

RESPONSE TO REVIEWERS

We are grateful to our editor, Laura Grigori, for handling the reviews of our paper, and to both reviewers for their constructive criticisms.

Our response to the reviews is below, and we have appended a diff file at the end that highlights which changes have been made.

Reviewer #1

1. line 5: “lower storage” → “lower storage cost” or “lower storage requirement”
Done.
2. line 13: Inconsistent use of “round off” and “round-off” (see, e.g., line 16)
Done.
3. line 11: Inconsistent hyphenation, e.g. “low precision” in line 11 and “high-precision” in line 21
Done.
4. line 25: What is the “weight of memory”? “4x” versus what?
We were referring to the physical weight for application like drones. However, it is not necessary for this sentence and therefore has been removed.
5. line 36: “computed up to 16x than that of fp64”— do you mean “16x faster”?
Done.
6. lines 36-37: “The existing rounding error analyses are built within ...” Existing rounding error analyses of what?
Done (added “of linear algebra routines”).
7. line 39: This is the first time that HH QR (or QR factorization at all) is even mentioned. The problem that you are considering should be stated earlier in the text.
We have included an additional sentence at the end of the first paragraph of the introduction that mentions the QR factorization.
8. lines 43-44: “... its block variants that partition the columns ... and those that partition the columns” Is there some typo here?
Fixed.
9. line 58: What does “sufficient in performance” mean?
This is now clarified.
10. line 137: The “(j,k)” in the max should not be a subscript.
Done.
11. line 144: State here what algorithms 5 and 6 are, since they have not been introduced yet.
Done.
12. line 145: Define what is meant by “ad hoc”.
We simply mean that MP Setting 2.3 is one we have constructed and there’s no hardware that implements it to the best of our knowledge currently.
13. line 155: “is in” → “in”
Done.
14. line 157: There is no “Assumption 2.3”. I assume you meant “MP Setting 2.3”.
Done.
15. line 157-160: It might help to give a verbal summary of this, i.e., “When two fp16 numbers are multiplied and stored in fp32, there is no roundoff error”.

Done.

16. line 168: The concept of “full precision” needs some additional constraint. By my understanding, this “full precision” argument only applies when $u^{(h)} \leq (u^{(l)})^2$. If so, then this constraint should be stated explicitly. This further constrains the values that the “disparity in the precisions” $M_{l,h}$ can have (line 192). It would be good to state these bounds on $M_{l,h}$.
We realize the impreciseness of “full precision” in general terms. We first encountered this term in the NVIDIA’s description of the TensorCore technology (see [4]), and determined that for IEEE fp16, fp32, and fp64, this means exact product. We have changed the wording the definition of MP Setting 2.3 to reflect this but still leave the analysis that shows that a full precision product of 2 fp16 numbers is exact when represented in fp32. Furthermore, the constraint on $M_{l,h}$ if $u^{(h)} \leq (u^{(l)})^2$ is required certainly improves the accuracy of the mixed precision arithmetic, and much of our analysis is inspired by TC bFMAs. However, we are trying to leave the analysis general so that other combinations of floating point numbers can be considered as well.
17. line 288: extra plus sign in “ $2(m+2)+$ ”.
Done.
18. line 288-289: “swept that minor difference between under” Some typo/extra word here.
Done.
19. Lemma 3.2: This use of “ δ ” is different than before, and different than that defined in Table 1. I would use some other symbol here.
Done. (τ)
20. line 341: I am confused as to where the “ r ” exponent is coming from.
We changed the product in Lemma 3.2 to range from $j = 1$ to r (it was previously n) to better fit how it is used in the proof of Lemma 3.3.
21. line 359 (and throughout the text): It is incorrect to refer to $\|\mathbf{A} - \hat{\mathbf{Q}}\hat{\mathbf{R}}\|$ as a backward error. This is better referred to as the “residual”. The backward error is the quantity which minimizes the difference $\mathbf{A} - \hat{\mathbf{Q}}\hat{\mathbf{R}}$ over all orthogonal matrices $\hat{\mathbf{Q}}$. The backward error may be bounded in terms of the loss of orthogonality and the residual.
We have added the definition of the backward error that is used in this manuscript in the introductory paragraph of Section 2. While $\|\mathbf{A} - \hat{\mathbf{Q}}\hat{\mathbf{R}}\|$ can be referred to as a residual, we prefer to continue to call it a backward error.
22. line 399: “negligible” spelling
Fixed.
23. line 416: Where is $\Delta X_k^{(j)}$ defined?
This is now explicitly defined.
24. line 441: missing period after displayed equation
Added.
25. line 494: $2^{(l)}$ should be 2^L , correct?
Yes! Thanks for catching that.
26. line 536 (and elsewhere): $m2^{-L}$ would look nicer as $2^{-L}m$
Done.
27. line 548: The γ notation here is not consistent with the rest of the text (gamma with a superscript in parenthesis is used for different precisions; the superscripts here should just be subscripts)
Fixed.

28. lines 552-553: *This case ...*” Ok, but do we really expect this in practice? Say something about this.
We have added: “Note that this likely is not the typical situation in practice.”
29. line 556: “ $2Ln$ ” typo
We got $2Ln$ from the approximation $L\gamma_{2n} = L\frac{2nu}{1-2nu} \approx 2Lnu$.
30. line 684-685: *“Therefore, the loss ...”* But again, there is some constraint on the disparity between precisions in order for the “full precision” assumption to apply, correct?
Yes, but now we have added an additional assumption that there is enough disparity such that full precision products are exact indeed. We have added further analysis to this paragraph to clarify how this can be used in practice.
31. line 694-695: *Why is the exact algorithm for mixed precision TSQR omitted?*
Given the length of the manuscript, we omitted this since the procedure of adding mixed precision to TSQR is similar to how we added mixed precision to BQR (which had been explicitly stated in description and in algorithm.)
32. lines 708-710: *Are there some factors dropped from the displayed equations here?*
We believe these are the correct coefficients for the $\tilde{\gamma}$ terms.
33. lines 778-780: *“Therefore, the numerical stability of mpBQR2 is guaranteed at smaller matrix sizes than the numerical stability of ... BQR in high precision”. I am confused as to how this can be the case.*
This has been reworded in the manuscript. We were pointing out that on top of mpBQR2 having more cast down operations and rounding errors, the small integer constant argument implicit in the $\tilde{\gamma}$ notation differs in the analysis of mpBQR2 than in the analyses of mpBQR3 and in high precision BQR.
34. lines 866: *“... particularly when HQR is unstable due to larger condition numbers”. What is meant by this? Error bounds for HQR do not depend on condition number. (In fact, I don’t see a condition number in any of the bounds in the entire paper).*
The condition number is not a part of our rounding error analysis. We use it in this experiment in our attempt to reach the “worst-case” scenario since that is what our deterministic error bounds bound. This additional explanation has been added.
35. line 870: *“mpTSQR2 can outperform mpHQR2” - “outperform” in terms of what?*
We have added that this is with respect to accuracy.
36. line 885: *“accurately bound” - I would remove or soften this claim, as you just said in the previous paragraph that the error bounds overestimate by 2-3 orders of magnitude (line 874).*
We have removed “accurately”.

Reviewer #2

Main Points:

1. In Lemma 2.2 and in various places in the analyses, $(1 + \theta_k)/(1 + \theta_j) = 1 + \theta_{k+2j}$ is used when $j > k$. As noted in [[2], bottom of p. 67], θ_{k+2j} can be improved to θ_{k+j} when the expressions of θ_k and θ_j are known and given by products of $(1 + \delta_i)^{\pm 1}$ terms. I have not checked all the details carefully but I believe in most if not all places in this article, the improved expression could be used? Could the authors please confirm whether this is correct? This may especially be important in the analysis of section 4.2.1 where a constant of 25 could be avoided.
We have changed the constants in Equation (3.6), (4.16)-(4.19). This has re-

duced the constant in section 4.2.1 from 25 to 14, and now we write $\tilde{\gamma}_{10n}$ instead of noting that the integer c is of a different order. This carries on to some analysis in 4.2.2 and 4.2.3 as well, notably in equations (4.24)-(4.26) and (4.27)-(4.29), but ultimately does not impact the comparison between mpBQR2, mpHQR2, and mpTSQR2.

2. In my first review I suggested using the $\tilde{\gamma}$ notation to hide small constants and simplify the analysis. I wonder whether it would however be useful/interesting to keep track of the constant in front of m , that is, to replace $\gamma_{c_1 m + c_2}$ by $\tilde{\gamma}_{c_1 m}$ instead of $\tilde{\gamma}_m$. I wonder about this because the analysis can apparently sometimes lead to quite large constants, although perhaps my first comment will make this issue go away.

This has been resolved by addressing the first comment.

3. Pages 6 and 7, there is a discussion of the speed benefits associated with MP Setting 2.3. I have found this discussion to be slightly confusing/questionable and not really useful, given this article focuses on error analysis and does not provide any performance experiments, as indicated on L212. I suggest removing any detailed discussion of the speed benefits of MP Setting 2.3 vs the TensorCore setting. In particular, I feel that the distinction of the 3 cases is not really meaningful and could be replaced by a simpler discussion comparing the mixed-precision bound with the uniform precisions bounds $\mu^{(l)}$ and $\mu^{(h)}$.

We have removed the discussion of speed in our analysis totally, but have kept the the 3 cases for clarity.

4. Figure 1 is hard to decipher, mainly due to the use of a color map for the error which can vary by several orders of magnitude. Can the authors use a color map using a log-scale? Or perhaps the figure can be replaced by a simpler, less rich figure with fixed n and varying m on the x -axis, error on the y -axis, and different plots for each variant.

We have replaced the left plot in Figure 1 with the right plot. The color map is on a log-scale and shows clearly that TSQR with 4 levels can exhibit smaller bounds than the other two.

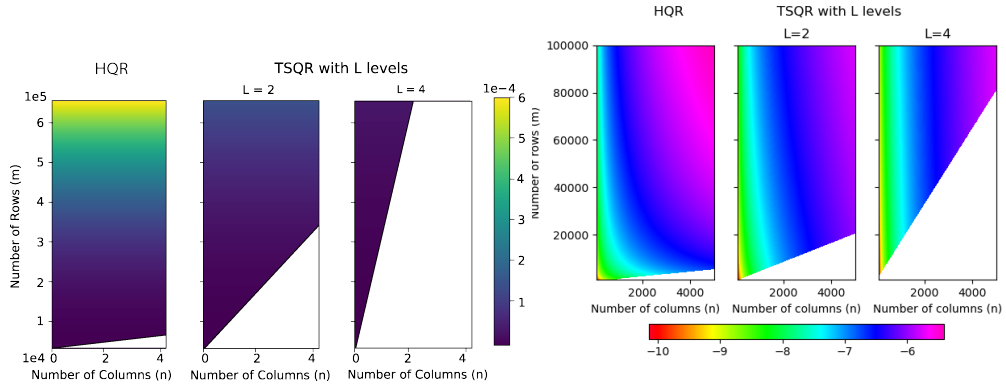


FIG. 1. Left plot: Old; Right plot: New

5. Section 4: I think the paper would read better by interchanging sections 4.1 and 4.2. The bounds of section 4.2 are weaker, and adding them after section 4.1 therefore seems artificial and unnecessary. Instead, giving first the bounds for the more general MP setting 2.3 would allow to present the bFMA/TensorCore setting as a further improvement, which would neatly emphasize the accuracy benefits brought by this type of hardware.

We have made this change to the manuscript. However, this change is not reflected in the diff.pdf I provided to aid with this submission.

6. Equation (4.5): $\gamma_{p/4}^{(fp32)}$ should be replaced by $\gamma_p^{(fp32)}$. In an early preprint version of [1], the former was indeed given under the assumption that NVIDIA tensor cores implement a “true FMA”, that is, that in (4.4) the additions in the 4×4 product AB are done in full precision. However this is not the case, the additions are performed in $fp32$, and so there are still p rounding errors that accumulate in a general product of inner dimension p , as indicated in the published version of [1] (cf. [1, Thm 3.1]). Fortunately I do not think this impacts the analysis of this paper beyond equation (4.5).

The change has been made to the manuscript.

7. The discussion on L684–691 is overall OK, although I feel that there are two points that could be better emphasized.
- First, the sentence “However, as r grows large. . . ” makes it seem as if r should be taken as small as possible for performance, but of course this is not the case since the factorization then also reduces to BLAS-2 operations. In practice, realistic values of r probably range between $\mathcal{O}(10)$ and $\mathcal{O}(100)$, depending on hardware, and not discussing double partitioning strategies. So it is entirely possible that the optimal value of r for performance is already quite large enough so that the accuracy is satisfying.
 - Second, for a fixed r , (4.11) shows that the loss of accuracy becomes less and less significant as m increases and so asymptotically, regardless of the practical value of r , the accuracy eventually will be satisfying. I feel this should be mentioned.

We have discussed the above two points in this part of the discussion.

8. I suggest adding a Table at the end of section 4 or beginning of section 5 summarizing the bounds obtained for each of the 9 algorithms: $\{H,B,TS\}QR$ and $mp\{H,B,TS\}QR\{2,3\}$. It would really be useful to have such a Table, for other articles to refer to, and as reference when comparing with the numerical experiments of section 5.

The table has been added.

