text{e}15$

TABLE 3  
Upper limits of meaningful relative error bounds in the  $\gamma^{(k)}$  notation.

120 can lead to an instability with fewer operations in comparison to higher precision types and prompts  
 122 us to evaluate whether existing algorithms can be naively adapted for mixed-precision arithmetic.

**2.2. Rounding Error Example for the Inner Product.** We now consider computing the inner product of two vectors to clearly illustrate how this situation restricts rounding error analysis in fp16. An error bound for an inner product of  $m$ -length vectors is

$$(2.7) \quad |\mathbf{x}^\top \mathbf{y} - \text{fl}(\mathbf{x}^\top \mathbf{y})| \leq \gamma_m |\mathbf{x}|^\top |\mathbf{y}|, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^m$$

as shown in [13]. While this result does not guarantee a high relative accuracy when  $|\mathbf{x}^\top \mathbf{y}| \ll |\mathbf{x}|^\top |\mathbf{y}|$ , high relative accuracy is expected in some special cases. For example, let  $\mathbf{x} = \mathbf{y}$ . Then we have exactly  $|\mathbf{x}^\top \mathbf{x}| = |\mathbf{x}|^\top |\mathbf{x}| = \|\mathbf{x}\|_2^2$ , which leads to a forward error:  $|\|\mathbf{x}\|_2^2 - \text{fl}(\|\mathbf{x}\|_2^2)| \leq \gamma_m \|\mathbf{x}\|_2^2$ . Since vectors of length  $m$  accumulate rounding errors that are bounded by  $\gamma_m$ , the dot products of vectors computed in fp16 already face a 100% relative error bound in the worst-case scenario ( $\gamma_{512}^{\text{fp16}} = 1$ ).

We present a simple numerical experiment that shows that the standard deterministic error bound is too pessimistic and cannot be practically used to approximate rounding error for half-precision arithmetic. In this experiment, we generated 2 million random half-precision vectors of length 512 from two random distributions: the standard normal distribution,  $N(0,1)$ , and the uniform distribution over  $(0,1)$ . Half precision arithmetic was simulated by calling `alg. 1`, which was proven to be a faithful simulation in [15], for every FLOP (multiplication and addition for the dot product). The relative error in this experiment is formulated as the LHS in Equation 2.7 divided by  $|\mathbf{x}|^\top |\mathbf{y}|$  and all operations outside of calculating  $\text{fl}(\mathbf{x}^\top \mathbf{y})$  are executed by casting up to fp64 and using fp64 arithmetic. Table 4 shows some statistics from computing the relative error for simulated half precision dot products of 512-length random vectors. We see that the inner products of vectors sampled from the standard normal distribution have backward relative errors that do not deviate much from the unit round-off ( $\mathcal{O}(1\text{e-}4)$ ), whereas the vectors sampled from the uniform distribution tend to accumulate larger errors on average ( $\mathcal{O}(1\text{e-}3)$ ). Even so, the theoretical upper error bound of 100% is far too pessimistic as the maximum relative error does not even meet 2% in this experiment. Recent work in developing probabilistic bounds on rounding errors of floating point operations (see [14, 17]) have shown that the inner product relative backward error for the conditions used for this experiment is bounded by  $5.466\text{e-}2$  with probability 0.99.

---

**Algorithm 1:**  $\mathbf{z}^{\text{fp16}} = \text{simHalf}(f, \mathbf{x}^{\text{fp16}}, \mathbf{y}^{\text{fp16}})$ . Simulate function  $f \in \text{OP} \cup \{\text{dot\_product}\}$  in half precision arithmetic given input variables  $\mathbf{x}, \mathbf{y}$ . Function `castup` converts half precision floats to single precision floats, and `castdown` converts single precision floats to half precision floats by rounding to the nearest half precision float.

---

**Input:**  $\mathbf{x}^{\text{fp16}}, \mathbf{y}^{\text{fp16}} \in \mathbb{F}_{\text{fp16}}^m, f : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}^n$

**Output:**  $\text{fl}(f(\mathbf{x}^{\text{fp16}}, \mathbf{y}^{\text{fp16}})) \in \mathbb{F}_{\text{fp16}}^n$

```

1  $\mathbf{x}^{\text{fp32}}, \mathbf{y}^{\text{fp32}} \leftarrow \text{castup}([\mathbf{x}^{\text{fp16}}, \mathbf{y}^{\text{fp16}}])$ 
2  $\mathbf{z}^{\text{fp32}} \leftarrow \text{fl}(f(\mathbf{x}^{\text{fp32}}, \mathbf{y}^{\text{fp32}}))$ 
3  $\mathbf{z}^{\text{fp16}} \leftarrow \text{castdown}(\mathbf{z}^{\text{fp32}})$ 
4 return  $\mathbf{z}^{\text{fp16}}$ 

```

---

Most importantly, no rounding error bounds (deterministic or probabilistic) allow flexibility in the precision types used for different operations. This restriction is the biggest obstacle in gaining an understanding of rounding errors to expect from computations done on emerging hardware that support mixed-precision such as GPUs that employ mixed-precision arithmetic.

Random Distribution	Average	Standard deviation	Maximum
Standard normal	1.627e-04	1.640e-04	2.838e-03
Uniform (0, 1)	2.599e-03	1.854e-03	1.399e-02

TABLE 4

Statistics from dot product backward relative error in for 512-length vectors stored in half-precision and computed in simulated half-precision from 2 million realizations.

We start by introducing some additional rules from [13] that build on Lemma 2.1 in Lemma 2.2. These rules summarize how to accumulate errors represented by  $\theta$ 's and  $\gamma$ 's in a *uniform precision* setting. These relations aid in writing clear and simpler error analyses. Regardless of the specific details of a mixed-precision setting, a rounding error analysis for mixed-precision arithmetic must support at least two different precision types. Thus, Lemma 2.3 allows low and high precision types and is a simple modification of Lemma 2.2. The rules for  $\theta$  allows us to keep track of the two precision types separately and the rules we present for  $\gamma$  were chosen to be useful for casting down to the lower of the two precisions, a pertinent procedure in our mixed-precision analysis in the later sections.

LEMMA 2.2. For any positive integer  $k$ , let  $\theta_k$  denote a quantity bounded according to  $|\theta_k| \leq \frac{ku}{1-ku} =: \gamma_k$ . The following relations hold for positive integers  $i, j$ , and nonnegative integer  $k$ . Arithmetic operations between  $\theta_k$ 's:

$$(2.8) \quad (1 + \theta_k)(1 + \theta_j) = (1 + \theta_{k+j}) \quad \text{and} \quad \frac{1 + \theta_k}{1 + \theta_j} = \begin{cases} 1 + \theta_{k+j}, & j \leq k \\ 1 + \theta_{k+2j}, & j > k \end{cases}$$

Operations on  $\gamma$ 's:

$$\begin{aligned} \gamma_k \gamma_j &\leq \gamma_{\min(k,j)}, \quad \text{for } \max_{(j,k)} u \leq \frac{1}{2}, \\ n\gamma_k &\leq \gamma_{nk}, \quad \text{for } n \leq \frac{1}{uk}, \\ \gamma_k + u &\leq \gamma_{k+1}, \\ \gamma_k + \gamma_j + \gamma_k \gamma_j &\leq \gamma_{k+j}. \end{aligned}$$

LEMMA 2.3. For any nonnegative integer  $k$  and some precision  $q$ , let  $\theta_k^q$  denote a quantity bounded according to  $|\theta_k^q| \leq \frac{ku^q}{1-ku^q} =: \gamma_k^q$ . The following relations hold for two precisions  $l$  (low) and  $h$  (high), positive integers,  $j_l, j_h$ , non-negative integers  $k_l$ , and  $k_h$ , and  $c > 0$ :

$$(2.9) \quad (1 + \theta_{k_l}^l)(1 + \theta_{j_l}^l)(1 + \theta_{k_h}^h)(1 + \theta_{j_h}^h) = (1 + \theta_{k_l+j_l}^l)(1 + \theta_{k_h+j_h}^h),$$

$$(2.10) \quad \frac{(1 + \theta_{k_l}^l)(1 + \theta_{j_l}^l)}{(1 + \theta_{j_l}^l)(1 + \theta_{j_h}^h)} = \begin{cases} (1 + \theta_{k_h+j_h}^h)(1 + \theta_{k_l+j_l}^l), & j_h \leq k_h, j_l \leq k_l, \\ (1 + \theta_{k_h+2j_h}^h)(1 + \theta_{k_l+j_l}^l), & j_h \leq k_h, j_l > k_l, \\ (1 + \theta_{k_h+j_h}^h)(1 + \theta_{k_l+2j_l}^l), & j_h > k_h, j_l \leq k_l, \\ (1 + \theta_{k_h+2j_h}^h)(1 + \theta_{k_l+2j_l}^l), & j_h > k_h, j_l > k_l. \end{cases}$$

Without loss of generality, let  $1 \gg u_l \gg u_h > 0$ . Let  $d$ , a nonnegative integer, and  $r \in [0, \lfloor \frac{u_l}{u_h} \rfloor]$  be numbers that satisfy  $k_h u_h = d u_l + r u_h$ . Alternatively,  $d$  can be defined by  $d := \lfloor \frac{k_h u_h}{u_l} \rfloor$ . Then,

$$(2.11) \quad \gamma_{k_h}^h \gamma_{k_l}^l \leq \gamma_{k_l}^l, \quad \text{for } k_l u^l \leq \frac{1}{2}$$

$$(2.12) \quad \gamma_{k_h}^h + u^l \leq \gamma_{d+2}^l$$

$$(2.13) \quad \gamma_{k_l}^l + u^h \leq \gamma_{k_l+1}^l$$

$$(2.14) \quad \gamma_{k_l}^l + \gamma_{k_h}^h + \gamma_{k_l}^l \gamma_{k_h}^h < \gamma_{k_l+d+1}^l.$$

We use these principles to establish a mixed-precision rounding error analysis for computing the dot product, which is crucial in many linear algebra routines such as the QR factorization. Let us define a mixed-precision setting that is similar to the TensorCore Fused Multiply-Add (FMA) block but works at the level of a dot product. While the FMA block in TensorCore is for matrix-matrix products (level-3 BLAS), we consider a vector inner product (level-2 BLAS) FMA as defined in Assumption 2.4.