9. Figure 2, left plot: I suggest fixing r to a constant rather than $n/4$, which I think is more realistic. Also, have the authors checked what the plots look like for fixed n rather than $n = m/4$? The bounds suggest n should not play a significant role in the relative accuracy of the MP block algorithms vs. the uniform precision ones, but it might be worth checking. **We fixed $n = 250$ (to accommodate for $L = 2$ for TSQR) and $r = 63$. Keeping n small reduced the errors for all of the BLAS-2 methods, so we updated the figure to this version. The high precision and BLAS-3 methods remained pretty much the same.**
10. Figure 2, right plot: the chosen m is a bit small, would it be worth giving the errors for a larger m (which would probably lead to the MP and uniform algorithms to be closer in accuracy)?

We had tried increasing m , but it looked that the matrices needed to be significantly larger to grow sufficiently closer together. Figures with larger matrix sizes are shown in Figure 2. Notably, the rightmost plot shows the $mpBQR3$ orthogonality error to drop below the unit round-off for $fp16$. Below are similar figures with larger matrix sizes. Our goal with this figure was to show that the errors do behave in the same trend as the error bounds suggest and therefore we have kept the figure.

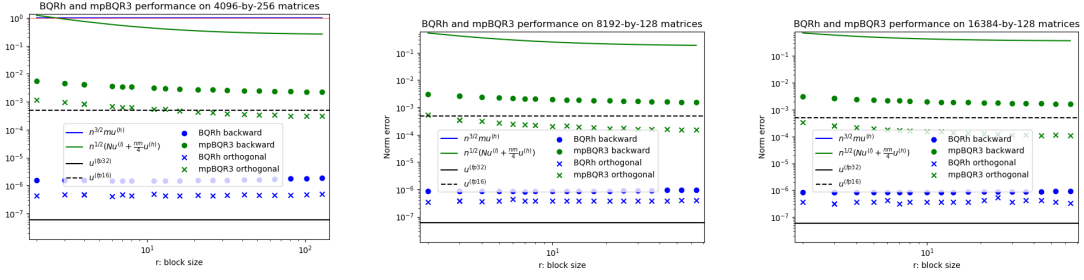


FIG. 2. Left: Old figure; Middle: Larger matrices; Right: Largest matrices.

Stylistic suggestions, typos, and other very minor things:

1. *Everywhere: a HH \rightarrow an HH.*
Done.
2. *L44: columns \rightarrow rows.*
Done.
3. *L120–123: the behavior of the error depending on the data distribution has been explained in the recent preprint [3], which could be referenced here.*
4. *L157: Assumption 2.3 \rightarrow MP Setting 2.3.*
Done.
5. *L175: the notation $x^{(l)}$, $y^{(l)}$ is introduced but does not seem to be used in the subsequent analysis.*
6. *L233–236: I did not fully understand these two sentences. I think the authors are saying that computations introduce error terms of the form (2.9) while castdowns introduce error terms of the form (2.10). Perhaps consider clarifying or removing these two sentences.*
7. *L255, L261, L382, L384, L733: zeros out \rightarrow zeroes out.*
Done.
8. *L288: remove “between”.*
Done.
9. *L294: (3.7) show \rightarrow (3.7) to show ?*
Done.
10. *L295: calculating of $\|x\| \rightarrow$ remove “of”.*
Done.
11. *L316, L317, L319: I suggest rewriting $v^\top x v$ either as $v v^\top x$ or $(v^\top x) v$.*
Done.
12. *(3.13): $\Delta y \rightarrow \Delta x$.*
I think this should be Δy as is stated in (3.10).
13. *L375: constants, $\beta \rightarrow$ missing space.*
Space added.
14. *L385: $k + 1^{st} \rightarrow (k + 1)^{th}$, idem at L420.*
Fixed.
15. *L397, L617: TensorCore \rightarrow TensorCores ?*
Fixed.
16. *L399: negligble \rightarrow negligible.*
Fixed.
17. *L761: “the inner product mixed precision setting yields higher error bounds”: specify than what (add a reference to Thm. 3.4 ?)*

Reference added.

18. L762: I suggest adding a sentence of the form “Before commenting on the significance of Theorem 4.1, we show that the same bounds hold for the BQR variant”. Otherwise section 4.2.1 ends a little abruptly and without any analysis of Thm 4.1 (which is in fact to be found later, L775–L783). In fact, perhaps consider merging sections 4.2.1 and 4.2.2 since the analysis and bounds are essentially the same?

The sentence has been added to better bridge the two sections, but they have been left as two distinct sections.

19. L798: what are $d1$ and $d2$ here?

Oops, an artifact from the first submission had survived! This has been reworded.

20. Sections 4.1.2 and 4.2.3: I suggest writing consistently either $2Ln + m2^{-L}$ (as in section 4.1.2) or $m2^{-L} + 2Ln$ (as in section 4.2.3). **Done.**

21. L885: analysis that accurately bound \rightarrow change to “analyses” or “bounds”.

Done.

22. Reference [5] has now appeared: cf. [1].

The reference has been updated.

REFERENCES

- [1] P. BLANCHARD, N. J. HIGHAM, F. LOPEZ, T. MARY, AND S. PRANESH, *Mixed Precision Block Fused Multiply-Add: Error Analysis and Application to GPU Tensor Cores*, SIAM Journal on Scientific Computing, 42 (2020), pp. C124–C141, <https://doi.org/10.1137/19M1289546>, <https://epubs.siam.org/doi/10.1137/19M1289546>.
- [2] N. J. HIGHAM, *Accuracy and Stability of Numerical Methods*, 2002, <https://doi.org/10.2307/2669725>.
- [3] N. J. HIGHAM AND T. MARY, *Sharper Probabilistic Backward Error Analysis for Basic Linear Algebra Kernels with Random Data*, MIMS EPrint, (2020).
- [4] NVIDIA, *Nvidia Tesla V100 GPU Architecture*, White Paper, (2017), p. 53, <https://images.nvidia.com/content/volta-architecture/pdf/Volta-Architecture-Whitepaper-v1.0.pdf>{%}0Ahttp://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf{%}0Ahttp://www.nvidia.com/content/gated-pdfs/Volta-Architecture-White.

ROUNDING ERROR ANALYSIS OF MIXED PRECISION BLOCK 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 cost 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 develop mixed precision variants of existing Householder QR algorithms and show error analyses supported by numerical experiments.

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~~ 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 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 low precision arithmetic and storage that are less sensitive to floating-point ~~round-off~~ round-off error, such as clustering or ranking graph algorithms [24] or training dense neural networks [20]. As a first step towards that goal, we investigate the use of mixed precision arithmetic for the QR factorization, a widely used linear algebra routine.

Many computing applications today require solutions quickly and often under low size, weight, and power constraints, such as in sensor formation, where low precision computation 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~~ high precision and still achieve speed, size, weight, and power constraint improvements. There have been several recent demonstrations of computing using IEEE half precision (fp16) achieving around half an order to an order of magnitude improvement of these categories in comparison to single and double precision (fp32, fp64). ~~Trivially, the size and weight of memory required for a specific problem is 4×~~. Additionally, there exist demonstrations that the power consumption improvement is similar [10]. Modern accelerators (e.g., GPUs, Knights Landing, or Xeon Phi) are able to achieve this factor or better speedup improvements. Several examples include: (i) 2-4× speedup in solving dense large linear equations [12, 13], (ii) 12× speedup in training dense neural networks, and (iii) 1.2-10× speedup in small batched dense matrix multiplication [1] (up to 26× for batches of tiny matrices). Training deep artificial neural networks by employing lower precision arithmetic to various tasks such as multiplication [6] and storage [7] can easily be implemented on GPUs and are a common practice in some 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 TensorCores perform block Fused Multiply Add operations (bFMAs), where matrix products of fp16 input data can be computed up to 16× faster than that of fp64. The existing rounding error analyses of linear algebra routines are built within what we call a *uniform*

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.

41 *precision* setting, which is the assumption that all arithmetic operations and storage are performed
 42 via the same precision. In this work, we develop mixed precision variants of existing Householder
 43 (HH) QR factorization algorithms and perform mixed precision error analysis. This work focuses on
 44 analyzing a few algorithms that use fp16/fp32 as the low/high precision types, but the error analysis
 45 can be easily modified for different floating point types (such as bfloat16 in [23]). The standard
 46 HH QR algorithm and its block variants that partition the columns (level-3 BLAS variant, see
 47 [11, 14]) and those that partition the ~~columns-rows~~ (communication-avoiding algorithms of [9]) are
 48 presented in section 3, then modified to support bFMAs and an ad hoc mixed precision setting
 49 that mimics NVIDIA TensorCores in section 4. Our key findings are that mixed precision error
 50 analyses produce tighter error bounds as supported by experiments in section 5, algorithms that
 51 utilize level-3 BLAS operations can easily be modified to incorporate TensorCore bFMAs, and a
 52 row partition block algorithm operates more robustly in mixed precision than non-block techniques
 53 in certain regimes.

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

$$57 \quad \mathbf{A} = \mathbf{QR} = [\mathbf{Q}_1 \quad \mathbf{Q}_2] \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0}_{m-n \times n} \end{bmatrix} = \mathbf{Q}_1 \mathbf{R}_1,$$

58 where an orthogonal $\mathbf{Q} \in \mathbb{R}^{m \times m}$ and an upper trapezoidal \mathbf{R} form a *full* QR factorization, and
 59 $\mathbf{Q}_1 \in \mathbb{R}^{m \times n}$, $\mathbf{R}_1 \in \mathbb{R}^{n \times n}$ form a *thin* QR factorization. If \mathbf{A} is full rank then the columns of \mathbf{Q}_1
 60 are orthonormal and \mathbf{R}_1 is upper triangular. In many applications, ~~computing the only a thin~~
 61 decomposition ~~requires-is needed as it produces an orthonormal basis for the column space of \mathbf{A}~~
 62 ~~while requiring~~ less computation and ~~is sufficient in performance. storage.~~

63 Lastly, suppose that \hat{y} is the result of implementing an algorithm that approximates $y = f(x)$.
 64 The forward error is $|\hat{y} - y|$, and the backward error is defined as $|\Delta x|$ or $\min |\Delta x|$ that satisfies
 65 $\hat{y} = f(x + \Delta x)$. We use the first definition of backward error for the remainder of this paper. While
 66 important definitions are stated explicitly in the text, Table 1 serves to establish basic notation.

| Symbol | Definition | Section |
|--|--|---------|
| $\mathbf{x}, \mathbf{A}, \mathbf{x} , \mathbf{A} $ | Vector, matrix, and absolute value of each component | 2 |
| $\ \mathbf{x}\ _p, \ \mathbf{A}\ _p$ | Vector, operator p -norms for $p = 2$, and Frobenius norm when $p = F$. | 2 |
| $\mathbf{x}[i], \mathbf{A}[i, j], :$ | i^{th} element of \mathbf{x} , i^{th} row and j^{th} column element of \mathbf{A} , all indices | 2 |
| $\mathbf{X}_{m \times n}, \mathbf{X}_n$ | m -by- n or n -by- n matrices for \mathbf{X} in $\{\mathbf{0}, \mathbf{I}\}$, $\mathbf{I}_{m \times n} = [\mathbf{I}_n \quad \mathbf{0}_{n \times (m-n)}]^\top$ | 1 |
| $\hat{\mathbf{e}}_i$ | i^{th} cardinal vector | 1 |
| \mathbf{Q}, \mathbf{R} | Factors resulting from Householder (HH) QR factorization algorithms | 2 |
| $\mathbf{P}_v, \mathbf{P}_i$ | HH transformation corresponding to \mathbf{v} , i^{th} HH transformation in HQR | 3 |
| $\mathbf{X}, \mathbf{W}, \mathbf{Y}$ | WY representation of successive HH transformations, $\mathbf{X} = \mathbf{I} - \mathbf{WY}^\top$ | |
| $\text{fl}(\mathbf{x}), \hat{\mathbf{x}}$ | Quantity \mathbf{x} calculated from floating point operations | 2 |
| μ, η | mantissa, exponent bits of a floating point number | 2 |
| $b_q, t_q, u^{(q)}$ | base, precision, unit round-off for precision q , $u^{(q)} := \frac{1}{2}b_q^{1-t_q}$ | 2 |
| $\delta^{(q)}$ | Quantity bounded by: $ \delta^{(q)} < u^{(q)}$ | 2 |
| $\gamma_k^{(q)}, \theta_k^{(q)}$ | $\frac{ku^{(q)}}{1-ku^{(q)}}$, Quantity bounded by: $ \tilde{\theta}_k^{(q)} \leq \tilde{\gamma}_k^{(q)}$ | 2 |
| $\tilde{\gamma}_k^{(q)}, \tilde{\theta}_k^{(q)}$ | $\frac{cku^{(q)}}{1-cku^{(q)}}$ for small integer $c > 0$, Quantity bounded by: $ \theta_k^{(q)} \leq \gamma_k^{(q)}$ | 2 |

TABLE 1

Basic definitions and where they first appear.

2.1. Basic rounding error analysis of floating point operations. We use and analyze the IEEE 754 Standard floating point number systems, shown in Table 2. 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 μ is any integer in $[0, b^t - 1]$ and η is an integer in $[\eta_{\min}, \eta_{\max}]$. 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 [14], which allow a relatively simple framework for formulating error bounds for complex linear algebra operations. An analysis of FLOPs (see Theorem 2.2 [14]) shows that the relative error is controlled by the unit round-off, $u := \frac{1}{2}b^{1-t}$ in uniform precision settings. In mixed precision settings we denote the higher precision unit round-off with $u^{(h)}$ (h for high) and the lower precision unit round-off with $u^{(l)}$ (l for low).

| Name | b | t | # of exponent bits | η_{\min} | η_{\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, δ , bounded

86 in the following sense:

$$87 \quad (2.2) \quad \text{fl}(x \text{ op } y) = (1 + \delta)(x \text{ op } y), \quad |\delta| \leq u, \quad \text{op} \in \{+, -, \times, \div\}.$$

88 We use (2.2) as a building block in accumulating errors from successive FLOPs. Successive operations introduce multiple rounding error terms, and keeping track of all errors is challenging. Lemma 89 2.1 introduces a convenient and elegant bound that simplifies accumulation of rounding error.

91 LEMMA 2.1 (Lemma 3.1 [14]). *Let $|\delta_i| < u$, $\rho_i = \pm 1$ for $i = 1 : k$, and $ku < 1$. Then,*

$$92 \quad (2.3) \quad \prod_{i=1}^k (1 + \delta_i)^{\rho_i} = 1 + \theta_k, \quad \text{where } |\theta_k| \leq \frac{ku}{1 - ku} =: \gamma_k.$$

Additionally, we define $\tilde{\theta}_k$ that satisfies $|\tilde{\theta}_k| \leq \tilde{\gamma}_k$, where $\tilde{\gamma}_k = \frac{cku}{1 - cku}$ for a small integer, $c > 0$.

93 In other words, θ_k represents the accumulation of rounding errors from k successive operations, 94 and it is bounded by γ_k . In more complicated routines shown in later sections, we use the tilde 95 notation ($\tilde{\gamma}_k$) to permit only keeping track of the leading order error terms. Applying this lemma 96 to the computation of $x + y + z$, where $x, y, z \in \mathbb{R}$, results in

$$97 \quad (2.4) \quad \text{fl}(x + y + z) = (1 + \delta')((1 + \delta)(x + y) + z) = (1 + \theta_2)(x + y) + (1 + \theta_1)z,$$

98 where $|\delta|, |\delta'| < u$. Since $|\theta_1| \leq \gamma_1 < \gamma_2$, we can further simplify (2.4) to

$$99 \quad (2.5) \quad \text{fl}(x + y + z) = (1 + \theta'_2)(x + y + z), \quad \text{where } |\theta'_2| \leq \gamma_2,$$

100 at the cost of a slightly larger upper bound. Note that both $|\theta_2|, |\theta'_2|$ are bounded above by γ_2 . 101 Typically, error bounds formed in the fashion of (2.5) are converted to relative errors in order to 102 put the error magnitudes in perspective. The relative error bound for our example is

$$103 \quad |(x + y + z) - \text{fl}(x + y + z)| \leq \gamma_2 |x + y + z|, \quad x + y + z \neq 0.$$

104 Although Lemma 2.1 requires $ku < 1$, we actually need $ku < \frac{1}{2}$ to maintain a meaningful 105 relative error bound as this assumption implies $\gamma_k < 1$ and guarantees a relative error below 106 100%. Since higher precision types have smaller unit round-offs, they can tolerate more successive 107 FLOPs than lower precision floating types before reaching $\gamma_m = 1$. For example, the IEEE types 108 introduced in Table 2 meet this requirement at $1/2 = 2^{10}u^{(\text{fp16})} = 2^{23}u^{(\text{fp32})} = 2^{52}u^{(\text{fp64})}$. Thus, 109 accumulated rounding errors in lower precision types can lead to an instability with fewer operations 110 in comparison to higher precision types and prompts us to evaluate whether existing algorithms 111 can be naively adapted for mixed precision arithmetic.

112 **2.2. Rounding Error Example for the Inner Product.** We now consider computing the 113 inner product of two vectors to clearly illustrate how this situation restricts rounding error analysis 114 in fp16. An error bound for an inner product of m -length vectors is

$$115 \quad (2.6) \quad |\mathbf{x}^\top \mathbf{y} - \text{fl}(\mathbf{x}^\top \mathbf{y})| \leq \gamma_m |\mathbf{x}^\top \mathbf{y}|, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^m$$

116 as shown in [14]. Since vectors of length m accumulate rounding errors that are bounded by γ_m , 117 dot products of vectors computed in fp16 already face a 100% relative error bound when $m = 1024$.

A simple numerical experiment shows that the standard deterministic error bound is too pessimistic and cannot be practically used to approximate rounding error for ~~half-precision-half~~ precision arithmetic. In this experiment, we generated 2 million random fp16 vectors of length 1024 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 [\[16\]](#), for every FLOP (multiplication and addition for the dot product). The relative error in this experiment is formulated as the LHS in Equation 2.6 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 3 shows some statistics from computing the relative error for simulated fp16 dot products.

| Random Distribution | Average | Stan. Dev. | Maximum |
|---------------------|-----------|------------|-----------|
| Standard normal | 1.621e-04 | 1.635e-04 | 3.204e-03 |
| Uniform (0, 1) | 6.904e-03 | 3.265e-03 | 2.447e-02 |

TABLE 3

Forward error statistics from experiment of dot products computed in simulated half precision.

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 [\[15, 18\]](#)) 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: $\mathbf{z}^{(\text{fp16})} = \text{simHalf}(f, \mathbf{x}^{(\text{fp16})}, \mathbf{y}^{(\text{fp16})})$. Given fp16 input variables \mathbf{x}, \mathbf{y} , perform function $f \in \text{OP} \cup \{\text{dot_product}\}$ in simulated fp16 arithmetic.

| | |
|---|---|
| <p>Input: $\mathbf{x}^{(\text{fp16})}, \mathbf{y}^{(\text{fp16})}, f$</p> <p>1 $[\mathbf{x}^{(\text{fp32})}, \mathbf{y}^{(\text{fp32})}] \leftarrow \text{castup}([\mathbf{x}^{(\text{fp16})}, \mathbf{y}^{(\text{fp16})}])$</p> <p>2 $\mathbf{z}^{(\text{fp32})} \leftarrow \text{fl}(f(\mathbf{x}^{(\text{fp32})}, \mathbf{y}^{(\text{fp32})}))$</p> <p>3 $\mathbf{z}^{(\text{fp16})} \leftarrow \text{castdown}(\mathbf{z}^{(\text{fp32})})$</p> <p>4 return $\mathbf{z}^{(\text{fp16})}$</p> | <p>Output: $\mathbf{z}^{(\text{fp16})} = \text{fl}_{\text{fp16}}(f(\mathbf{x}^{(\text{fp16})}, \mathbf{y}^{(\text{fp16})}))$</p> <p>// Convert input vars to fp32.</p> <p>// Perform fp32 arithmetic.</p> <p>// Convert result to fp16.</p> |
|---|---|

Most importantly, we need error analysis that allows flexibility in precision in order to better our understanding of the impact of rounding errors on computations done on emerging hardware (i.e. GPUs) that support mixed precision. We start by introducing some additional rules from [\[14\]](#) that build on [Lemma 2.1](#) in [Lemma 2.2](#). These rules summarize how to accumulate errors represented by θ 's and γ 's in a *uniform precision* setting.

LEMMA 2.2. *For any positive integer k , let θ_k denote a quantity bounded according to $|\theta_k| \leq \frac{k\epsilon}{1-k\epsilon} =: \gamma_k$. The following relations hold for positive integers j, n and nonnegative integer k .*

143 Arithmetic operations between bounded terms, θ_k 's, are:

$$144 \quad (2.7) \quad (1 + \theta_k)(1 + \theta_j) = (1 + \theta_{k+j}) \quad \text{and} \quad \frac{1 + \theta_k}{1 + \theta_j} = \begin{cases} 1 + \theta_{k+j}, & j \leq k \\ 1 + \theta_{k+2j}, & j > k \end{cases}.$$

145 If $\max_{(j,k)} u \leq \frac{1}{2}$ and $n \leq \frac{1}{uk}$, the operations on the bounds, γ 's, are:

$$146 \quad \gamma_k \gamma_j \leq \gamma_{\min(k,j)}, \quad n\gamma_k \leq \gamma_{nk}, \\ 147 \quad \gamma_k + u \leq \gamma_{k+1}, \quad \gamma_k + \gamma_j + \gamma_k \gamma_j \leq \gamma_{k+j}.$$

149 Note that all the rules hold when replaced by $\tilde{\gamma}$'s, but result in looser bounds.

150 We define two mixed precision settings that we use in section 4. In subsection 4.1, we present
151 the block Fused Multiply-Add (bFMA) of NVIDIA's TensorCore (TC) technology, which computes
152 matrix-matrix multiply and accumulate for 4-by-4 blocks, and incorporate it into algs. 5 and 6. Here,
153 we introduce an ad hoc mixed precision setting (MP Setting) which we use in subsection 4.2. This
154 is explicitly defined in MP Setting 2.3 and is a level-2 BLAS variant of the TC bFMA. Both mixed
155 precision settings define how inner products are computed although the bFMA is only applicable to
156 inner products within matrix products and uses fp16 and fp32 whereas our ad hoc mixed precision
157 setting is applicable to all inner products with any two precision types.

158 ~~Although our analysis concerns accuracy and stability and leaves~~
159 ~~out timing results of various hardware, we add a general timing statement to MP Setting 2.3 that~~
160 ~~is analogous to that of TC: the mixed precision FMA inner product performs at least 2 times faster~~
161 ~~than the inner product in the higher precision.~~ Note that TCs perform matrix-matrix multiply
162 and accumulate up to 8 times faster than fp32, and up to 16 times faster than fp64 (see [19]), and our
163 ad hoc timing assumption is in conservative in comparison. Nonetheless, this gives a vague insight
164 into the trade-offs between speediness and accuracy from some mixed precision computations.

165 The full precision multiplication in Assumption 2.3 is exact. The exact product in MP Setting 2.3
166 is the result of using full precision products when the low precision type is fp16 and the high precision
167 type of is fp32 due to their precisions and exponent ranges as is in TC bFMAs. As a quick proof,
168 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 $\eta_x, \eta_y \in [-15, 16]$,
169 and note that the significand and exponent ranges for fp32 are $[0, 2^{24} - 1]$ and $[-127, 128]$. Then
170 the product in full precision is

$$171 \quad x^{(\text{fp16})} y^{(\text{fp16})} = \pm \mu_x \mu_y 2^{\eta_x + \eta_y + 2 - 24},$$

172 where $\mu_x \mu_y \in [0, (2^{11} - 1)^2] \subseteq [0, 2^{24} - 1]$ and $\eta_x + \eta_y + 2 \in [-28, 34] \subseteq [-127, 128]$. Thus, when
173 two fp16 numbers are multiplied and stored in fp32, there is no roundoff error, and therefore is
174 exact. Thus, the summation and the final cast down operations are the only sources of rounding
175 error in this inner product scheme.

176 MP SETTING 2.3. Let l and h each denote low and high precision types with unit round-off
177 values $u^{(l)}$ and $u^{(h)}$, where $1 \gg u^{(l)} \gg u^{(h)} > 0$. Consider an FMA operation for inner products that
178 take vectors stored in precision l , compute products in full precision exactly, and sum the products
179 in precision h . Finally, the result is then cast back down to precision l . Furthermore, we expect this
180 procedure to be approximately twice as fast as if it were done entirely in the higher precision, and
181 about the same as if it were done entirely in the lower precision.

We now analyze the rounding error for the inner product scheme described in [MP Setting 2.3](#) and hypothesize that the guaranteed accuracy for this mixed precision inner product should be better than that of the low precision inner product and worse than that of the high precision inner product. Let $\mathbf{x}^{(l)}, \mathbf{y}^{(l)}$ be m -length vectors stored in a low precision type (\mathbb{F}_l^m), s_k be the exact k^{th} partial sum, and \hat{s}_k be s_k computed with FLOPs. Then the first three partial sums are,

$$\begin{aligned}\hat{s}_1 &= \text{fl}(\mathbf{x}[1]\mathbf{y}[1]) = \mathbf{x}[1]\mathbf{y}[1], \quad \hat{s}_2 = \text{fl}(\hat{s}_1 + \mathbf{x}[2]\mathbf{y}[2]) = (\mathbf{x}[1]\mathbf{y}[1] + \mathbf{x}[2]\mathbf{y}[2]) (1 + \delta_1^{(h)}), \\ \hat{s}_3 &= \text{fl}(\hat{s}_2 + \mathbf{x}[3]\mathbf{y}[3]) = \left[(\mathbf{x}[1]\mathbf{y}[1] + \mathbf{x}[2]\mathbf{y}[2]) (1 + \delta_1^{(h)}) + \mathbf{x}[3]\mathbf{y}[3] \right] (1 + \delta_2^{(h)}).\end{aligned}$$

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

$$(2.8) \quad \hat{s}_m = (\mathbf{x}[1]\mathbf{y}[1] + \mathbf{x}[2]\mathbf{y}[2]) \prod_{k=1}^{m-1} (1 + \delta_k^{(h)}) + \sum_{i=3}^n \mathbf{x}[i]\mathbf{y}[i] \left(\prod_{k=i-1}^{m-1} (1 + \delta_k^{(h)}) \right).$$

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

$$(2.9) \quad \text{fl}(\mathbf{x}^\top \mathbf{y}) = (\mathbf{x} + \Delta \mathbf{x})^\top \mathbf{y} = \mathbf{x}^\top (\mathbf{y} + \Delta \mathbf{y}) \quad \text{for } |\Delta \mathbf{x}| \leq \gamma_{m-1}^{(h)} |\mathbf{x}|, \quad |\Delta \mathbf{y}| \leq \gamma_{m-1}^{(h)} |\mathbf{y}|.$$

Casting down to \mathbb{F}_l without underflow or overflow results in backward errors,

$$(2.10) \quad \text{castdown}(\text{fl}(\mathbf{x}^\top \mathbf{y})) = (\mathbf{x} + \Delta \mathbf{x} + \tilde{\Delta} \mathbf{x})^\top \mathbf{y} = \mathbf{x}^\top (\mathbf{y} + \Delta \mathbf{y} + \tilde{\Delta} \mathbf{y}),$$

where $|\Delta \mathbf{x} + \tilde{\Delta} \mathbf{x}| \leq ((1 + u^{(l)})(1 + \gamma_{m-1}^{(h)}) - 1) |\mathbf{x}|$ and $|\Delta \mathbf{y} + \tilde{\Delta} \mathbf{y}| \leq ((1 + u^{(l)})(1 + \gamma_{m-1}^{(h)}) - 1) |\mathbf{y}|$. Our hypothesis is indeed true since,

$$\gamma_m^{(h)} < u^{(l)} + \gamma_{m-1}^{(h)} + u^{(l)} \gamma_{m-1}^{(h)} < \gamma_m^{(l)},$$

where the lower and upper bounds are derived from the uniform precision error bound in (2.6). Equation (2.10) shows us that the two larger error terms are from the higher precision summation, $\gamma_{m-1}^{(h)}$, and the cast down operation, $u^{(l)}$. We can measure the impact of the cast down step relative to the length of the vector, m , and the disparity in the two precisions, $M_{l,h} := u^{(l)}/u^{(h)}$, since these two factors determine which one of $u^{(l)}$ and $mu^{(h)}$ is the leading order term. We consider 3 cases.