**ASSUMPTION 2.4.** Let  $l$  and  $h$  each denote low and high precision types with unit round-off values  $u^l$  and  $u^h$ , where  $1 \gg u^l \gg u^h > 0$ . Consider an FMA operation for inner products that take vectors stored in precision  $l$ , compute products in full precision, and sum the products in precision  $h$ . Finally, the result is then cast back down to precision  $l$ .

The full precision multiplication in Assumption 2.4 is exact when the low precision type is fp16 and the high precision type of fp32 due to their precisions and exponent ranges. As a quick proof, consider  $x^{\text{fp16}} = \pm \mu_x 2^{\eta_x - 11}$ ,  $y^{\text{fp16}} = \pm \mu_y 2^{\eta_y - 11}$  where  $\mu_x, \mu_y \in [0, 2^{11} - 1]$  and  $\eta_x, \eta_y \in [-15, 16]$ , and note that the significand and exponent range for fp32 are  $[0, 2^{24} - 1]$  and  $[-127, 128]$ . Then the product in full precision is

$$x^{\text{fp16}} y^{\text{fp16}} = \pm \mu_x \mu_y 2^{\eta_x + \eta_y + 2 - 24},$$

where  $\mu_x \mu_y \in [0, (2^{11} - 1)^2] \subseteq [0, 2^{24} - 1]$  and  $\eta_x + \eta_y + 2 \in [-28, 34] \subseteq [-127, 128]$ , and therefore is exact. Thus, the summation and the final cast down operations are the only sources of rounding error.

Let  $\mathbf{x}^{\text{fp16}}, \mathbf{y}^{\text{fp16}}$  be  $m$ -length vectors stored in fp16,  $s_k$  be the  $k^{\text{th}}$  partial sum, and  $\hat{s}_k$  be  $s_k$  computed with FLOPs. Then,

$$\begin{aligned} \hat{s}_1 &= \text{fl}(\mathbf{x}_1 \mathbf{y}_1) = \mathbf{x}_1 \mathbf{y}_1, \\ \hat{s}_2 &= \text{fl}(\hat{s}_1 + \mathbf{x}_2 \mathbf{y}_2) = (\mathbf{x}_1 \mathbf{y}_1 + \mathbf{x}_2 \mathbf{y}_2) (1 + \delta_1^h), \\ \hat{s}_3 &= \text{fl}(\hat{s}_2 + \mathbf{x}_3 \mathbf{y}_3) = [(\mathbf{x}_1 \mathbf{y}_1 + \mathbf{x}_2 \mathbf{y}_2) (1 + \delta_1^h) + \mathbf{x}_3 \mathbf{y}_3] (1 + \delta_2^h). \end{aligned}$$

We can see a pattern emerging. The error for a general  $m$ -length vector dot product is then

$$(2.15) \quad \hat{s}_m = (\mathbf{x}_1 \mathbf{y}_1 + \mathbf{x}_2 \mathbf{y}_2) \prod_{k=1}^{m-1} (1 + \delta_k^h) + \sum_{i=3}^n \mathbf{x}_i \mathbf{y}_i \left( \prod_{k=i-1}^{m-1} (1 + \delta_k^h) \right).$$

Using Lemma 2.1, we further simplify and form componentwise backward errors with

$$(2.16) \quad \text{fl}(\mathbf{x}^\top \mathbf{y}) = (\mathbf{x} + \Delta \mathbf{x})^\top \mathbf{y} = \mathbf{x}^\top (\mathbf{y} + \Delta \mathbf{y}), \quad \text{for } |\Delta \mathbf{x}| \leq \gamma_{m-1}^h |\mathbf{x}|, \quad |\Delta \mathbf{y}| \leq \gamma_{m-1}^h |\mathbf{y}|.$$



210 Casting this down to fp16, then we incur a rounding error quantified by  $d := \lfloor \frac{(m-1)u^h}{u^l} \rfloor$ . The  
 211 resulting componentwise backward errors are

$$212 \quad (2.17) \quad \text{fl}(\mathbf{x}^\top \mathbf{y}) = (\mathbf{x} + \Delta \mathbf{x})^\top \mathbf{y} = \mathbf{x}^\top (\mathbf{y} + \Delta \mathbf{y}), \quad \text{for } |\Delta \mathbf{x}| \leq \gamma_{d+1}^l |\mathbf{x}|, \quad |\Delta \mathbf{y}| \leq \gamma_{d+1}^l |\mathbf{y}|.$$

213 Equations (2.16) and (2.17) are crucial for our analysis in section 4 since the TensorCore  
 214 technology outputs a matrix product in fp16 or fp32. Consider matrices  $\mathbf{A} \in \mathbb{F}_{\text{fp16}}^{p \times m}$  and  $\mathbf{B} \in \mathbb{F}_{\text{fp16}}^{m \times q}$ ,  
 215 and  $\mathbf{C} = \mathbf{AB} \in \mathbb{F}_{\text{fp16}}^{p \times q}$ . If  $\text{fl}(\mathbf{C})$  is desired in fp16, then each component of that matrix incurs  
 216 rounding errors as shown in (2.17) and if it is desired in fp32, the componentwise rounding error  
 217 is given by (2.16). Similarly, we could consider other mixed-precision algorithms that cast down at  
 218 various points within the algorithm to take advantage of better storage properties of lower precision  
 219 types. Error bounds in the fashion of (2.16) can be used before the cast down operations, and the  
 220 action of the cast down is best represented by error bounds similar to (2.17).

221 In section 3, we introduce various Householder QR algorithms as well as a skeleton for rounding  
 222 error analysis for these algorithms that we will modify for different mixed precision assumptions in  
 223 section 4.

224 **3. Algorithms and existing round-off error analyses.** We introduce the Householder QR  
 225 factorization algorithm (HQR) in subsection 3.1 and two block variants that use HQR within the  
 226 block in subsections 3.2 and 3.3. The blocked HQR (BQR) in subsection 3.2 partitions the columns  
 227 of the target matrix and utilizes mainly level-3 BLAS operations and is a well-known algorithm that  
 228 uses the WY representation of [4]. In contrast, the Tall-and-Skinny QR (TSQR) in subsection 3.3  
 229 partitions rows of the matrix and takes a communication-avoiding divide-and-conquer approach  
 230 that can be easily parallelized (see [7]). We also present the crucial results in standard rounding  
 231 error analysis of these algorithms that excludes any mixed-precision assumptions. These building  
 232 steps of round-off error analysis will be easily tweaked for various mixed-precision assumptions in  
 233 section 4.

234 **3.1. Householder QR (HQR).** The HQR algorithm uses Householder transformations to  
 235 zero out elements below the diagonal of a matrix (see [16]). We present this as zeroing out all but  
 236 the first element of some vector,  $\mathbf{x} \in \mathbb{R}^m$ .

237 **LEMMA 3.1.** *Given vector  $\mathbf{x} \in \mathbb{R}^m$ , there exist Householder vector,  $\mathbf{v}$ , and Householder trans-*  
 238 *formation matrix,  $\mathbf{P}_{\mathbf{v}}$ , such that  $\mathbf{P}_{\mathbf{v}}$  zeros out  $\mathbf{x}$  below the first element.*

$$239 \quad (3.1) \quad \begin{aligned} \sigma &= -\text{sign}(x_1) \|\mathbf{x}\|_2, & \mathbf{v} &= \mathbf{x} - \sigma \mathbf{e}_1, \\ \beta &= \frac{2}{\mathbf{v}^\top \mathbf{v}} = -\frac{1}{\sigma v_1}, & \mathbf{P}_{\mathbf{v}} &= \mathbf{I}_m - \beta \mathbf{v} \mathbf{v}^\top. \end{aligned}$$

240 *The transformed vector,  $\mathbf{P}_{\mathbf{v}} \mathbf{x}$ , has the same 2-norm as  $\mathbf{x}$  since Householder transformations are*  
 241 *orthogonal:  $\mathbf{P}_{\mathbf{v}} \mathbf{x} = \sigma \mathbf{e}_1$ . In addition,  $\mathbf{P}_{\mathbf{v}}$  is symmetric and orthogonal,  $\mathbf{P}_{\mathbf{v}} = \mathbf{P}_{\mathbf{v}}^\top = \mathbf{P}_{\mathbf{v}}^{-1}$ .*

242 **3.1.1. HQR: Algorithm.** Given  $\mathbf{A} \in \mathbb{R}^{m \times n}$  and Lemma 3.1, HQR is done by repeating the  
 243 following processes until only an upper triangle matrix remains. For  $i = 1, 2, \dots, n$ ,  
 244 Step 1) Compute  $\mathbf{v}$  and  $\beta$  that zeros out the  $i^{\text{th}}$  column of  $\mathbf{A}$  beneath  $a_{ii}$  (see alg. 2), and  
 245 Step 2) Apply  $\mathbf{P}_{\mathbf{v}}$  to the bottom right partition,  $\mathbf{A}[i : m, i : n]$  (lines 4-6 of alg. 3).

246 Consider the following 4-by-3 matrix example adapted from [13]. Let  $\mathbf{P}_i$  represent the  $i^{\text{th}}$



247 Householder transformation of this algorithm.

$$\begin{aligned}
248 \quad \mathbf{A} &= \begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix} \xrightarrow{\text{apply } \mathbf{P}_1 \text{ to } \mathbf{A}} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \end{bmatrix} \xrightarrow{\text{apply } \mathbf{P}_2 \text{ to } \mathbf{P}_1 \mathbf{A}} \\
249 \quad & \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & \times \end{bmatrix} \xrightarrow{\text{apply } \mathbf{P}_3 \text{ to } \mathbf{P}_2 \mathbf{P}_1 \mathbf{A}} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & 0 \end{bmatrix} = \mathbf{P}_3 \mathbf{P}_2 \mathbf{P}_1 \mathbf{A} =: \mathbf{R} \\
250 \quad &
\end{aligned}$$

251 Then, the  $\mathbf{Q}$  factor for a full QR factorization is  $\mathbf{Q} := \mathbf{P}_1 \mathbf{P}_2 \mathbf{P}_3$  since  $\mathbf{P}_i$ 's are symmetric, and the  
252 thin factors for a general matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  are

$$253 \quad (3.2) \quad \mathbf{Q}_{\text{thin}} = \mathbf{P}_1 \cdots \mathbf{P}_n \mathbf{I}_{m \times n} \quad \text{and} \quad \mathbf{R}_{\text{thin}} = \mathbf{I}_{m \times n}^\top \mathbf{P}_n \cdots \mathbf{P}_1 \mathbf{A}.$$

---

**Algorithm 2:**  $\beta, \mathbf{v}, \sigma = \text{hh\_vec}(\mathbf{x})$ . Given a vector  $\mathbf{x} \in \mathbb{R}^n$ , return  $\mathbf{v}, \beta, \sigma$  that satisfy  $(I - \beta \mathbf{v} \mathbf{v}^\top) \mathbf{x} = \sigma \hat{\mathbf{e}}_1$  and  $\mathbf{v}_1 = 1$  (see [2, 13]).

---

**Input:**  $\mathbf{x} \in \mathbb{R}^m$   
**Output:**  $\mathbf{v} \in \mathbb{R}^m$ , and  $\sigma, \beta \in \mathbb{R}$  such that  $(I - \beta \mathbf{v} \mathbf{v}^\top) \mathbf{x} = \pm \|\mathbf{x}\|_2 \hat{\mathbf{e}}_1 = \sigma \hat{\mathbf{e}}_1$

254 1  $\mathbf{v} \leftarrow \text{copy}(\mathbf{x})$   
2  $\sigma \leftarrow -\text{sign}(\mathbf{x}_1) \|\mathbf{x}\|_2$   
3  $\mathbf{v}_1 \leftarrow \mathbf{x}_1 - \sigma$   
4  $\beta \leftarrow -\frac{\mathbf{v}_1}{\sigma}$   
5 **return**  $\beta, \mathbf{v}/\mathbf{v}_1, \sigma$

---



---

**Algorithm 3:**  $\mathbf{V}, \beta, \mathbf{R} = \text{HQR2}(A)$ . A Level-2 BLAS implementation of the Householder QR algorithm. Given a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  where  $m \geq n$ , return matrix  $\mathbf{V} \in \mathbb{R}^{m \times n}$ , vector  $\beta \in \mathbb{R}^n$ , and upper triangular matrix  $\mathbf{R}$ . An orthogonal matrix  $\mathbf{Q}$  can be generated from  $\mathbf{V}$  and  $\beta$ , and  $\mathbf{QR} = \mathbf{A}$ .

---

**Input:**  $A \in \mathbb{R}^{m \times n}$  where  $m \geq n$ .  
**Output:**  $\mathbf{V}, \beta, \mathbf{R}$

1  $\mathbf{V}, \beta \leftarrow \mathbf{0}_{m \times n}, \mathbf{0}_m$   
2 **for**  $i = 1 : n$  **do**  
3      $\mathbf{v}, \beta, \sigma \leftarrow \text{hh\_vec}(\mathbf{A}[i : \text{end}, i])$  /\* Algorithm 2 \*/  
4      $\mathbf{V}[i : \text{end}, i], \beta_i, \mathbf{A}[i, i] \leftarrow \mathbf{v}, \beta, \sigma$   
5      $\mathbf{A}[i + 1 : \text{end}, i] \leftarrow \text{zeros}(m - i)$   
6      $\mathbf{A}[i : \text{end}, i + 1 : \text{end}] \leftarrow \mathbf{A}[i : \text{end}, i + 1 : \text{end}] - \beta \mathbf{v} \mathbf{v}^\top \mathbf{A}[i : \text{end}, i + 1 : \text{end}]$   
7 **return**  $\mathbf{V}, \beta, \mathbf{A}[1 : n, 1 : n]$

---

255 **3.1.2. HQR: Rounding Error Analysis.** Now we present an error analysis for [alg. 3](#) by  
256 keeping track of the different operations of [alg. 2](#) and [alg. 3](#).

257 *Calculating the  $i^{\text{th}}$  Householder vector and constant.* In [alg. 3](#), the  $i^{\text{th}}$  Householder vector  
 258 shares all but the first component with the target column,  $\mathbf{A}[i : m, i]$ . We first calculate  $\sigma$  as is  
 259 implemented in line 2 of [alg. 2](#).

$$260 \quad (3.3) \quad \text{fl}(\sigma) = \hat{\sigma} = \text{fl}(-\text{sign}(\mathbf{A}_{i,i})\|\mathbf{A}[i : m, i]\|_2) = \sigma + \Delta\sigma, \quad |\Delta\sigma| \leq \gamma_{m-i+1}|\sigma|.$$

261 Note that the backward error incurred here is simply that an inner product of a vector in  $\mathbb{R}^{m-i+1}$   
 262 with itself. Let  $\tilde{\mathbf{v}}_1 \equiv \mathbf{A}_{i,i} - \sigma$ , the penultimate value  $\mathbf{v}_1$ . The subtraction adds a single additional  
 263 rounding error via

$$264 \quad \text{fl}(\tilde{\mathbf{v}}_1) = \tilde{\mathbf{v}}_1 + \Delta\tilde{\mathbf{v}}_1 = (1 + \delta)(\mathbf{A}_{i,i} - \sigma - \Delta\sigma) = (1 + \tilde{\theta}_{m-i+2})(\mathbf{A}_{i,i} - \sigma)$$

265 where the last equality is granted because the sign of  $\sigma$  is chosen to prevent cancellation. For the sake  
 266 of simplicity, we write  $|\Delta\tilde{\mathbf{v}}_1| \leq \tilde{\gamma}_{m-i+1}|\tilde{\mathbf{v}}_1|$  even though a tighter relative upper bound is  $\theta_{m-i+2}$   
 267 We sweep that minor difference (in comparison to  $\mathcal{O}(m-i)$ ) under the our use of the  $\tilde{\gamma}$  notation  
 268 defined in [Lemma 2.1](#). Since [alg. 2](#) normalizes the Householder vector so that its first component  
 269 is 1, the remaining components of  $\mathbf{v}$  are divided by  $\text{fl}(\tilde{\mathbf{v}}_1)$  incurring another single rounding error.  
 270 As a result, the rounding errors in  $\mathbf{v}$  are

$$271 \quad (3.4) \quad \text{fl}(\mathbf{v}_j) = \mathbf{v}_j + \Delta\mathbf{v}_j \text{ where } |\Delta\mathbf{v}_j| \leq \begin{cases} 0, & j = 1 \\ \tilde{\gamma}_{m-i+1}|\mathbf{v}_j|, & j = 2 : m - i + 1. \end{cases}$$

272 Next, we consider the Householder constant,  $\beta$ , as is computed in line 4 of [alg. 2](#).

$$273 \quad (3.5) \quad \hat{\beta} = \text{fl}\left(-\frac{\tilde{\mathbf{v}}_1}{\hat{\sigma}}\right) = -(1 + \delta)\frac{\tilde{\mathbf{v}}_1 + \Delta\tilde{\mathbf{v}}_1}{\sigma + \Delta\sigma}$$

$$274 \quad (3.6) \quad = \frac{(1 + \delta)(1 + \theta_{m-i+1})}{(1 + \theta_{m-i+2})}\beta = (1 + \theta_{3(m-i+2)})\beta$$

$$275 \quad (3.7) \quad = \beta + \Delta\beta, \text{ where } |\Delta\beta| \leq \tilde{\gamma}_{m-i+1}\beta.$$

277 We have shown [\(3.5\)](#) to keep our analysis simple in [section 4](#) and [\(3.6\)](#) and [\(3.7\)](#) show that the error  
 278 incurred from calculating of  $\|\mathbf{A}[i : m, i]\|_2$  accounts for the vast majority of the rounding error so  
 279 far.