Case 1: ($m \ll M_{l,h}$) The leading order term is $u^{(l)}$. The mixed precision inner product has a smaller worst case error bound than the bound of the low precision inner product ($mu^{(l)}$) ~~with no apparent improvements in speed.~~ On the other hand, $u^{(l)}$ is a larger upper bound than that of the high precision inner product ($mu^{(h)} = \frac{m}{M_{l,h}} u^{(l)}$), ~~although it was computed approximately twice as fast.~~ It is likely that this factor of $M_{l,h}/m$ increase in the worst case error bound is unwanted ~~even when considering the speed-up.~~

Case 2: ($m = M_{l,h}$) Both terms are now leading order. This is still an improvement in comparison to the lower precision arithmetic as the error bound is reduced from $mu^{(l)}$ to $2u^{(l)}$. Comparing this to the high precision inner product shows that the error bound has doubled from $mu^{(h)}$ to $2mu^{(h)}$; ~~but gained a factor of 2 in speed instead. One can argue that the loss in accuracy guarantee and the improvement in speed cancel each other out especially if $2mu^{(h)} \ll 1$ or if the speed-up greatly exceeds a factor of 2.~~ **Case 3:** ($m \gg M_{l,h}$) Now $\gamma_{m-1}^{(h)}$ is the leading order term. As in the above two cases, this is an improvement in the context of the low precision accuracy since the error

has been reduced from $\gamma_m^{(l)}$ to $\gamma_{m/M_{l,h}}^{(l)} \equiv \gamma_m^{(h)}$. Since $u^{(l)} = M_{l,h}u^{(h)} \ll mu^{(h)}$, the mixed precision error bound has the same *order* as the error bound from carrying the computation out in the higher precision. Therefore, we can expect about the same level of accuracy ~~but a factor of 2 or greater reduction in speed when compared to the higher precision.~~

~~While the above cases establish 3 regimes of trade-offs between accuracy and speed in mixed precision computing, the remainder of this paper focuses only on accuracy and does not consider the impact of mixed precision computations on speed.~~ Finally, we present alternative representations of the error bound in (2.10),

$$(1 + u^{(l)})(1 + \gamma_{m-1}^{(h)}) - 1 \leq \gamma_{M_{l,h}m-1}^{(h)} = \gamma_{1+(m-1)/M_{l,h}}^{(l)}, \quad M_{l,h} = u^{(l)}/u^{(h)},$$

$$(1 + u^{(l)})(1 + \gamma_{m-1}^{(h)}) - 1 \leq u^{(l)} + \gamma_{m-1}^{(h)} + \min\{u^{(l)}, \gamma_{m-1}^{(h)}\}, \quad \gamma_{m-1}^{(h)} < 1,$$

where the rules from Lemma 2.2 were directly applied. Both alternative bounds are only slightly larger than the original bound shown on the LHS and remain in the same order. The first is useful when comparing against the low or the high precision, whereas the second keeps track of the error bounds in both precisions. We summarize these ways of combining γ terms of different precisions in Lemma 2.4,

LEMMA 2.4. *For any nonnegative integers k_l, k_h and some precision q defined with respect to the unit round-off, $u^{(q)}$, define $\gamma_k^{(q)} := \frac{ku^{(q)}}{1-ku^{(q)}}$. Consider a low precision and a high precision where $1 \gg u^{(l)} \gg u^{(h)} > 0$, and k_l, k_h that satisfy $\max\{\gamma_{k_h}^{(h)}, \gamma_{k_l}^{(l)}\} < 1/2$. Then the following rules help us accumulate γ 's of different precisions,*

$$(2.11) \quad \gamma_{k_h}^{(h)} \gamma_{k_l}^{(l)} \leq \min\{\gamma_{k_h}^{(h)}, \gamma_{k_l}^{(l)}\},$$

$$(2.12) \quad (1 + \tilde{\gamma}_{k_l}^{(l)})(1 + \tilde{\gamma}_{k_h}^{(h)}) - 1 = \tilde{\gamma}_{k_l}^{(l)} + \tilde{\gamma}_{k_h}^{(h)}.$$

Note that (2.12) drops the term $\tilde{\gamma}_{k_l}^{(l)}\tilde{\gamma}_{k_h}^{(h)}$ since both $\tilde{\gamma}_{k_l}^{(l)}$ and $\tilde{\gamma}_{k_h}^{(h)}$ are larger than their product and this product can be swept under the small integer $c > 0$ assumption implicitly included in the tilde notation. the two mixed precision settings add **castdown** operations at different parts of the HQR algorithms we consider. In general, error bounds in the fashion of (2.9) can be used before the cast down operations and the action of the cast down is best represented by error bounds similar to (2.10).

We have demonstrated a need for rounding error analysis that is accurate for mixed precision procedures and analyzed the inner product in an ad hoc mixed precision inner product that mimics the TensorCore bFMA. We will use this to analyze various Householder (HH) QR factorization algorithms. Algorithms and the general framework for the standard rounding error analysis for these algorithms are introduced in section 3, and both are modified to meet different mixed precision assumptions in section 4.

3. Algorithms and existing round-off error analyses. We introduce the Householder QR factorization algorithm (HQR) in subsection 3.1 and two block variants that use HQR within the block in subsections 3.2 and 3.3. The blocked HQR (BQR) in subsection 3.2 partitions the columns of the target matrix and is a well-known algorithm that uses the WY representation of [4] that utilizes mainly level-3 BLAS operations. In contrast, the Tall-and-Skinny QR (TSQR) in subsection 3.3 partitions the rows and takes a communication-avoiding divide-and-conquer approach that can be easily parallelized (see [8]). We present the standard rounding error analysis of these algorithms (see [14, 21]) which will be tweaked for various mixed precision assumptions in section 4.

3.1. Householder QR (HQR). The HQR algorithm uses HH transformations to zero out elements below the diagonal of a matrix (see [17]). We present this as zeroing out all but the first element of some vector, $\mathbf{x} \in \mathbb{R}^m$.

LEMMA 3.1. *Given vector $\mathbf{x} \in \mathbb{R}^m$, there exist an HH vector, \mathbf{v} , and an HH constant, β , that define the HH transformation matrix, $\mathbf{P}_{\mathbf{v}} := \mathbf{I}_m - \beta \mathbf{v} \mathbf{v}^\top$, such that $\mathbf{P}_{\mathbf{v}}$ zeroes out \mathbf{x} below the first element. The HH vector and constant are defined via*

$$(3.1) \quad \sigma = -\text{sign}(\mathbf{x}[1]) \|\mathbf{x}\|_2, \quad \mathbf{v} = \mathbf{x} - \sigma \hat{\mathbf{e}}_1, \quad \text{and} \quad \beta = \frac{2}{\mathbf{v}^\top \mathbf{v}} = -\frac{1}{\sigma \mathbf{v}[1]}.$$

The transformed vector, $\mathbf{P}_{\mathbf{v}} \mathbf{x} = \sigma \hat{\mathbf{e}}_1$, has the same 2-norm as \mathbf{x} since $\mathbf{P}_{\mathbf{v}} = \mathbf{P}_{\mathbf{v}}^\top = \mathbf{P}_{\mathbf{v}}^{-1}$.

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

Consider the following 4-by-3 matrix example adapted from [14]. Let \mathbf{P}_i represent the i^{th} HH transformation of this algorithm.

$$\mathbf{A} = \begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix} \xrightarrow{\mathbf{P}_1 \mathbf{A}} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \end{bmatrix} \xrightarrow{\mathbf{P}_2 \mathbf{P}_1 \mathbf{A}} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & \times \end{bmatrix} \xrightarrow{\mathbf{P}_3 \mathbf{P}_2 \mathbf{P}_1 \mathbf{A}} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & 0 \end{bmatrix}$$

The resulting matrix is the \mathbf{R} factor, $\mathbf{R} := \mathbf{P}_3 \mathbf{P}_2 \mathbf{P}_1 \mathbf{A}$, and 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. The thin factors for a general matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ are

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

Algorithm 2: $\beta, \mathbf{v}, \sigma = \text{hhvec}(\mathbf{x})$. Given a vector $\mathbf{x} \in \mathbb{R}^m$, return $\mathbf{v} \in \mathbb{R}^m$ and $\beta, \sigma \in \mathbb{R}$ that satisfy $(\mathbf{I} - \beta \mathbf{v} \mathbf{v}^\top) \mathbf{x} = \sigma \hat{\mathbf{e}}_1$ and $\mathbf{v}[1] = 1$ (see [2, 14]).

Input: \mathbf{x}

Output: \mathbf{v}, σ , and β

```

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  $\mathbf{v} \leftarrow \mathbf{v} / \mathbf{v}[1]$ 
6 return  $\beta, \mathbf{v}, \sigma$ 
```

3.1.2. HQR: Rounding Error Analysis. Now we present an error analysis for alg. 3 by keeping track of the different operations of alg. 2 and alg. 3. We follow the analysis of [14] and modify it for the variant where $\mathbf{v}[1]$ is set to 1. The goal of this section is to present the basic steps of the standard error analysis for HQR so that we modify them easily in section 4 for different mixed precision settings.

Calculating the i^{th} HH vector and constant. In alg. 3, we compute the HH vector and constant by using alg. 2 to $\mathbf{A}[i : m, i]$. For now, consider zeroing out any vector $\mathbf{x} \in \mathbb{R}^m$ below its first component with an HH transformation. We first calculate σ as is implemented in line 2 of alg. 2.

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

Algorithm 3: $\mathbf{V}, \beta, \mathbf{R} = \text{HQR2}(\mathbf{A})$. A Level-2 BLAS implementation of HQR. 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} . The orthogonal factor \mathbf{Q} can be generated from \mathbf{V} and β .

| | |
|----------------------------|--|
| Input: \mathbf{A} | Output: $\mathbf{V}, \beta, \mathbf{R}$ |
|----------------------------|--|

```

1 Initialize  $\mathbf{V} \leftarrow \mathbf{0}_{m \times n}$ ,  $\beta \leftarrow \mathbf{0}_m$ 
2 for  $i = 1 : n$  do
3    $\mathbf{v}, \beta, \sigma \leftarrow \text{hhvec}(\mathbf{A}[i : \text{end}, i])$  /* 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  $\mathbf{V}, \beta, \mathbf{A}[1 : n, 1 : n]$ 

```

288 Note that the backward error incurred here accounts for an inner product of a vector in \mathbb{R}^m with
 289 itself and a square root operation to get the 2-norm. Let $\mathbf{v}'[1] \equiv \mathbf{x}[i] - \sigma$, the penultimate value
 290 $\mathbf{v}[1]$ held. The subtraction adds a single additional rounding error via

$$291 \quad (3.4) \quad \text{fl}(\mathbf{v}'[1]) = \mathbf{v}'[1] + \Delta \mathbf{v}'[1] = (1 + \delta)(\mathbf{x}[i] - \sigma - \Delta \sigma) = (1 + \theta_{m+2})\mathbf{v}'[1]$$

292 where the last equality is granted because the sign of σ is chosen to prevent cancellation. Since
 293 [alg. 2](#) normalizes the HH vector so that its first component is 1, the remaining components of \mathbf{v}
 294 are divided by $\text{fl}(\tilde{\mathbf{v}}_1)$ incurring another single rounding error. As a result, the components of \mathbf{v}
 295 computed with FLOPs have error $\text{fl}(\mathbf{v}[j]) = \mathbf{v}[j] + \Delta \mathbf{v}[j]$ where

$$296 \quad (3.5) \quad |\Delta \mathbf{v}[j]| \leq \gamma_{1+2(m+2)} |\mathbf{v}[j]| = \tilde{\gamma}_m |\mathbf{v}[j]| \quad j = 2 : m - i + 1,$$

297 and $|\Delta \mathbf{v}[1]| = 0$. Since $\text{fl}(\mathbf{v}[1]) = \mathbf{v}[1] + \Delta \mathbf{v}[1] = \mathbf{v}[1] + \Delta \sigma$, we have swept that minor
 298 difference ~~between~~ under our use of the $\tilde{\gamma}$ notation defined in [Lemma 2.1](#). Next, we consider the
 299 HH constant, β , as is computed in line 4 of [alg. 2](#).

$$300 \quad (3.6) \quad \hat{\beta} = \text{fl}(-\mathbf{v}'[1]/\hat{\sigma}) = -(1 + \delta) \frac{\mathbf{v}'[1] + \Delta \mathbf{v}'[1]}{\sigma + \Delta \sigma} = \frac{(1 + \delta)(1 + \theta_{m+2})}{(1 + \theta_{m+1})} \beta$$

$$301 \quad (3.7) \quad = (1 + \theta_{3m+52m+4}) \beta = \beta + \Delta \beta, \text{ where } |\Delta \beta| \leq \tilde{\gamma}_m \beta.$$

302

303 We have shown (3.6) to keep our analysis simple in [section 4](#) and (3.7) to show that the error
 304 incurred from calculating $\text{fl}(\|\mathbf{x}\|_2)$ accounts for the vast majority of the rounding error so far. At
 305 iteration i , we replace \mathbf{x} with $\mathbf{A}[i : m, i] \in \mathbb{R}^{m-i+1}$ and the i^{th} HH constant and vector $(\hat{\beta}_i, \mathbf{v}_i)$ both
 306 have errors bounded by $\tilde{\gamma}_{m-i+1}$.