280 *Applying a Single Householder Transformation.* Now we consider lines 4-6 of [alg. 3](#). Since  
 281 the entries in  $\mathbf{A}[i + 1 : m, i]$  are simply zeroed out and  $\mathbf{A}_{i,i}$  is replaced by  $\sigma$ , we only need to  
 282 calculate the errors for applying a Householder transformation with the computed Householder  
 283 vector and constant. This is the most crucial building block of the rounding error analysis for any  
 284 variant of HQR because the  $\mathbf{Q}$  factor is formed by applying the Householder transformations to  
 285 the identity and both of the blocked versions in [subsection 3.2](#) and [subsection 3.3](#) require efficient  
 286 implementations of this step. In this section, we only consider a level-2 BLAS implementation  
 287 of applying the Householder transformation, but in [subsection 3.2](#) we introduce a level-3 BLAS  
 288 implementation.

289 A Householder transformation is applied through a series of inner and outer products, since  
 290 Householder matrices are rank-1 updates of the identity. That is, computing  $\mathbf{P}_\mathbf{v}\mathbf{x}$  for any  $\mathbf{x} \in \mathbb{R}^m$   
 291 is as simple as computing  $\mathbf{y} := \mathbf{x} - (\beta\mathbf{v}^\top\mathbf{x})\mathbf{v}$ . Let us assume that  $\mathbf{x}$  is an exact vector and there  
 292 were errors incurred in forming  $\mathbf{v}$  and  $\beta$ . The errors incurred from computing  $\mathbf{v}$  and  $\beta$  need to be  
 293 included in addition to the new rounding errors accumulating from the action of applying  $\mathbf{P}_\mathbf{v}$  to

a column. In practice,  $\mathbf{x}$  would be a column in  $\mathbf{A}^{(i-1)}[i+1:m, i+1:n]$ , where the superscript  $(i-1)$  indicates that this submatrix of  $\mathbf{A}$  has already been transformed by  $i-1$  Householder transformations that zeroed out components below  $\mathbf{A}_{j,j}$  for  $j = 1:i-1$ . We show the error for forming  $\text{fl}(\hat{\mathbf{v}}^\top \mathbf{x})$  where we continue to let  $\mathbf{v}, \mathbf{x} \in \mathbb{R}^{m-i+1}$  as would be in the  $i^{\text{th}}$  iteration of the for-loop in [alg. 3](#):

$$\text{fl}(\hat{\mathbf{v}}^\top \mathbf{x}) = (1 + \theta_{m-i+1})(\mathbf{v} + \Delta \mathbf{v})^\top \mathbf{x}.$$

Set  $\mathbf{w} := \beta \mathbf{v}^\top \mathbf{x} \mathbf{v}$ . Then,

$$\hat{\mathbf{w}} = (1 + \theta_{m-i+1})(1 + \delta)(1 + \tilde{\delta})(\beta + \Delta \beta)(\mathbf{v} + \Delta \mathbf{v})^\top \mathbf{x}(\mathbf{v} + \Delta \mathbf{v}),$$

where  $\theta_{m-i+1}$  is from computing the inner product  $\hat{\mathbf{v}}^\top \mathbf{x}$ , and  $\delta$  and  $\tilde{\delta}$  are from multiplying  $\beta$ ,  $\text{fl}(\hat{\mathbf{v}}^\top \mathbf{x})$ , and  $\hat{\mathbf{v}}$  together. Finally, we can add in the vector subtraction operation and complete the rounding error analysis of applying a Householder transformation to any vector:

$$(3.8) \quad \text{fl}(\mathbf{x} - \hat{\mathbf{w}}) = (1 + \delta)(\mathbf{x} - \mathbf{w} - \Delta \mathbf{w}) = (1 + \tilde{\theta}_{m-i+1})\mathbf{y}.$$

We can easily switch between forward and errors from [\(3.8\)](#) via

$$\mathbf{y} + \Delta \mathbf{y} = (1 + \tilde{\theta}_{m-i+1})\mathbf{y} = (1 + \tilde{\theta}_{m-i+1})\mathbf{P}_\mathbf{v} \mathbf{x} = (\mathbf{P}_\mathbf{v} + \Delta \mathbf{P}_\mathbf{v})\mathbf{x},$$

where  $|\Delta \mathbf{y}| \leq \tilde{\gamma}_{m-i+1}|\mathbf{y}|$  and  $|\Delta \mathbf{P}_\mathbf{v}| \leq \tilde{\gamma}_{m-i+1}|\mathbf{P}_\mathbf{v}|$ .

Even though we never explicitly form  $\mathbf{P}_\mathbf{v}$ , forming the normwise error bound for this matrix makes the analysis for HQR simpler. Therefore, we now transition from componentwise error to matrix norm errors: the 2-norm and the Frobenius norm.

First, we transition from componentwise forward error to the 2-norm forward error via

$$(3.9) \quad \|\Delta \mathbf{y}\|_2 = \left( \sum_{i=1}^m \Delta \mathbf{y}_i^2 \right)^{1/2} \leq \left( (\tilde{\gamma}_{m-i+1})^2 \sum_{i=1}^m |\mathbf{y}_i|^2 \right)^{1/2} = \tilde{\gamma}_{m-i+1} \|\mathbf{y}\|_2.$$

In exact arithmetic, we are guaranteed  $\|\mathbf{y}\|_2 = \|\mathbf{P}_\mathbf{v} \mathbf{x}\|_2 \leq \|\mathbf{P}\|_2 \|\mathbf{x}\|_2 = \|\mathbf{x}\|_2$  since  $\mathbf{P}_\mathbf{v}$  is orthogonal and preserves norms. Combining this with [\(3.9\)](#) we find

$$(3.10) \quad \frac{\|\Delta \mathbf{y}\|_2}{\|\mathbf{x}\|_2} \leq \tilde{\gamma}_{m-i+1}.$$

Now we convert this to a normwise backward error. Since  $\Delta \mathbf{P}$  is exactly  $\frac{1}{\mathbf{x}^\top \mathbf{x}} \Delta \mathbf{y} \mathbf{x}^\top$ , we can compute its Frobenius norm by using  $\Delta \mathbf{P}_{ij} = \frac{1}{\|\mathbf{x}\|_2^2} \Delta \mathbf{y}_i \mathbf{x}_j$ ,

$$(3.11) \quad \|\Delta \mathbf{P}\|_F = \left( \sum_{i=1}^m \sum_{j=1}^m \left( \frac{1}{\|\mathbf{x}\|_2^2} \Delta \mathbf{y}_i \mathbf{x}_j \right)^2 \right)^{1/2} = \frac{\|\Delta \mathbf{y}\|_2}{\|\mathbf{x}\|_2} \leq \tilde{\gamma}_{m-i+1},$$

where the last inequality is a direct application of [\(3.10\)](#). We summarize these results in [Lemma 3.2](#).

**LEMMA 3.2.** *Let  $\mathbf{x} \in \mathbb{R}^m$  and consider the computation of  $\hat{\mathbf{y}} = \text{fl}(\mathbf{P}_\mathbf{v} \mathbf{x})$  via*

$$\mathbf{y} + \Delta \mathbf{y} = \text{fl}(\mathbf{P}_\mathbf{v} \mathbf{x}) = \text{fl}(\mathbf{x} - \hat{\beta} \hat{\mathbf{v}} \hat{\mathbf{v}}^\top \mathbf{x})$$

323 and rounding errors incurred in forming  $\hat{\mathbf{v}}$  and  $\hat{\beta}$  are expressed componentwise via  $\hat{\mathbf{v}} = \mathbf{v} + \Delta\mathbf{v}$   
 324 and  $\hat{\beta} = \beta + \Delta\beta$ . Let us write the componentwise forward error bound as  $|\Delta\mathbf{y}| \leq \gamma_y |\mathbf{y}|$ . Then, the  
 325 normwise forward and backward errors are

$$326 \quad \|\Delta\mathbf{y}\|_2 \leq \gamma_y \|\mathbf{y}\|_2, \quad \|\mathbf{P}_\mathbf{v}\|_F \leq \gamma_y.$$

327 Note that in a uniform precision setting this bound is represented as  $\gamma_y = \tilde{\gamma}_m$ , where the majority  
 328 of the round-off errors are attributed to inner product computations for forming  $\hat{\beta}$  and  $\mathbf{v}$ .

*Applying many successive Householder transformations.* Consider applying a sequence of transformations in the set  $\{\mathbf{P}_i\}_{i=1}^r \subset \mathbb{R}^{m \times m}$  to  $\mathbf{x} \in \mathbb{R}^m$ , where  $\mathbf{P}_i$ 's are all Householder transformations computed with  $\tilde{\mathbf{v}}_i$ 's and  $\hat{\beta}_i$ 's. This is directly applicable to HQR as  $\mathbf{Q} = \mathbf{P}_1 \cdots \mathbf{P}_n \mathbf{I}$  and  $\mathbf{R} = \mathbf{Q}^\top \mathbf{A} = \mathbf{P}_n \cdots \mathbf{P}_1 \mathbf{A}$ . Let us define

$$\mathbf{Q} + \Delta\mathbf{Q}' \equiv \prod_{i=1}^r (\mathbf{P}_i + \Delta\mathbf{P}_i)$$

329 in the context of applying this matrix to a vector,  $\mathbf{x} \in \mathbb{R}^m$ , where  $\Delta\mathbf{Q}'^\top$  represents the backward  
 330 error of forming  $\mathbf{R}$ , instead of the forward error of the  $\mathbf{Q}$  factor. The forward error for  $\mathbf{Q}$  is  
 331 denoted as  $\Delta\mathbf{Q} \equiv \text{fl}(\mathbf{Q}) - \mathbf{Q}$  where  $\text{fl}(\mathbf{Q})$  is formed via HQR. That is, if  $\mathbf{y} = \mathbf{Q}^\top \mathbf{x}$ , then  $\text{fl}(\mathbf{y}) =$   
 332  $\mathbf{y} + \Delta\mathbf{y} = (\mathbf{Q} + \Delta\mathbf{Q}')^\top \mathbf{x}$ . Even though an efficient implementation would use that  $\mathbf{P}_i$ 's are applied  
 333 to successively shorter vectors ( $\mathbf{P}_i$  is left multiplied to  $\mathbf{A}[i : m, i+1 : n]$ , which is equivalent to  
 334  $n-i$  vectors of length  $m-i+1$ ), we assume  $\{\mathbf{P}_i\}_{i=1}^r \subset \mathbb{R}^{m \times m}$  to allow for a simpler analysis while  
 335 forming a looser bound. We will now use Lemma 3.7 from [13] to bound  $\Delta\mathbf{Q}'$  with the Frobenius  
 336 norm.

$$337 \quad (3.11) \quad \|\Delta\mathbf{Q}'^\top\|_F = \left\| \prod_{i=1}^r (\mathbf{P}_i + \Delta\mathbf{P}_i) - \prod_{i=r}^1 \mathbf{P}_i \right\|_F,$$

$$338 \quad (3.12) \quad \leq \left( \prod_{i=1}^r (1 + \tilde{\gamma}_m) - 1 \right) \prod_{i=r}^1 \|\mathbf{P}_i\|_2 = (1 + \tilde{\gamma}_m)^r - 1.$$

340 The last equality results from the orthogonality of Householder matrices, and we further reduce the  
 341 last term. Generalizing the last rule in Lemma 2.2 yields

$$342 \quad (1 + \tilde{\gamma}_m)^r = (1 + \tilde{\gamma}_m)^{r-2} (1 + \tilde{\gamma}_m) (1 + \tilde{\gamma}_m) \leq (1 + \tilde{\gamma}_m)^{r-2} (1 + \tilde{\gamma}_{2m}) \leq \cdots \leq (1 + \tilde{\gamma}_{rm}).$$

343 Now we will use the following equivalent algebraic inequalities to get the final result.

$$344 \quad (3.13) \quad 0 < a < b < 1 \Leftrightarrow 1 - a > 1 - b \Leftrightarrow \frac{1}{1-a} < \frac{1}{1-b} \Leftrightarrow \frac{a}{1-a} < \frac{b}{1-b}$$

345 In addition, we assume  $r\tilde{\gamma}_m < \frac{1}{2}$ , such that

$$346 \quad (3.14) \quad (1 + \tilde{\gamma}_m)^r - 1 \leq \gamma_w^{(r\tilde{z})} = \frac{r\tilde{z}u_w}{1 - r\tilde{z}u_w} \quad (\text{by definition})$$

$$347 \quad (3.15) \quad \leq \frac{r\tilde{\gamma}_m}{1 - r\tilde{\gamma}_m}, \text{ since } r\tilde{z}u_w < r\tilde{\gamma}_m \quad (\text{by Equation 3.13})$$

$$348 \quad (3.16) \quad \leq 2r\tilde{\gamma}_m \quad (\text{since } r\tilde{\gamma}_m < \frac{1}{2} \text{ implies } \frac{1}{1 - r\tilde{\gamma}_m} < 2)$$

$$349 \quad (3.17) \quad = r\tilde{\gamma}_m,$$

350 Therefore, we have  $(1 + \tilde{\gamma}_m)^r - 1 \leq r\tilde{\gamma}_m$  and

$$352 \quad (3.18) \quad \|\Delta\mathbf{Q}'\|_2 \leq \|\Delta\mathbf{Q}'\|_F = \|\Delta\mathbf{Q}'^\top\|_F \leq r\tilde{\gamma}_m$$

353 In this current uniform precision error analysis, the important quantity  $\tilde{\gamma}_m$  is derived from the  
 354 backward error of applying one Householder transformation. To easily generalize this section for  
 355 mixed-precision analysis, we benefit from alternatively denoting this quantity as  $\tilde{\gamma}_{\mathbf{P}}$  with the un-  
 356 derstanding that  $\tilde{\gamma}_{\mathbf{P}}$  will be some combination of  $\tilde{\gamma}$ 's of differing precisions. Equation (3.17) would  
 357 then be

$$358 \quad (3.19) \quad (1 + \tilde{\gamma}_{\mathbf{P}})^r - 1 \leq r\tilde{\gamma}_{\mathbf{P}}.$$

359 Next, we apply (3.18) to the  $i^{\text{th}}$  columns of  $\mathbf{Q}, \mathbf{R}$  and set  $r = n$  for a full rank matrix,  $\mathbf{A}$ . Then,

$$360 \quad \|\Delta\mathbf{R}[:, i]\|_2 = \|\Delta\mathbf{Q}'^\top \mathbf{A}[:, i]\|_2 \leq \|\Delta\mathbf{Q}'\|_2 \|\mathbf{A}[:, i]\|_2 \leq n\tilde{\gamma}_m \|\mathbf{A}[:, i]\|_2,$$

$$361 \quad \|\Delta\mathbf{Q}[:, i]\|_2 = \|\Delta\mathbf{Q}'\mathbf{I}[:, i]\|_2 \leq \|\Delta\mathbf{Q}'\|_2 \leq n\tilde{\gamma}_m.$$

362 These columnwise bounds can now be transformed into matrix norms as follows:

$$364 \quad \|\Delta\mathbf{R}\|_F = \left( \sum_{i=1}^n \|\Delta\mathbf{R}[:, i]\|_2^2 \right)^{1/2} \leq \left( \sum_{i=1}^n n^2 \tilde{\gamma}_m^2 \|\mathbf{A}[:, i]\|_2^2 \right)^{1/2} = n\tilde{\gamma}_m \|\mathbf{A}\|_F,$$

$$365 \quad \|\Delta\mathbf{Q}\|_F = \left( \sum_{i=1}^n \|\Delta\mathbf{Q}[:, i]\|_2^2 \right)^{1/2} \leq \left( \sum_{i=1}^n \tilde{\gamma}_m^2 \right)^{1/2} = n^{3/2} \tilde{\gamma}_m.$$

366 We gather these results into Theorem 3.3.

367 THEOREM 3.3. Let  $\mathbf{A} \in \mathbb{R}^{m \times n}$  with  $m \geq n$  have full rank,  $n$ . Let  $\hat{\mathbf{Q}} \in \mathbb{R}^{m \times n}$  and  $\hat{\mathbf{R}} \in \mathbb{R}^{n \times n}$  be  
 368 the thin QR factors of  $\mathbf{A}$  obtained via alg. 3, defined via

$$370 \quad \hat{\mathbf{R}} = \mathbf{R} + \Delta\mathbf{R} = \text{fl}(\hat{\mathbf{P}}_n \cdots \hat{\mathbf{P}}_1 \mathbf{A}), \quad n\tilde{\gamma}_m \|\mathbf{A}\|_F$$

$$371 \quad \hat{\mathbf{Q}} = \mathbf{Q} + \Delta\mathbf{Q} = \text{fl}(\hat{\mathbf{P}}_1 \cdots \hat{\mathbf{P}}_n \mathbf{I}), \quad \|\Delta\mathbf{Q}\|_F \leq n^{3/2} \tilde{\gamma}_m.$$

372 Let  $\mathbf{A} + \Delta\mathbf{A} = \hat{\mathbf{Q}}\hat{\mathbf{R}}$ , where  $\hat{\mathbf{Q}}$  and  $\hat{\mathbf{R}}$  are obtained via Algorithm 3. Then the backward error is

$$373 \quad (3.20) \quad \|\Delta\mathbf{A}\|_F \leq n^{3/2} \tilde{\gamma}_m \|\mathbf{A}\|_F.$$

374 The content of this section is largely derived directly from [13], but we kept the analysis  
 375 general by employing quantities denoted via  $\Delta\beta$ ,  $\Delta\mathbf{v}$ ,  $\tilde{\gamma}_y$ , and  $\tilde{\gamma}_{\mathbf{P}}$ . These quantities account for  
 376 various forward and backward errors formed in computing essential components of HQR, namely the  
 377 Householder constant and vector, as well as normwise errors of the action of applying Householder  
 378 transformations. In the next sections, we present blocked variants of HQR that use alg. 3.

**3.2. Block HQR with partitioned columns (BQR).** We refer to the blocked variant of HQR where the columns are partitioned as BQR. Note that this algorithm relies on the WY representation described in [4] instead of the storage-efficient version of [19].