307 *Applying a Single HH Transformation.* Now we consider lines 4-6 of [alg. 3](#). At iteration i ,
 308 we set $\mathbf{A}[i + 1 : m, :]$ to zero and replace $\mathbf{A}[i, i]$ with σ computed from [alg. 2](#). Therefore, we
 309 now need to calculate the errors for applying ~~a~~ [an](#) HH transformation to the remaining columns,
 310 $\mathbf{A}[i : m, i + 1 : n]$ with the computed HH vector and constant. This is the most crucial building
 311 block of the rounding error analysis for any variant of HQR because the \mathbf{R} factor is formed by
 312 applying the HH transformations to \mathbf{A} and the \mathbf{Q} factor is formed by applying them in reverse
 313 order to the identity. Both of the blocked versions in [subsection 3.2](#) and [subsection 3.3](#) also require
 314 slightly different but efficient implementations of this step. For example, BQR in [alg. 5](#) uses level-3

BLAS operations to apply multiple HH transformations at once whereas the variant of HQR in [alg. 3](#) can only use level-2 BLAS operations to apply HH transformations.

A HH transformation is applied through a series of inner and outer products, since HH matrices are rank-1 updates of the identity. That is, computing $\mathbf{P}_\mathbf{v}\mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^m$ is as simple as computing

$$(3.8) \quad \mathbf{y} := \mathbf{P}_\mathbf{v}\mathbf{x} = \mathbf{x} - (\beta \mathbf{v}^\top \mathbf{x}) \mathbf{v}.$$

Let us assume that \mathbf{x} is an exact vector and there were errors incurred in forming \mathbf{v} and β . The errors incurred from computing \mathbf{v} and β need to be 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} is any 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 $i-1$ HH transformations that zeroed out components below $\mathbf{A}[j, j]$ for $j = 1:i-1$. We show the error for forming $\hat{\mathbf{w}}$ where $\hat{\mathbf{w}} := \beta \mathbf{v}^\top \mathbf{x} \mathbf{v} - \mathbf{w} := \beta (\mathbf{v}^\top \mathbf{x}) \mathbf{v}$ and $\mathbf{v}, \mathbf{x} \in \mathbb{R}^m$,

$$\hat{\mathbf{w}} = \text{fl}(\hat{\beta} \text{fl}(\hat{\mathbf{v}}^\top \mathbf{x}) \hat{\mathbf{v}}) = (1 + \theta_m)(1 + \delta)(1 + \delta')(\beta + \Delta\beta) ((\mathbf{v} + \Delta\mathbf{v})^\top \mathbf{x}) (\mathbf{v} + \Delta\mathbf{v}),$$

where θ_m is from computing the inner product $\hat{\mathbf{v}}^\top \mathbf{x}$, and δ and δ' are from multiplying β , $\text{fl}(\hat{\mathbf{v}}^\top \mathbf{x})$, and $\hat{\mathbf{v}}$. The forward error is $\hat{\mathbf{w}} = \mathbf{w} + \Delta\mathbf{w}$, where $|\Delta\mathbf{w}| \leq \tilde{\gamma}_m |\beta| |\mathbf{v}| |\mathbf{v}^\top \mathbf{x}| |\mathbf{v}|$ and $|\Delta\mathbf{w}| \leq \tilde{\gamma}_m |\beta| (|\mathbf{v}^\top \mathbf{x}|) |\mathbf{v}|$. Subtracting $\hat{\mathbf{w}}$ from \mathbf{x} yields the HH transformation with forward error,

$$(3.9) \quad \text{fl}(\mathbf{P}_\mathbf{v}\mathbf{x}) = \text{fl}(\mathbf{x} - \hat{\mathbf{w}}) = (1 + \delta)(\mathbf{x} - \mathbf{w} - \Delta\mathbf{w}) = \mathbf{y} + \Delta\mathbf{y} = (\mathbf{P}_\mathbf{v} + \Delta\mathbf{P}_\mathbf{v})\mathbf{x},$$

where $|\Delta\mathbf{y}| \leq u|\mathbf{x}| + \tilde{\gamma}_m |\beta| |\mathbf{v}| |\mathbf{v}^\top \mathbf{x}|$. Using $\sqrt{2/\beta} = \|\mathbf{v}\|_2$, we form a normwise bound,

$$(3.10) \quad \|\Delta\mathbf{y}\|_2 \leq \tilde{\gamma}_m \|\mathbf{x}\|_2.$$

Since $\Delta\mathbf{P}_\mathbf{v}[i, j] = \frac{1}{\|\mathbf{x}\|_2^2} \Delta\mathbf{y}[i] \mathbf{x}[j]$, we can compute its Frobenius norm,

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

where the last inequality is a direct application of (3.10).

Applying many successive HH 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 HH transformations computed with $\hat{\mathbf{v}}_i$'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 $\mathbf{R} = \mathbf{Q}^\top \mathbf{A} = \mathbf{P}_n \cdots \mathbf{P}_1 \mathbf{A}$. [Lemma 3.2](#) is very useful for any sequence of transformations, where each transformation has a known bound. We will invoke this lemma to prove [Lemma 3.3](#), and use it in future sections for other consecutive transformations.

LEMMA 3.2. *If $\mathbf{X}_j + \Delta\mathbf{X}_j \in \mathbb{R}^{m \times m}$ satisfies $\|\Delta\mathbf{X}_j\|_F \leq \delta_j \|\mathbf{X}_j\|_2$ for all j , then $\|\Delta\mathbf{X}_j\|_F \leq \tau_j \|\mathbf{X}_j\|_2$ for all $j = 1, \dots, r$, then*

$$\left\| \prod_{j=1}^r (\mathbf{X}_j + \Delta\mathbf{X}_j) - \prod_{j=1}^r \mathbf{X}_j \right\|_F \leq \left(-1 + \prod_{j=1}^r (1 + \delta_j) \right) \prod_{j=1}^r \|\mathbf{X}_j\|_2.$$

LEMMA 3.3. *Consider applying a sequence of transformations $\mathbf{Q} = \mathbf{P}_r \cdots \mathbf{P}_2 \mathbf{P}_1$ onto vector $\mathbf{x} \in \mathbb{R}^m$ to form $\hat{\mathbf{y}} = \text{fl}(\hat{\mathbf{P}}_r \cdots \hat{\mathbf{P}}_2 \hat{\mathbf{P}}_1 \mathbf{x})$, where $\hat{\mathbf{P}}_k$'s are HH transformations constructed from $\hat{\beta}_k$ and*

344 $\hat{\mathbf{v}}_k$. These HH vectors and constants are computed via [alg. 2](#) and the rounding errors are bounded
 345 by (3.5) and (3.7). If each transformation is computed via (3.8), then

$$346 \quad (3.12) \quad \hat{\mathbf{y}} = \mathbf{Q}(\mathbf{x} + \Delta\mathbf{x}) = (\mathbf{Q} + \Delta\mathbf{Q})\mathbf{x} = \hat{\mathbf{Q}}\mathbf{x},$$

$$347 \quad (3.13) \quad \|\Delta\mathbf{y}\|_2 \leq r\tilde{\gamma}_m\|\mathbf{x}\|_2, \quad \|\Delta\mathbf{Q}\|_F \leq r\tilde{\gamma}_m.$$

349 *Proof.* Applying [Lemma 3.2](#) directly to \mathbf{Q} yields

$$350 \quad \|\Delta\mathbf{Q}\|_F = \left\| \prod_{j=1}^r (\mathbf{P}_j + \Delta\mathbf{P}_j) - \prod_{j=1}^r \mathbf{P}_j \right\|_F \leq \left(-1 + \prod_{j=1}^r (1 + \tilde{\gamma}_{m-j+1})^r \right) \prod_{j=1}^n \|\mathbf{P}_j\|_2 \leq -1 + (1 + \tilde{\gamma}_m)^r,$$

351 since \mathbf{P}_j 's are orthogonal and have 2-norm, 1, and $m - j + 1 \leq m$. While we omit the details here,
 352 we can show that $(1 + \tilde{\gamma}_m)^r - 1 \leq r\tilde{\gamma}_m$ using the argument from [Lemma 2.1](#) if $r\tilde{\gamma}_m \leq 1/2$. \square

353 In this error analysis, the prevailing bound for errors at various stages of forming and applying
 354 a [an](#) HH transformation is $\tilde{\gamma}_m$ where m corresponds to the dimension of the transformed vectors.
 355 In [Lemma 3.3](#), a factor of r is introduced for applying r HH transformations to form the term
 356 $r\tilde{\gamma}_m \approx rmu$. Therefore, we can expect that the columnwise norm error for a thin QR factorization
 357 should be $\mathcal{O}(mnu)$ for a full rank matrix. In [Theorem 3.4](#), we formalize this by applying [Lemma 3.3](#)
 358 directly and also show a conversion of columnwise bounds to a matrix norm bound,

$$359 \quad \|\Delta\mathbf{R}\|_F = \left(\sum_{i=1}^n \|\Delta\mathbf{R}[:, i]\|_2^2 \right)^{1/2} \leq \left(\sum_{i=1}^n n^2 \tilde{\gamma}_m^2 \|\mathbf{A}[:, i]\|_2^2 \right)^{1/2} = n\tilde{\gamma}_m \|\mathbf{A}\|_F.$$

360 We gather these results into [Theorem 3.4](#).

361 **THEOREM 3.4.** 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
 362 the thin QR factors of \mathbf{A} obtained via [alg. 3](#). Then,

$$363 \quad \hat{\mathbf{R}} = \mathbf{R} + \Delta\mathbf{R} = \text{fl}(\hat{\mathbf{P}}_n \cdots \hat{\mathbf{P}}_1 \mathbf{A}), \quad \|\Delta\mathbf{R}[:, j]\|_2 \leq n\tilde{\gamma}_m \|\mathbf{A}[:, j]\|_2, \quad \|\Delta\mathbf{R}\|_F \leq n\tilde{\gamma}_m \|\mathbf{A}\|_F$$

$$364 \quad \hat{\mathbf{Q}} = \mathbf{Q} + \Delta\mathbf{Q} = \text{fl}(\hat{\mathbf{P}}_1 \cdots \hat{\mathbf{P}}_n \mathbf{I}), \quad \|\Delta\mathbf{Q}[:, j]\|_2 \leq n\tilde{\gamma}_m, \quad \|\Delta\mathbf{Q}\|_F \leq n^{3/2} \tilde{\gamma}_m.$$

366 In future sections, we show the forward error columnwise bounds for each factor which can be easily
 367 converted to matrix norm bounds. The numerical experiments in [section 5](#) measure backward errors
 368 with $\|\hat{\mathbf{Q}}\hat{\mathbf{R}} - \mathbf{A}\|_F$ and the orthogonality of the \mathbf{Q} factor with $\|\hat{\mathbf{Q}}^\top \hat{\mathbf{Q}} - \mathbf{I}\|_2$. The content of this section
 369 shows the standard rounding error analysis in [\[14\]](#) where some important stages are summarized
 370 in (3.5), (3.7), and (3.13), which we will modify to different mixed precision settings in [section 4](#).
 371 These quantities account for various forward and backward errors formed in computing essential
 372 components of HQR, namely the HH constant and vector, as well as normwise errors of the action
 373 of applying HH transformations. In the next sections, we present blocked variants of HQR that use
 374 [alg. 3](#).

375 **3.2. Block HQR with partitioned columns (BQR).** We refer to the blocked variant
 376 of HQR where the columns are partitioned as BQR. Note that this section relies on the WY
 377 representation described in [\[4\]](#) instead of the storage-efficient version of [\[22\]](#), even though both are
 378 known to be just as numerically stable as HQR.

3.2.1. The WY Representation. A convenient matrix representation that accumulates r HH reflectors is known as the WY representation (see [4, 11]). Lemma 3.5 shows how to update a rank- j update of the identity, $\mathbf{Q}^{(j)}$, with a ~~a~~-an HH 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 HH transformations as shown in Algorithm 4. This algorithm assumes that the HH vectors, \mathbf{V} , and constants, β , have already been computed. Since the \mathbf{Y} factor is exactly \mathbf{V} , we only need to compute the \mathbf{W} factor.