**3.2.1. The WY Representation.** A convenient matrix representation that accumulates  $r$  Householder reflectors is known as the WY representation (see [4, 10]). Lemma 3.4 shows how to update a rank- $j$  update of the identity,  $\mathbf{Q}^{(j)}$ , with a Householder transformation,  $\mathbf{P}$ , to produce a rank- $(j+1)$  update of the identity,  $\mathbf{Q}^{(j+1)}$ . With the correct initialization of  $\mathbf{W}$  and  $\mathbf{Y}$ , we can build the WY representation of successive Householder transformations as shown in Algorithm 4. This algorithm assumes that the Householder vectors,  $\mathbf{V}$ , and constants,  $\beta$ , have already been computed. Since the  $\mathbf{Y}$  factor is exactly  $\mathbf{V}$ , we only need to compute the  $\mathbf{W}$  factor.

LEMMA 3.4. Suppose  $\mathbf{Q}^{(j)} = \mathbf{I} - \mathbf{W}^{(j)}\mathbf{Y}^{(j)\top} \in \mathbb{R}^{m \times m}$  is an orthogonal matrix with  $\mathbf{W}^{(j)}, \mathbf{Y}^{(j)} \in \mathbb{R}^{m \times j}$ . Let us define  $\mathbf{P} = \mathbf{I} - \beta \mathbf{v}\mathbf{v}^\top$  for some  $\mathbf{v} \in \mathbb{R}^m$  and let  $\mathbf{z}^{(j+1)} = \beta \mathbf{Q}^{(j)}\mathbf{v}$ . Then,

$$\mathbf{Q}^{(j+1)} = \mathbf{Q}^{(j)}\mathbf{P} = \mathbf{I} - \mathbf{W}^{(j+1)}\mathbf{Y}^{(j+1)\top},$$

where  $\mathbf{W}^{(j+1)} = [\mathbf{W}^{(j)}|\mathbf{z}]$  and  $\mathbf{Y}^{(j+1)} = [\mathbf{Y}^{(j)}|\mathbf{v}]$  are each  $m$ -by- $(j+1)$ .

---

**Algorithm 4:**  $\mathbf{W}, \mathbf{Y} \leftarrow \text{buildWY}(\mathbf{V}, \beta)$ : Given a set of householder vectors  $\{\mathbf{V}[:, i]\}_{i=1}^r$  and their corresponding constants  $\{\beta_i\}_{i=1}^r$ , form the final  $\mathbf{W}$  and  $\mathbf{Y}$  factors of the WY representation of  $\mathbf{P}_1 \cdots \mathbf{P}_r$ , where  $\mathbf{P}_i := \mathbf{I}_m - \beta_i \mathbf{v}_i \mathbf{v}_i^\top$

---

**Input:**  $\mathbf{V} \in \mathbb{R}^{m \times r}$ ,  $\beta \in \mathbb{R}^r$  where  $m > r$ .

**Output:**  $\mathbf{W}$

```

1 Initialize:  $\mathbf{W} := \beta_1 \mathbf{V}[:, 1]$ . /*  $\mathbf{Y}$  is  $\mathbf{V}$ . */
2 for  $j = 2 : r$  do
3    $\mathbf{z} \leftarrow \beta_j [\mathbf{V}[:, j] - \mathbf{W}(\mathbf{V}[:, 1:j-1]^\top \mathbf{V}[:, j])]$ 
4    $\mathbf{W} \leftarrow [\mathbf{W} \quad \mathbf{z}]$  /* Update  $\mathbf{W}$  to an  $m$ -by- $j$  matrix. */
5 return  $\mathbf{W}$ 
```

---

In HQR,  $\mathbf{A}$  is transformed into an upper triangular matrix  $\mathbf{R}$  by identifying a Householder transformation that zeros out a column below the diagonal, then applying that Householder transformation to the bottom right partition. For example, the  $k^{\text{th}}$  Householder transformation finds an  $m - k + 1$  sized Householder transformation that zeros out column  $k$  below the diagonal and then applies it to the  $(m - k + 1)$ -by- $(n - k)$  partition of the matrix,  $\mathbf{A}[k : m, k + 1 : n]$ . Since the  $k + 1^{\text{st}}$  column is transformed by the  $k^{\text{th}}$  Householder transformation, this algorithm must be executed serially as shown in alg. 3. The highest computational burden at each iteration falls on alg. 3 line 6, which requires Level-2 BLAS operations when computed efficiently.

In contrast, BQR replaces this step with Level-3 BLAS operations by partitioning  $\mathbf{A}$  by groups of columns. Let  $\mathbf{A} = [\mathbf{C}_1 \cdots \mathbf{C}_N]$  where  $\mathbf{C}_1, \dots, \mathbf{C}_{N-1}$  are each  $m$ -by- $r$ , and  $\mathbf{C}_N$  holds the remaining columns. The  $k^{\text{th}}$  block,  $\mathbf{C}_k$ , is transformed using HQR (alg. 3) while building the WY representation of  $\mathbf{P}_{(k-1)r+1} \cdots \mathbf{P}_{kr} = \mathbf{I}_m - \mathbf{W}_k \mathbf{Y}_k^\top$  as in alg. 4. Note that both algs. 3 and 4 are rich in Level-2 BLAS operations. Then,  $\mathbf{I} - \mathbf{Y}_k \mathbf{W}_k^\top$  is applied to  $[\mathbf{C}_2 \cdots \mathbf{C}_N]$  with two Level-3 BLAS operations as shown in line 6 of alg. 5. BQR performs approximately  $1 - \mathcal{O}(1/N)$  fraction of its FLOPs in Level-3 BLAS operations (see section 5.2.3 of [10]), and can reap the benefits from the accelerated matrix-matrix-multiply and accumulate technology of TensorCore. Note that BQR

does require more FLOPs when compared to HQR, but these additional FLOPs are negligible in high precision. A pseudoalgorithm for BQR is shown in [alg. 5](#). Note that the subscripts on  $\mathbf{W}_i$  indicate the WY representation for the Householder transformations on the  $i^{th}$  block of  $\mathbf{A}$ ,  $\mathbf{C}_k$ , whereas the superscripts on  $\mathbf{W}^{(j)}$  in [Lemma 3.4](#) refers to the  $j^{th}$  update within building a WY representation.

---

**Algorithm 5:**  $\mathbf{Q}, \mathbf{R} \leftarrow \text{blockHQR}(\mathbf{A}, r)$ : Perform Householder QR factorization of matrix  $\mathbf{A}$  with column partitions of size  $r$ .

---

**Input:**  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $r \in \mathbb{R}$  where  $r < n$ .  
**Output:**  $\mathbf{Q}, \mathbf{R}$

```

1  $N = \lceil \frac{n}{r} \rceil$ 
  // Let  $\mathbf{A} = [\mathbf{C}_1 \cdots \mathbf{C}_N]$  where all blocks except  $\mathbf{C}_N$  are  $m$ -by- $r$  sized.
2 for  $i = 1 : N$  do
3    $\mathbf{V}_i, \beta_i, \mathbf{C}_i \leftarrow \text{hhQR}(\mathbf{C}_i)$                                 /* Algorithm 3 */
4    $\mathbf{W}_i \leftarrow \text{buildWY}(\mathbf{V}_i, \beta_i)$                                 /* Algorithm 4 */
5   if  $i < N$  then
6      $[\mathbf{C}_{i+1} \cdots \mathbf{C}_N] \leftarrow \mathbf{V}_i \left( \mathbf{W}_i^\top [\mathbf{C}_{i+1} \cdots \mathbf{C}_N] \right)$  /* update the rest: BLAS-3 */
  //  $\mathbf{A}$  has been transformed into  $\mathbf{R} = \mathbf{Q}^\top \mathbf{A}$ .
  // Now build  $\mathbf{Q}$  using level-3 BLAS operations.
7  $\mathbf{Q} \leftarrow \mathbf{I}$                                                     /*  $\mathbf{I}_m$  if full QR, and  $\mathbf{I}_{m \times n}$  if thin QR. */
8 for  $i = N : -1 : 1$  do
9    $\mathbf{Q}[(i-1)r+1 : m, (i-1)r+1 : n] \leftarrow \mathbf{W}_i \left( \mathbf{V}_i^\top \mathbf{Q}[(i-1)r+1 : m, (i-1)r+1 : n] \right)$ 
10 return  $\mathbf{Q}, \mathbf{A}$ 
```

---

**3.2.2. BQR: Rounding Error Analysis.** We now present the basic structure for the rounding error analysis for [alg. 5](#), which consist of: 1)HQR, 2)building the  $\mathbf{W}$  factor, and 3) updating the remaining blocks with the WY representation. We have adapted the analysis from [\[13\]](#) to fit this exact variant. Furthermore, we assume that  $n$  is divisible by  $r$  so that  $N = \lceil n/r \rceil = n/r$  to make our error analysis simple. In practice, an efficient implementation might require  $r$  to be a power of two or a product of small prime factors and result a thinner  $N^{th}$  block compared to the rest. This discrepancy is easily fixed by padding the matrix with zeros, a standard procedure for standard algorithms like the Fast Fourier Transform (FT).