LEMMA 3.5. Suppose $\mathbf{X}^{(j)} = \mathbf{I} - \mathbf{W}^{(j)}\mathbf{Y}^{(j)\top} \in \mathbb{R}^{m \times m}$ 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 $\mathbf{z}^{(j+1)} = \beta \mathbf{X}^{(j)} \mathbf{v}$. Then,

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

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 ~~a~~-an HH transformation that ~~zeros-zeroes~~ out a column below the diagonal, then applying that HH transformation to the bottom right partition. For example, the k^{th} HH transformation finds an $m - k + 1$ sized HH transformation that ~~zeros-zeroes~~ 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 ~~$k+1^{\text{st}}$~~ $(k+1)^{\text{th}}$ column is transformed by the k^{th} HH 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} into blocks 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^{th} block, \mathbf{C}_k , is transformed with HQR (alg. 3), and the WY representation of these r successive HH transformations is constructed as in alg. 4. We write the WY update as

$$(3.14) \quad \mathbf{X}_k = \mathbf{I}_m - \mathbf{W}_k \mathbf{Y}_k^\top = \mathbf{P}_k^{(1)} \cdots \mathbf{P}_k^{(r)}.$$

Thus far, algs. 3 and 4 are rich in Level-2 BLAS operations. Next, $\mathbf{I} - \mathbf{Y}_k \mathbf{W}_k^\top$ is applied to $[\mathbf{C}_2 \cdots \mathbf{C}_N]$ with two Level-3 BLAS operations as shown in line 5 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 [11]), and can reap the benefits from the accelerated block FMA feature of ~~TensorCore~~ TensorCores. Note that BQR does require strictly more FLOPs when compared to HQR, but these additional

408 FLOPs are ~~negligible~~ ~~negligible~~ ~~negligible~~ in standard precision and does not impact the numerical
409 stability. A pseudoalgorithm for BQR is shown in [alg. 5](#) where we assume that $n = Nr$ to make
410 our error analysis in [subsection 3.2.2](#) simple. In practice, an efficient implementation might require
411 r to be a power of two or a product of small prime factors and result a thinner N^{th} block com-
412 pared to the rest. This discrepancy is easily fixed by padding the matrix with zeros, a standard
413 procedure for standard algorithms like the Fast Fourier Transform (FFT). For any variable x in
 $\{\mathbf{X}, \mathbf{W}, \mathbf{Y}, \mathbf{z}, \beta, \mathbf{v}, \mathbf{P}\}$, $x_k^{(j)}$ corresponds to the j^{th} update for the k^{th} block.

Algorithm 5: $\mathbf{Q}, \mathbf{R} \leftarrow \text{blockHQR}(\mathbf{A}, r)$: Perform HH 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  $N = \frac{n}{r}$ 
  // 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    $[\mathbf{C}_{i+1} \cdots \mathbf{C}_N] \leftarrow \mathbf{V}_i (\mathbf{W}_i^\top [\mathbf{C}_{i+1} \cdots \mathbf{C}_N])$     /* 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.
6  $\mathbf{Q} \leftarrow \mathbf{I}$                                                     /*  $\mathbf{I}_m$  if full QR, and  $\mathbf{I}_{m \times n}$  if thin QR. */
7 for  $i = N : -1 : 1$  do
8    $\mathbf{Q}[(i-1)r+1 : m, (i-1)r+1 : n] \leftarrow \mathbf{W}_i (\mathbf{V}_i^\top \mathbf{Q}[(i-1)r+1 : m, (i-1)r+1 : n])$ 
9 return  $\mathbf{Q}, \mathbf{A}$ 
```

414
415 **3.2.2. BQR: Rounding Error Analysis.** We now present the basic structure for the round-
416 ing error analysis for [alg. 5](#), which consist of: 1)HQR, 2)building the W factor, and 3) updating
417 the remaining blocks with the WY representation. We have adapted the analysis from [\[14\]](#) to fit
418 this exact variant, and denote $\hat{\mathbf{Q}}_{BQR}, \hat{\mathbf{R}}_{BQR}$ to be the outputs from [alg. 5](#). First, we analyze the
419 error accumulated from updating $\mathbf{X}_k^{(j-1)}$ to $\mathbf{X}_k^{(j)}$, which applies a rank-1 update via the subtrac-
420 tion of the outer product $\hat{\mathbf{z}}_k^{(j)} \hat{\mathbf{v}}_k^{(j)\top}$. Since $\mathbf{z}_k^{(j)} = \beta_k^{(j)} \mathbf{X}_k^{(j-1)} \mathbf{v}_k^{(j)}$, this update requires a single HH
421 transformation on the right side in the same efficient implementation that is discussed in [\(3.8\)](#),

$$422 \quad (3.15) \quad \hat{\mathbf{X}}_k^{(j)} = \hat{\mathbf{X}}_k^{(j-1)} - \text{fl}(\hat{\beta}_k^{(j-1)} \hat{\mathbf{X}}_k^{(j-1)} \hat{\mathbf{v}}_k^{(j-1)} \hat{\mathbf{v}}_k^{(j)\top}) = \hat{\mathbf{X}}_k^{(j-1)} (\mathbf{P}_k^{(j)} + \Delta \mathbf{P}_k^{(j)}),$$

423 where $\|\Delta \mathbf{P}_k^{(j)}\|_F \leq \tilde{\gamma}_{m-(k-1)r}$. Since $\hat{\mathbf{X}}_k^{(1)} = \mathbf{I} - \hat{\beta}_k^{(1)} \hat{\mathbf{v}}_k^{(1)} \hat{\mathbf{v}}_k^{(1)\top} = \mathbf{P}_k^{(1)} + \Delta \mathbf{P}_k^{(1)}$, we can travel up
424 the recursion relation in [\(3.15\)](#) and use [Lemma 3.2](#) to form

$$425 \quad (3.16) \quad \|\Delta \mathbf{X}_k^{(j)}\|_F := \|\hat{\mathbf{X}}_k^{(j)} - \mathbf{X}_k^{(j)}\|_F \leq j \tilde{\gamma}_{m-(k-1)r}.$$

426 *HQR within each block:* line 3 of [alg. 5](#). We apply [Algorithm 3](#) to the k^{th} block, $\hat{\mathbf{X}}_{k-1} \cdots \hat{\mathbf{X}}_1 \mathbf{C}_k$,
427 which applies r more HH transformations to columns that had been transformed by $(k-1)$ WY
428 transformations in prior iterations. The upper trapezoidal factor that results from applying HQR

429 to $\mathbf{C}_k^{((k-1)r)}$ corresponds to the $(k-1)r+1^{st}$ to kr^{th} columns of $\hat{\mathbf{R}}_{BQR}$, and applying [Lemmas 3.2](#)
 430 and [3.3](#) yields

$$431 \quad \|\hat{\mathbf{R}}_{BQR}[:, j] - \mathbf{R}[:, j]\|_2 \leq r\tilde{\gamma}_m \|\hat{\mathbf{X}}_{k-1} \cdots \hat{\mathbf{X}}_1^\top \mathbf{C}_k[:, j]\|_2, \quad j = (k-1)r+1 : kr.$$

432 *Build WY at each block: line 4 of [alg. 5](#).* We now calculate the rounding errors incurred from
 433 building the WY representation when given a set of HH vectors and constants as shown in [alg. 4](#).
 434 Since the columns of $\hat{\mathbf{Y}}_k$ are simply $\{\hat{\mathbf{v}}_k^{(j)}\}$ built in [alg. 3](#) the errors for forming these are shown in
 435 [\(3.5\)](#) where m should be replaced by $m - (k-1)r$. The HH constants, $\hat{\beta}_k^{(j)}$ are bounded by [\(3.7\)](#)
 436 modified similarly. Thus, $\mathbf{z}_k^{(j)}$ is the only newly computed quantity. Using [\(3.5\)](#), [\(3.7\)](#), and [\(3.16\)](#),
 437 we find

$$438 \quad \|\Delta \mathbf{z}_k^{(j)}\|_2 = \|\Delta \mathbf{X}_k^{(j-1)} \hat{\beta}_k^{(j)} \hat{\mathbf{v}}_k^{(j)}\|_2 \leq \|\Delta \mathbf{X}_k^{(j-1)}\|_2 \|\hat{\beta}_k^{(j)} \hat{\mathbf{v}}_k^{(j)}\|_2 \leq \|\Delta \mathbf{X}_k^{(j-1)}\|_F \|\hat{\beta}_k^{(j)} \hat{\mathbf{v}}_k^{(j)}\|_2 \\
 439 \leq ((1 + (j-1)\tilde{\gamma}_{m-(k-1)r})(1 + \tilde{\gamma}_{m-(k-1)r}) - 1) \|\beta_k^{(j)} \mathbf{v}_k^{(j)}\|_2 \leq j\tilde{\gamma}_{m-(k-1)r} \|\mathbf{z}_k^{(j)}\|_2.$$

441 Componentwise bounds follow immediately, and are summarized in [Lemma 3.6](#).

442 **LEMMA 3.6.** *Consider the construction of the WY representation for the k^{th} partition of matrix*
 443 $\mathbf{A} \in \mathbb{R}^{m \times n}$ *given a set of HH constants and vectors, $\{\beta_k^{(j)}\}_{j=1}^r$ and $\{\mathbf{v}_k^{(j)}\}$ via [alg. 4](#). Then,*

$$444 \quad (3.17) \quad \hat{\mathbf{z}}_k^{(j)} = \mathbf{z}_k^{(j)} + \Delta \mathbf{z}_k^{(j)}, \quad |\Delta \mathbf{z}_k^{(j)}| \leq j\tilde{\gamma}_{m-(k-1)r} |\mathbf{z}_k^{(j)}|, \quad \|\Delta \mathbf{z}_k^{(j)}\|_2 \leq j\tilde{\gamma}_{m-(k-1)r} \|\mathbf{z}_k^{(j)}\|_2.$$

445 Most importantly, this shows that constructing the WY update is just as numerically stable as
 446 applying successive HH transformations (see Section 19.5 of [\[14\]](#)).

447 *Update blocks to the right: line 5 of [alg. 5](#).* We now consider applying $\mathbf{X}_k := \mathbf{I} - \mathbf{W}_k \mathbf{Y}_k^\top$ to
 448 some matrix, \mathbf{B} . In practice, \mathbf{B} is the bottom right submatrix, $[\mathbf{C}_{k+1} \cdots \mathbf{C}_N][(k-1)r+1 : m, :]$.
 449 We can apply [\(3.16\)](#) directly to the columns of \mathbf{B} ,

$$450 \quad (3.18) \quad \|\mathbf{f}(\hat{\mathbf{X}}_k \mathbf{B}[:, j])\|_2 = \|\mathbf{f}(\hat{\mathbf{X}}_k^{(r)} \mathbf{B}[:, j])\|_2 \leq r\tilde{\gamma}_{m-(k-1)r} \|\mathbf{B}[:, j]\|_2.$$

452 A normwise bound for employing a general matrix-matrix multiplication operation is stated in
 453 section 19.5 of [\[14\]](#).

454 *Multiple WY updates: line 8-9 of [alg. 5](#).* All that remains is to consider the application of
 455 successive WY updates to form the QR factorization computed with BQR denoted as \mathbf{Q}_{BQR} and
 456 \mathbf{R}_{BQR} . We can apply [Lemma 3.2](#) directly by setting $\mathbf{X}_k := \mathbf{I} - \mathbf{W}_k \mathbf{Y}_k^\top$ and consider the backward
 457 errors for applying the sequence to a vector, $\mathbf{x} \in \mathbb{R}^m$, as we did for [Lemma 3.3](#). Since $\mathbf{X}_k =$
 458 $\mathbf{P}_{(k-1)r+1} \cdots \mathbf{P}_{kr}$, is simply a sequence of HH transformations, it is orthogonal, i.e. $\|\mathbf{X}_k\|_2 = 1$. We
 459 only need to replace with \mathbf{x} with $\mathbf{A}[:, i]$'s to form the columnwise bounds for \mathbf{R}_{BQR} , and apply the
 460 transpose to $\hat{\mathbf{e}}_i$'s to form the bounds for \mathbf{Q}_{BQR} . Then,

$$461 \quad (3.19) \quad \left\| \prod_{k=1}^N (\mathbf{X}_k + \Delta \mathbf{X}_k) - \prod_{k=1}^N \mathbf{X}_k \right\|_F \leq \left(-1 + \sum_{k=1}^N (1 + r\tilde{\gamma}_{m-(k-1)r}) \right) \leq rN\tilde{\gamma}_m \equiv n\tilde{\gamma}_m,$$

$$462 \quad (3.20) \quad \|\hat{\mathbf{Q}}_{BQR} - \mathbf{Q}\|_F \leq n^{3/2} \tilde{\gamma}_m.$$

464 We can also form the normwise bound for the j'^{th} column of $\hat{\mathbf{Q}}_{BQR}, \hat{\mathbf{R}}_{BQR}$. If we let $k' = \lceil j'/r \rceil^{th}$,
 465 then the j'^{th} column is the result of applying $k'-1$ WY updates and an additional HQR. Applying

466 Lemma 3.2 yields

$$467 \quad (3.21) \quad \|\Delta \mathbf{R}_{BQR}[:, j']\|_2 \leq rk' \tilde{\gamma}_m \|\mathbf{A}[:, j']\|_2, \quad \|\Delta \mathbf{R}_{BQR}\|_F \leq n \tilde{\gamma}_m \|\mathbf{A}\|_F$$

$$468 \quad (3.22) \quad \|\Delta \mathbf{Q}_{BQR}[:, j']\|_2 \leq rk' \tilde{\gamma}_m, \quad \|\Delta \mathbf{Q}_{BQR}\|_F = r \tilde{\gamma}_m \sum_{j=1}^n \lceil j/r \rceil = n^{3/2} \tilde{\gamma}_m.$$

469 and near orthogonality of the \mathbf{Q} factor is still achieved.

470 *BQR and HQR error bound comparison.* BQR under exact arithmetic is equivalent to HQR,
 471 and it is often referred to as the level-3 BLAS version of HQR. Furthermore, the error analysis of
 472 this section shows that BQR is as numerically stable as HQR despite requiring more FLOPs. In
 473 fact, many linear algebra libraries such as LAPACK use a variant of BQR as the QR factorization
 474 algorithm (see `dgeqrf` of [2]). The primary goal of the analysis presented in this section is to
 475 provide the basic skeleton for the standard BQR rounding error analysis to make the generalization
 476 to mixed precision settings in section 4 easier. Readers should refer to [11, 14] for full details.

477 **3.3. Block HQR with partitioned rows : Tall-and-Skinny QR (TSQR).** Some im-
 478 portant problems that require QR factorizations of overdetermined systems include least squares
 479 problems, eigenvalue problems, low rank approximations, as well as other matrix decompositions.
 480 Although Tall-and-Skinny QR (TSQR) broadly refers to block QR factorization methods with row
 481 partitions, we will discuss a specific variant of TSQR which is also known as the AllReduce algo-
 482 rithm [21]. In this paper, the TSQR/AllReduce algorithm refers to the most parallel variant of
 483 the block QR factorization algorithms discussed in [9]. A detailed description and rounding error
 484 analysis of this algorithm can be found in [21], and we present a pseudocode for the algorithm in
 485 alg. 6. Our initial interest in this algorithm came from its parallelizable nature, which is particu-
 486 larly suitable to implementation on GPUs. Additionally, our numerical simulations (discussed in
 487 section 5) show that TSQR can not only increase the speed but also outperform the traditional
 488 HQR factorization in low precisions.

489 **3.3.1. TSQR/AllReduce Algorithm.** Algorithm 6 partitions the rows of a tall-and-skinny
 490 matrix, \mathbf{A} . HQR is performed on each of those blocks and pairs of \mathbf{R} factors are combined to form
 491 the next set of \mathbf{A} matrices to be QR factorized. This process is repeated until only a single \mathbf{R} factor
 492 remains, and the \mathbf{Q} factor is built from all of the HH constants and vectors stored at each level.
 493 The most gains from parallelization can be made in the initial level where the maximum number
 494 of independent HQR factorizations occur. Although more than one configuration of this algorithm
 495 may be available for a given tall-and-skinny matrix, the number of nodes available and the shape of
 496 the matrix eliminate some of those choices. For example, a 1600-by-100 matrix can be partitioned
 497 into 2, 4, 8, or 16 initial row-blocks but may be restricted by a machine with only 4 nodes, and a
 498 1600-by-700 matrix can only be partitioned into 2 initial blocks. Our numerical experiments show
 499 that the choice in the initial partition, which directly relates to the recursion depth of TSQR, has
 500 an impact in the accuracy of the QR factorization.

501 We refer to *level* as the number of recursions in a particular TSQR implementation. An L -
 502 level TSQR algorithm partitions the original matrix into $2^{(L)} 2^{(L)}$ submatrices in the initial or 0^{th}
 503 level of the algorithm, and 2^{L-i} QR factorizations are performed in level i for $i = 1, \dots, L$. The
 504 set of matrices that are QR factorized at each level i are called $\mathbf{A}_j^{(i)}$ for $j = 1, \dots, 2^{L-i}$, where
 505 superscript (i) corresponds to the level and the subscript j indexes the row-blocks within level
 506 i . In the following sections, alg. 6 (`tsqr`) will find a TSQR factorization of a matrix $A \in \mathbb{R}^{m \times n}$
 507 where $m \gg n$. The inline function `qr` refers to alg. 3 and we use alg. 2 as a subroutine of `qr`.

Algorithm 6: $\mathbf{Q}, \mathbf{R} = \text{tsqr}(\mathbf{A}, L)$. Finds a QR factorization of a tall, skinny matrix, \mathbf{A} .

Input: $\mathbf{A} \in \mathbb{R}^{m \times n}$ where $m \gg n$, $L \leq \lfloor \log_2(\frac{m}{n}) \rfloor$, and 2^L is the initial number of blocks.

Output: $\mathbf{Q} \in \mathbb{R}^{m \times n}$, $\mathbf{R} \in \mathbb{R}^{n \times n}$ such that $\mathbf{Q}\mathbf{R} = \mathbf{A}$.

```

1  $h \leftarrow m2^{-L}$   $h \leftarrow 2^{-L}m$  // Number of rows.
/* Split  $\mathbf{A}$  into  $2^L$  blocks. Note that level  $(i)$  has  $2^{L-i}$  blocks. */
2 for  $j = 1 : 2^L$  do
3    $\mathbf{A}_j^{(0)} \leftarrow \mathbf{A}[(j-1)h + 1 : jh, :]$ 
/* Store HH vectors as columns of matrix  $\mathbf{V}_j^{(i)}$ , HH constants as components of
   vector  $\beta_j^{(i)}$ , and set up the next level. */
4 for  $i = 0 : L - 1$  do
/* The inner loop can be parallelized. */
5   for  $j = 1 : 2^{L-i}$  do
6      $\mathbf{V}_{2j-1}^{(i)}, \beta_{2j-1}^{(i)}, \mathbf{R}_{2j-1}^{(i)} \leftarrow \text{qr}(\mathbf{A}_{2j-1}^{(i)})$ 
7      $\mathbf{V}_{2j}^{(i)}, \beta_{2j}^{(i)}, \mathbf{R}_{2j}^{(i)} \leftarrow \text{qr}(\mathbf{A}_{2j}^{(i)})$ 
8      $\mathbf{A}_j^{(i+1)} \leftarrow \begin{bmatrix} \mathbf{R}_{2j-1}^{(i)} \\ \mathbf{R}_{2j}^{(i)} \end{bmatrix}$ 
9  $\mathbf{V}_1^{(L)}, \beta_1^{(L)}, \mathbf{R} \leftarrow \text{qr}(\mathbf{A}_1^{(L)})$  // The final  $\mathbf{R}$  factor is built.
10  $\mathbf{Q}_1^{(L)} \leftarrow \text{hh\_mult}(\mathbf{V}_1^{(L)}, I_{2n \times n})$ 
/* Compute  $\mathbf{Q}^{(i)}$  factors by applying  $\mathbf{V}^{(i)}$  to  $\mathbf{Q}^{(i+1)}$  factors. */
11 for  $i = L - 1 : -1 : 1$  do
12   for  $j = 1 : 2^{L-i}$  do
13      $\mathbf{Q}_j^{(i)} \leftarrow \text{hh\_mult} \left( \mathbf{V}_j^{(i)}, \begin{bmatrix} \tilde{\mathbf{Q}}_{\alpha(j), \phi(j)}^{(i+1)} \\ \mathbf{0} \end{bmatrix} \right)$ 
14  $\mathbf{Q} \leftarrow \mathbf{I}$ ; // Construct the final  $\mathbf{Q}$  factor.
15 for  $j = 1 : 2^L$  do
16    $\mathbf{Q} \leftarrow \begin{bmatrix} \mathbf{Q} \\ \text{hh\_mult} \left( \mathbf{V}_j^{(0)}, \begin{bmatrix} \tilde{\mathbf{Q}}_{\alpha(j), \phi(j)}^{(1)} \\ \mathbf{0} \end{bmatrix} \right) \end{bmatrix}$ 
17 return  $\mathbf{Q}, \mathbf{R}$ 

```

TSQR Notation. We introduce new notation due to the multi-level nature of the TSQR algorithm. In the final task of constructing \mathbf{Q} , $\mathbf{Q}_j^{(i)}$ factors are aggregated from each block at each level. Each $\mathbf{Q}_j^{(i)}$ factor from level i is partitioned such that two corresponding $\mathbf{Q}^{(i-1)}$ factors from level $i-1$ can be applied to them. The partition (approximately) splits $\mathbf{Q}_j^{(i)}$ into two halves, $[\tilde{\mathbf{Q}}_{j,1}^{(i)\top} \tilde{\mathbf{Q}}_{j,2}^{(i)\top}]^\top$. The functions $\alpha(j)$ and $\phi(j)$ are defined such that $\mathbf{Q}_j^{(i)}$ is applied to the correct blocks from the level below: $\tilde{\mathbf{Q}}_{\alpha(j), \phi(j)}^{(i+1)}$. For $j = 1, \dots, 2^{L-i}$ at level i , we need $j = 2(\alpha(j) - 1) + \phi(j)$, where $\alpha(j) = \lceil \frac{j}{2} \rceil$ and $\phi(j) = 2 + j - 2\alpha(j) \in \{1, 2\}$. subsection 3.3.2 shows full linear algebra details for a single-level ($L = 1, 2$ initial blocks) example. The reconstruction of \mathbf{Q} can be implemented more efficiently (see [3]), but the reconstruction method in alg. 6 is presented for a clear, straightforward explanation.

519 **3.3.2. Single-level Example.** In the single-level version of this algorithm, we first bisect \mathbf{A}
 520 into $\mathbf{A}_1^{(0)}$ and $\mathbf{A}_2^{(0)}$ and compute the QR factorization of each of those submatrices. We combine the
 521 resulting upper-triangular matrices (see below) which is QR factorized, and the process is repeated:

$$522 \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}_1^{(0)} \\ \mathbf{A}_2^{(0)} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_1^{(0)} \mathbf{R}_1^{(0)} \\ \mathbf{Q}_2^{(0)} \mathbf{R}_2^{(0)} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_1^{(0)} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2^{(0)} \end{bmatrix} \begin{bmatrix} \mathbf{R}_1^{(0)} \\ \mathbf{R}_2^{(0)} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_1^{(0)} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2^{(0)} \end{bmatrix} \mathbf{A}_1^{(1)} = \begin{bmatrix} \mathbf{Q}_1^{(0)} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2^{(0)} \end{bmatrix} \mathbf{Q}_1^{(1)} \mathbf{R}.$$

523 The \mathbf{R} factor of $\mathbf{A}_1^{(1)}$ is the final \mathbf{R} factor of the QR factorization of the original matrix, \mathbf{A} . However,
 524 the final \mathbf{Q} still needs to be constructed. Bisecting $\mathbf{Q}_1^{(1)}$ into two submatrices, i.e. $\tilde{\mathbf{Q}}_{1,1}^{(1)}$ and $\tilde{\mathbf{Q}}_{1,2}^{(1)}$,
 525 allows us to write and compute the product more compactly,

$$526 \quad \mathbf{Q} := \begin{bmatrix} \mathbf{Q}_1^{(0)} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2^{(0)} \end{bmatrix} \mathbf{Q}_1^{(1)} = \begin{bmatrix} \mathbf{Q}_1^{(0)} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2^{(0)} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{Q}}_{1,1}^{(1)} \\ \tilde{\mathbf{Q}}_{1,2}^{(1)} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_1^{(0)} \tilde{\mathbf{Q}}_{1,1}^{(1)} \\ \mathbf{Q}_2^{(0)} \tilde{\mathbf{Q}}_{1,2}^{(1)} \end{bmatrix}.$$

527 More generally, [alg. 6](#) takes a tall-and-skinny matrix \mathbf{A} and level L and finds a QR factorization
 528 by initially partitioning \mathbf{A} into $2^{(L)}$ row-blocks and includes the building of \mathbf{Q} . For simplicity, we
 529 assume that m is exactly $h2^{(L)}$ so that the initial partition yields $2^{(L)}$ blocks of equal sizes, h -by- n .
 530 Also, note that `hh_mult` refers to the action of applying multiple HH transformations given a set
 531 of HH vectors and constants, which can be performed by iterating line 6 of [alg. 3](#). This step can
 532 be done in a level-3 BLAS operation via a WY update if [alg. 6](#) was modified to store the WY
 533 representation at the QR factorization of each block of each level, $\mathbf{A}_j^{(i)}$.

534 **3.3.3. TSQR: Rounding Error Analysis.** The TSQR algorithm presented in [alg. 6](#) is a
 535 divide-and-conquer strategy for the QR factorization that uses the HQR within the subproblems.
 536 Divide-and-conquer methods can naturally be implemented in parallel and accumulate less rounding
 537 errors. For example, the single-level TSQR decomposition of a tall-and-skinny matrix, \mathbf{A} requires
 538 3 total HQRs of matrices of sizes $\lfloor \log_2(\frac{m}{n}) \rfloor$ -by- n , $\lfloor \log_2(\frac{m}{n}) \rfloor$ -by- n , and $2n$ -by- n . The single-level
 539 TSQR strictly uses more FLOPs, but the dot product subroutines may accumulate smaller rounding
 540 errors (and certainly have smaller upper bounds) since they are performed on shorter vectors, and
 541 lead to a more accurate solution overall. These concepts are elucidated in [\[21\]](#) and we summarize
 542 the main results in [Theorem 3.7](#).

543 **THEOREM 3.7.** *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m \geq n$ have full rank, n , and $\hat{\mathbf{Q}}_{TSQR} \in \mathbb{R}^{m \times n}$ and
 544 $\hat{\mathbf{R}}_{TSQR} \in \mathbb{R}^{n \times n}$ be the thin QR factors of \mathbf{A} obtained via [alg. 6](#) with L levels. Let us further
 545 assume that m is divisible by 2^L and $n\tilde{\gamma}_{m2^{-L}}, n\tilde{\gamma}_{2n} \ll 1$. Then, 2-norm error
 546 bound for the j^{th} column ($j = 1 : n$) of $\hat{\mathbf{R}}_{TSQR}$ and the Frobenius norm error bound for $\hat{\mathbf{Q}}_{TSQR}$
 547 are*

$$548 \quad (3.23) \quad \|\hat{\mathbf{R}}_{TSQR}[:, j] - \mathbf{R}[:, j]\|_2 \leq n(\tilde{\gamma}_{m2^{-L}} + L\tilde{\gamma}_{2n})\|\mathbf{A}[:, j]\|_2,$$

$$549 \quad (3.24) \quad \|\hat{\mathbf{Q}}_{TSQR} - \mathbf{Q}\|_F \leq n^{3/2}(\tilde{\gamma}_{m2^{-L}} + L\tilde{\gamma}_{2n}).$$

551 Note that the $n\tilde{\gamma}_{m2^{-L}}$ and $n\tilde{\gamma}_{2n}$ terms correspond to errors from applying HQR to the
 552 blocks in the initial partition and to the blocks in levels 1 through L respectively. We can easily
 553 replace these with analogous mixed precision terms and keep the analysis accurate. Both level-2
 554 and level-3 BLAS implementations will be considered in [section 4](#).

TSQR and HQR error bound comparison. We compare the error bounds for HQR and TSQR. Consider the bounds for $\|\hat{\mathbf{Q}} - \mathbf{Q}\|_F$ in [Theorems 3.4](#) and [3.7](#). TSQR has a lower worst-case error bound than HQR when integers $m, n > 0$, and $L \geq 0$ satisfy

$$1 \gg n^{3/2} \gamma^{(m)} m \gg n^{3/2} (\gamma^{(\frac{m}{2^L})} 2^{-L} m + L \gamma^{(2n)} 2n).$$