*HQR within each block: line 3 of [alg. 5](#).* Consider the  $k^{th}$  block,  $\mathbf{C}_k$ . We first apply [Algorithm 3](#) to rows  $k : m$ , which is equivalent to  $r$  Householder transformations of size  $m - (r-1)k$  if  $k < N$  and  $n - r(i-1)$  Householder transformations of size  $m - (N-1)k$  if  $k = N$ . The general rounding error incurred is represented by [\(3.11\)](#) and a uniform-precision rounding error for this step is shown in [\(3.12\)](#). Let  $\mathbf{R}_k$  be the  $\mathbf{R}$  factor acquired from applying HQR on  $\mathbf{C}_k$  and let  $\Delta \mathbf{Q}'_k$  represent the backward error incurred in this step:

$$\mathbf{R}_k = (\mathbf{Q}_k + \Delta \mathbf{Q}'_k)^\top \mathbf{C}_k.$$

Adjusting [\(3.18\)](#) for  $k < N$  yields

$$(3.21) \quad \|\Delta \mathbf{Q}'\|_2 \leq \|\Delta \mathbf{Q}'\|_F = \|\Delta \mathbf{Q}'^\top\|_F \leq r \tilde{\gamma}_{m-(i-1)k}$$



432 Since the  $j^{th}$  Householder vector for  $\mathbf{C}_k$  zeros out the  $r(k-1)+j^{th}$  column of  $\mathbf{A}$  below the diagonal,  
 433 it has length  $m - (k-1)r - j + 1$ . This corresponds to setting  $i = r(k-1) + j$  in (3.4) and (3.7)

*Build WY at each block: line 4 of alg. 5.* We now calculate the rounding errors incurred from building the WY representation when given a set of Householder vectors and constants as shown in alg. 4. Consider the  $j^{th}$  update of the WY representation for block  $\mathbf{C}_k$ . Let us denote the  $j^{th}$  Householder constant and vector of the  $k^{th}$  block computed with FLOPs as with  $\hat{\beta}_k^{(j)}$  and  $\hat{\mathbf{v}}_k^{(j)}$  and we write  $\mathbf{V}_k[:, 1:j] =: \mathbf{Y}_k^{(j)}$  to make a clear connection to Lemma 3.4. Error accumulated from forming  $\mathbf{v}_k^{(j)}$  is represented in (3.4) and we only need to consider  $\mathbf{z}_k^{(j)} = \beta_k^{(j)}(\mathbf{I} - \mathbf{W}_k^{(j-1)}\mathbf{Y}_k^{(j-1)\top})\mathbf{v}_k^{(j)}$ . An implementation of this computes

$$\mathbf{z}_k^{(j)}\hat{\beta}_k^{(j)}\left(\hat{\mathbf{v}}_k^{(j)} - \hat{\mathbf{W}}_k^{(j-1)}\hat{\mathbf{Y}}_k^{(j-1)\top}\hat{\mathbf{v}}_k^{(j)}\right),$$

434 and we consider just the computation of the intermediate term,  $\hat{\mathbf{W}}_k^{(j-1)}(\hat{\mathbf{Y}}_k^{(j-1)\top}\hat{\mathbf{v}}_k^{(j)})$ . For  
 435 simplicity, we assume  $\hat{\mathbf{v}}_k^{(j)}$ ,  $\hat{\mathbf{W}}_k^{(j)}$ , and  $\hat{\mathbf{Y}}_k^{(j)}$  all have  $m - (k-1)r$  rows. The rounding error  
 436 incurred in this step is a componentwise error that results from inner products of two different  
 437 lengths,  $m - (k-1)r$  and  $j$ :

$$438 \quad (3.22) \quad \text{fl}(\hat{\mathbf{W}}_k^{(j-1)}(\hat{\mathbf{Y}}_k^{(j-1)\top}\hat{\mathbf{v}}_k^{(j)})) = (1 + \theta_{m-(k-1)r})(1 + \theta_j)\hat{\mathbf{W}}_k^{(j-1)}(\hat{\mathbf{Y}}_k^{(j-1)\top}\hat{\mathbf{v}}_k^{(j)}),$$

$$439 \quad (3.23) \quad = (1 + \tilde{\theta}_{m-(k-1)r})\hat{\mathbf{W}}_k^{(j-1)}(\hat{\mathbf{Y}}_k^{(j-1)\top}\hat{\mathbf{v}}_k^{(j)}),$$

441 where they are lumped into  $\tilde{\theta}_{m-(k-1)r}$  together. Now we delve into how the errors from construct-  
 442 ing the previous updates,  $\hat{\mathbf{W}}_k^{(j-1)}$  and  $\hat{\mathbf{Y}}_k^{(j-1)}$ , are carried onto the current updates in exact  
 443 arithmetic. We compute

$$444 \quad \hat{\mathbf{W}}_k^{(j-1)}(\hat{\mathbf{Y}}_k^{(j-1)\top}\hat{\mathbf{v}}_k^{(j)}) = \sum_{j'=1}^{j-1} \mathbf{z}_k^{(j')}\hat{\mathbf{v}}_k^{(j')\top}\hat{\mathbf{v}}_k^{(j)},$$

$$445 \quad = \sum_{j'=1}^{j-1} (\mathbf{z}_k^{(j')} + \Delta\mathbf{z}_k^{(j')})(\mathbf{v}_k^{(j')} + \Delta\mathbf{v}_k^{(j')})^\top(\mathbf{v}_k^{(j)} + \Delta\mathbf{v}_k^{(j)}),$$

$$446$$

where  $|\Delta\mathbf{v}_k^{(j')}| \leq \tilde{\gamma}_{m-(k-1)r+j'}|\mathbf{v}_k^{(j')}|$  and  $|\Delta\mathbf{v}_k^{(j)}| \leq \tilde{\gamma}_{m-(k-1)r+j}|\mathbf{v}_k^{(j)}|$ . Since  $j' < j$ , we use Lemma 2.2 and that  $j < r$  to result in

$$\hat{\mathbf{v}}_k^{(j')\top}\hat{\mathbf{v}}_k^{(j)} = \mathbf{v}_k^{(j')\top}\mathbf{v}_k^{(j)} + \Delta\mathbf{v}_{k,j},$$

447 where  $|\Delta\mathbf{v}_{k,j}| \leq \tilde{\gamma}_{m-(k-2)r}|\mathbf{v}_k^{(j')\top}\mathbf{v}_k^{(j)}|$ . Then, we are left only with the errors in  $\mathbf{z}_k^{(j')}$ 's as shown  
 448 below:

$$449 \quad (3.24) \quad \hat{\mathbf{W}}_k^{(j-1)}(\hat{\mathbf{Y}}_k^{(j-1)\top}\hat{\mathbf{v}}_k^{(j)}) = (1 + \tilde{\theta}_{m-(k-2)r}) \sum_{j'=1}^{j-1} (\mathbf{z}_k^{(j')} + \Delta\mathbf{z}_k^{(j')})\mathbf{v}_k^{(j')\top}\mathbf{v}_k^{(j)}.$$

450 Since the rounding errors from FLOPs shown in (3.23) are of order only  $\tilde{\theta}_{m-(k-1)r}$  and only 2  
 451 additional rounding errors are added from the multiplication of  $\hat{\beta}_k^{(j)}$  and the vector subtraction, we

452 can state

$$453 \quad (3.25) \quad \mathbf{f}(\mathbf{z}_{\mathbf{k}}^{(j)}) = (1 + \tilde{\theta}_{m-(k-2)r})\beta_k^{(j)} \left( \mathbf{v}_{\mathbf{k}}^{(j)} - \sum_{j'=1}^{j-1} (\mathbf{z}_{\mathbf{k}}^{(j')} + \Delta\mathbf{z}_{\mathbf{k}}^{(j')})\mathbf{v}_{\mathbf{k}}^{(j')\top}\mathbf{v}_{\mathbf{k}}^{(j)} \right).$$

The errors for  $\mathbf{z}_{\mathbf{k}}^{(j')}$ 's remain, and we solve for these recursively. Note that  $\mathbf{z}_{\mathbf{k}}^{(1)}$  only incurs a single rounding error since it is initialized with  $\hat{\beta}_k^{(1)}\hat{\mathbf{v}}_{\mathbf{k}}^{(1)}$ . That is,

$$\hat{\mathbf{z}}_{\mathbf{k}}^{(1)} = \mathbf{z}_{\mathbf{k}}^{(1)} + \Delta\mathbf{z}_{\mathbf{k}}^{(1)}, \quad |\Delta\mathbf{z}_{\mathbf{k}}^{(1)}| \leq \tilde{\gamma}_{m-(k-1)r}|\mathbf{z}_{\mathbf{k}}^{(1)}|,$$

454 accounting for  $\Delta\beta_k^{(1)}$ ,  $\Delta\mathbf{v}_{\mathbf{k}}^{(1)}$ , and the multiplication operation. Using (3.25) for  $j = 2$  yields

$$455 \quad (3.26) \quad \mathbf{z}_{\mathbf{k}}^{(2)} + \Delta\mathbf{z}_{\mathbf{k}}^{(2)} = (1 + \tilde{\theta}_{m-(k-2)r})\beta_k^{(2)} \left( \mathbf{v}_{\mathbf{k}}^{(2)} - (\mathbf{z}_{\mathbf{k}}^{(1)} + \Delta\mathbf{z}_{\mathbf{k}}^{(1)})\mathbf{v}_{\mathbf{k}}^{(1)\top}\mathbf{v}_{\mathbf{k}}^{(2)} \right)$$

$$456 \quad (3.27) \quad = (1 + \tilde{\theta}_{m-(k-2)r})(1 + \tilde{\theta}_{m-(k-2)r})\mathbf{z}_{\mathbf{k}}^{(2)},$$

458 where we do not consolidate the two  $\tilde{\theta}_{m-(k-2)r}$ 's into just one  $\tilde{\theta}_{m-(k-2)r}$  under the small integer  
459 rule of the tilde notation, since we will continue this recursive relation. At  $j = r$ , we have

$$460 \quad (3.28) \quad \mathbf{z}_{\mathbf{k}}^{(r)} + \Delta\mathbf{z}_{\mathbf{k}}^{(r)} = (1 + \tilde{\theta}_{m-(k-2)r})^r \mathbf{z}_{\mathbf{k}}^{(r)}.$$

461 **3.3. Block HQR with partitioned rows : Tall-and-Skinny QR (TSQR).**

462 **4. Mixed-precision error analysis.**

463 **4.1. Round down at the end of the factorization.**

464 **4.2. Round down at block-level (BLAS-3).**

465 **4.3. Round down at inner-product level (BLAS-2).**

466 **5. Numerical Experiments.**

467 **6. Conclusion.** Though the use of lower precision naturally reduces the bandwidth and stor-  
468 age needs, the development of GPUs to optimize low precision floating point arithmetic have ac-  
469 celerated the interest in half precision and mixed-precision algorithms. Loss in precision, stability,  
470 and representable range offset for those advantages, but these shortcomings may have little to no  
471 impact in some applications. It may even be possible to navigate around those drawbacks with  
472 algorithmic design.

473 The existing rounding error analysis cannot accurately bound the behavior of mixed-precision  
474 arithmetic. We have developed a new framework for mixed-precision rounding error analysis and  
475 applied it to HQR, a widely used linear algebra routine, and implemented it in an iterative eigen-  
476 solver in the context of spectral clustering. The mixed-precision error analysis builds from the  
477 inner product routine, which can be applied to many other linear algebra tools as well. The new  
478 error bounds more accurately describe how rounding errors are accumulated in mixed-precision set-  
479 tings. We also found that TSQR, a communication-avoiding, easily parallelizable QR factorization  
480 algorithm for tall-and-skinny matrices, can outperform HQR in mixed-precision settings for ill-  
481 conditioned, extremely overdetermined cases, which suggests that some algorithms are more robust  
482 against lower precision arithmetic. As QR factorizations of tall-and-skinny matrices are common

in spectral clustering, we experimented with introducing mixed-precision settings into graph partitioning problems. In particular, we applied DBSCAN to the spectral basis of a graph identified via subspace iteration that used our simulated mixed-precision HQR, which yielded clustering results tantamount to results from employing double-precision entirely.

Although this work is focused on QR factorizations and applications in spectral clustering, the mixed precision round-off error analysis can be applied to other tasks and applications that can benefit from employing low precision computations. While the emergence of technology that support low precision floats combats issues dealing with storage, now we need to consider how low precision affects stability of numerical algorithms.

Future work is needed to test larger, more ill-conditioned problems with different mixed-precision settings, and to explore other divide-and-conquer methods like TSQR that can harness parallel capabilities of GPUs while withstanding lower precisions.

## REFERENCES

- [1] A. ABDELFAH, S. TOMOV, AND J. DONGARRA, *Fast batched matrix multiplication for small sizes using half-precision arithmetic on GPUs*, in 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), May 2019, pp. 111–122, <https://doi.org/10.1109/IPDPS.2019.00022>.
- [2] E. ANDERSON, Z. BAI, C. BISCHOF, L. S. BLACKFORD, J. DEMMEL, J. J. DONGARRA, J. DU CROZ, S. HAMMARLING, A. GREENBAUM, A. MCKENNEY, AND D. SORESENSEN, *LAPACK Users' Guide (Third Ed.)*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999; also available online from <http://www.netlib.org>.
- [3] J. APPLEYARD AND S. YOKIM, *Programming Tensor Cores in CUDA 9*, 2017, <https://devblogs.nvidia.com/programming-tensor-cores-cuda-9/> (accessed 2018-07-30).
- [4] C. BISCHOF AND C. VAN LOAN, *The WY Representation for Products of Householder Matrices*, SIAM Journal on Scientific and Statistical Computing, 8 (1987), pp. s2–s13, <https://doi.org/10.1137/0908009>.
- [5] M. COURBARIAUX, Y. BENGIO, AND J.-P. DAVID, *Training deep neural networks with low precision multiplications*, arXiv preprint, arXiv:1412.7024, (2014).
- [6] M. COURBARIAUX, J.-P. DAVID, AND Y. BENGIO, *Low precision storage for deep learning*, arXiv preprint arXiv:1412.7024, (2014).
- [7] J. DEMMEL, I. DUMITRIU, AND O. HOLTZ, *Fast linear algebra is stable*, Numerische Mathematik, 108 (2007), pp. 59–91, <https://doi.org/10.1007/s00211-007-0114-x>, <https://arxiv.org/abs/0612264>.
- [8] J. DEMMEL, L. GRIGORI, M. HOEMMEN, AND J. LANGOU, *Communication-optimal parallel and sequential QR and LU factorizations*, SIAM Journal on Scientific Computing, 34 (2012), <https://doi.org/10.1137/080731992>, <https://arxiv.org/abs/0808.2664>.
- [9] M. FAGAN, J. SCHLACHTER, K. YOSHII, S. LEYFFER, K. PALEM, M. SNIR, S. M. WILD, AND C. ENZ, *Overcoming the power wall by exploiting inexactness and emerging COTS architectural features: Trading precision for improving application quality*, in 2016 29th IEEE International System-on-Chip Conference (SOCC), Sep. 2016, pp. 241–246, <https://doi.org/10.1109/SOCC.2016.7905477>.
- [10] G. H. GOLUB AND C. F. VAN LOAN, *Matrix computations*, JHU press, 4 ed., 2013.
- [11] A. HAIDAR, A. ABDELFAH, M. ZOUNON, P. WU, S. PRANESH, S. TOMOV, AND J. DONGARRA, *The Design of Fast and Energy-Efficient Linear Solvers: On the Potential of Half-Precision Arithmetic and Iterative Refinement Techniques*, June 2018, pp. 586–600, [https://doi.org/10.1007/978-3-319-93698-7\\_45](https://doi.org/10.1007/978-3-319-93698-7_45).
- [12] A. HAIDAR, S. TOMOV, J. DONGARRA, AND N. J. HIGHAM, *Harnessing GPU tensor cores for fast fp16 arithmetic to speed up mixed-precision iterative refinement solvers*, in Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, SC '18, Piscataway, NJ, USA, 2018, IEEE Press, pp. 47:1–47:11, <https://doi.org/10.1109/SC.2018.00050>, <https://doi.org/10.1109/SC.2018.00050>.
- [13] N. J. HIGHAM, *Accuracy and Stability of Numerical Methods*, 2002, <https://doi.org/10.2307/2669725>.
- [14] N. J. HIGHAM AND T. MARY, *A New Approach to Probabilistic Rounding Error Analysis*, SIAM Journal on Scientific Computing, 41 (2019), pp. A2815–A2835, <https://doi.org/10.1137/18M1226312>, <https://epubs.siam.org/doi/10.1137/18M1226312>.
- [15] N. J. HIGHAM AND S. PRANESH, *Simulating Low Precision Floating-Point Arithmetic*, SIAM Journal on Scientific Computing, 41 (2019), pp. C585–C602, <https://doi.org/10.1137/19M1251308>, <https://epubs.siam.org/doi/10.1137/19M1251308>.

- org/doi/10.1137/19M1251308.
- [16] A. S. HOUSEHOLDER, *Unitary triangularization of a nonsymmetric matrix*, Journal of the ACM (JACM), 5 (1958), pp. 339–342.
  - [17] I. C. F. IPSEN AND H. ZHOU, *Probabilistic Error Analysis for Inner Products*, (2019), <http://arxiv.org/abs/1906.10465>, <https://arxiv.org/abs/1906.10465>.
  - [18] P. MICKEVICIUS, S. NARANG, J. ALBEN, G. DIAMOS, E. ELSEEN, D. GARCIA, B. GINSBURG, M. HOUSTON, O. KUCHAIEV, G. VENKATESH, AND H. WU, *Mixed precision training*, in International Conference on Learning Representations, 2018, <https://openreview.net/forum?id=r1gs9JgRZ>.
  - [19] R. SCHREIBER AND C. VAN LOAN, *A Storage-Efficient  $WY$  Representation for Products of Householder Transformations*, SIAM Journal on Scientific and Statistical Computing, 10 (1989), pp. 53–57, <https://doi.org/10.1137/0910005>.
  - [20] G. TAGLIAVINI, S. MACH, D. ROSSI, A. MARONGIU, AND L. BENIN, *A transprecision floating-point platform for ultra-low power computing*, in 2018 Design, Automation Test in Europe Conference Exhibition (DATE), March 2018, pp. 1051–1056, <https://doi.org/10.23919/DATE.2018.8342167>.
  - [21] U. VON LUXBURG, *A tutorial on spectral clustering*, Statistics and Computing, 17 (2007), pp. 395–416, <https://doi.org/10.1007/s11222-007-9033-z>, <https://doi.org/10.1007/s11222-007-9033-z>.