Let us consider as an example the case when $\frac{m}{2^L} = 2n$. Then, the HQR bound is $2^L/(L+1)$ larger than the bound for TSQR with L levels. For example, in single precision, a HQR of a 2^{15} -by- 2^6 matrix results in an upper bound relative backward error ($\|\mathbf{A} - \hat{\mathbf{Q}}\hat{\mathbf{R}}\|_F/\|\mathbf{A}\|_F$) of ≈ 1.002 , but a TSQR with $L = 8$ is bounded by $\approx 3.516\text{e-}02$. This case exemplifies a situation in which stability is not guaranteed in HQR, but the method is stable when using TSQR, even in the worst-case. Note that this likely is not the typical situation in practice. Now consider some 2^{20} -by- 2^{12} matrix and QR factorizations performed with double precision. The error bound for HQR is $1.686\text{e-}7$, whereas the error bound for TSQR with 12 levels is $5.351\text{e-}10$. In general, we can conjecture that values of L that can make $m2^{-L}2^{-L}m$ and $2Ln$ much smaller than m , should produce a TSQR that outperforms HQR in worst-case scenarios, at least in uniform precision settings. However, the range of matrix sizes that TSQR can accommodate decreases as L grows larger. Figure 1 shows the matrix sizes HQR, 2-level TSQR, and 4-level TSQR can accommodate as well as their respective error bounds.

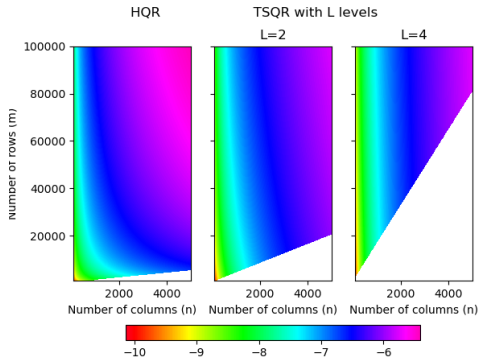


FIG. 1. Non-white space indicates allowable matrix sizes for each scheme, and color map represents error bounds for $\|\Delta\mathbf{Q}\|_F$ for uniform precision error analysis when using double precision arithmetic in log₁₀ scale.

4. Mixed precision error analysis. In this section, we consider three different mixed precision settings for the QR factorization, all of which take in a matrix \mathbf{A} stored in low precision and return \mathbf{Q}, \mathbf{R} both represented in low precision. First, we consider a trivial mixed precision setting where HQR, BQR, and TSQR are computed in high precision after casting up the input matrix at the beginning, and casting down the resulting high precision factors to low precision. Then in [subsection 4.1](#), we modify BQR and TSQR to utilize level-3 BLAS operations and TensorCore bFMAs for the matrix product subroutines. Finally, we impose [MP Setting 2.3](#) in [subsection 4.2](#) to see how a mixed precision inner product impacts HQR, BQR, and TSQR when applied in level-2 BLAS operations.

Backward error of casting down vectors. First, consider casting down a vector $\mathbf{x} \in \mathbb{F}_h^{(m)}$. The componentwise forward error is,

$$\text{castdown}_l(\mathbf{x}) = \mathbf{x} + \Delta\mathbf{x}, \quad |\Delta\mathbf{x}| < u^{(l)}|\mathbf{x}|.$$

We use this to represent the backward error of a casting down a vector with a linear transformation, $\mathbf{I}^{(l)} := \mathbf{I} + \mathbf{E} \in \mathbb{R}^{m \times m}$, a diagonal perturbation of the identity. We write,

$$(4.1) \quad \mathbf{x}^{(l)} := \text{castdown}(\mathbf{x}^{(h)}) = \mathbf{I}^{(l)}\mathbf{x}^{(h)} = (\mathbf{I} + \mathbf{E})\mathbf{x}^{(h)} = \mathbf{x}^{(h)} + \Delta\mathbf{x},$$

where $|\Delta \mathbf{x}| \leq u^{(l)} |\mathbf{x}^{(h)}|$ and $\|\Delta \mathbf{x}\|_2 \leq u^{(l)} \|\mathbf{x}^{(h)}\|_2$. Thus, $\mathbf{E} = \Delta \mathbf{x} \mathbf{x}^\top / \|\mathbf{x}\|_2^2$ and we can use the same argument as in (3.11) to form a backward matrix norm bound,

$$(4.2) \quad \|\mathbf{E}\|_F \leq u^{(l)}.$$

Casting down after HQR in high precision. Let us consider the trivial case of carrying out HQR in high precision and casting down at the very end. This is useful for the analysis of mixed precision block algorithms as will be shown in subsection 4.1. If the two floating point types \mathbb{F}_l and \mathbb{F}_h satisfy $\mathbb{F}_l \subseteq \mathbb{F}_h$ and the matrix to be factorized is stored with low precision numbers, $\mathbf{A} \in \mathbb{F}_l^{m \times n}$, then casting up adds no rounding errors. Therefore, we can directly apply the analysis that culminated in Theorem 3.4, and we only consider the columnwise forward error in the \mathbf{Q} factor. Then, the j^{th} column of $\hat{\mathbf{Q}}_{HQR} = \mathbf{Q} + \Delta \mathbf{Q}_{HQR}$ is bounded normwise via $\|\Delta \mathbf{Q}_{HQR}[:, j]\|_2 \leq n \tilde{\gamma}_m^h$, and incurs an extra rounding error when $\hat{\mathbf{Q}}_{HQR} \in \mathbb{F}_h^{m \times n}$ is cast down to $\mathbb{F}_l^{m \times n}$. Using this in Lemma 3.2 to analyze the forward norm error for the j^{th} column of the \mathbf{Q} factor computed with alg. 3 yields

$$(4.3) \quad \|(\text{castdown}(\hat{\mathbf{Q}}_{HQR}) - \mathbf{Q})[:, j]\|_2 = \|(\mathbf{I}^{(l)} \hat{\mathbf{P}}_1 \cdots \hat{\mathbf{P}}_n - \mathbf{P}_1 \cdots \mathbf{P}_n) \hat{\mathbf{e}}_j\|_2 \leq u^{(l)} + n \tilde{\gamma}_m^{(h)} + n u^{(l)} \tilde{\gamma}_m^{(h)}.$$

The final castdown operation increases the upper bound by $u^{(l)}$ and the size of \mathbf{A} has no impact on this extra rounding error. Applying this trivial mixed precision setting to BQR and TSQR would simply increase the error bound by approximately $u^{(l)}$ all the while taking an even longer time than the high precision implementation due the extra cast down and cast up operations. Therefore, we do not analyze the rounding error analysis of this mixed precision variant of BQR and TSQR. However, we will use this mixed precision HQR as a subroutine of the mixed precision BQR and TSQR in the following section.

4.1. Round down at block-level: level-3 BLAS mixed precision setting. The mixed precision setting in this section is designed to meet the below requirements.

1. Modify Algorithms 5 and 6 to maximize level-3 BLAS operations and use TensorCore bFMAs.
 2. Apply (4.3) to all instances of HQR to the error analyses for BQR and TSQR in section 3.
 3. Cast down quantities at every block/level and the insertion of low precision errors $u^{(l)}$ should be somewhat correlated to the number of blocks and levels.
 4. Both input and output of the various QR factorization algorithms are given in the low precision.
- TensorCore's bFMA computes

$$(4.4) \quad \hat{\mathbf{D}} = \text{fl}_{TC}(\mathbf{C} + \mathbf{A}\mathbf{B}), \quad \mathbf{C}, \mathbf{D} \in \mathbb{F}_{\text{fp16}}^{4 \times 4} \text{ or } \mathbb{F}_{\text{fp32}}^{4 \times 4}, \text{ and } \mathbf{A}, \mathbf{B} \in \mathbb{F}_{\text{fp16}}^{4 \times 4},$$

and employs *full* precision products and fp32 summation accumulate. Here, the *full* precision multiplication is exact as explained in section 2. In [7][5], the authors investigate all four possible matrix-matrix multiplication routines in **TensorCoreTensorCores**, which depend on whether \mathbf{C} and \mathbf{D} are computed in fp16 or fp32. They also note that matrices larger than 4-by-4 can still be computed using this block FMA by accumulating matrix sums with $\mathbf{C} \in \mathbb{F}_{\text{fp32}}^{4 \times 4}$. Suppose that we aim to compute a fp16 matrix product of two fp16 matrices, $\mathbf{X} \in \mathbb{F}_{(fp16)}^{m \times p}$, $\mathbf{Y} \in \mathbb{F}_{(fp16)}^{p \times n}$, and $\mathbf{Z} = \mathbf{X}\mathbf{Y} \in \mathbb{F}_{\text{fp16}}^{m \times n}$. We pad \mathbf{X}, \mathbf{Y} with zeros so that all matrix dimensions are multiples of 4 and the matrix product can be computed with the TensorCore block FMA. Let $\mathbf{Q}_{[i,j]} := \mathbf{Q}[4(i-1)+1 : 4i, 4(j-1)+1 : 4j]$ refer to the $(i, j)^{\text{th}}$ 4-by-4 block for any $\mathbf{Q} \in \{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$. Then, we compute

635 $\mathbf{Z}_{[i,j]}$ via

$$636 \quad \mathbf{Z}_{[i,j]} = \sum_{k=1}^{\lceil p/4 \rceil} \mathbf{X}_{[i,k]} \mathbf{Y}_{[k,j]},$$

637 where we use (4.4) by initializing with $\mathbf{A}^{(1)} := \mathbf{X}_{[i,1]}$, $\mathbf{B}^{(1)} := \mathbf{Y}_{[1,j]}$, and $\mathbf{C}^{(1)} := \mathbf{0}_{4 \times 4}$ and setting
 638 $\mathbf{A}^{(k)} := \mathbf{X}_{[i,k]}$, $\mathbf{B}^{(k)} := \mathbf{Y}_{[k,j]}$, and $\mathbf{C}^{(k)} := \mathbf{D}^{(k-1)}$ for $k = 2 : \lceil p/4 \rceil$. By setting $\mathbf{C}^{(k)}, \mathbf{D}^{(k)} \in \mathbb{F}_{\text{fp32}}^{4 \times 4}$
 639 for $k > 1$ and only casting down at the end via $\mathbf{Z}_{[i,j]} = \text{fp16}(\mathbf{D}^{(\lceil p/4 \rceil)})$, we maximize our use of
 640 fp32 arithmetic. This computes the most accurate mixed precision matrix product routine possible
 641 using TensorCore bFMAs whose inputs and output are required to be stored in fp16. For example,
 642 take $p = 8$. Then the $[i, j]^{th}$ 4-by-4 block of the product is computed via,

$$643 \quad \mathbf{D}^{(1)} = \text{fl}_{TC}(\mathbf{X}_{[i,1]} \mathbf{Y}_{[1,j]}), \quad \mathbf{D}^{(2)} = \text{fl}_{TC}(\mathbf{X}_{[i,2]} \mathbf{Y}_{[2,j]} + \mathbf{D}^{(1)}) \in \mathbb{F}_{\text{fp32}}^{4 \times 4}$$

$$644 \quad \mathbf{Z}_{[i,j]} = \text{castdown}(\mathbf{D}^{(2)}) \in \mathbb{F}_{\text{fp16}}^{4 \times 4}.$$

646 Adapting the rounding error analysis in [?] [5] into this specific mixed precision matrix product
 647 setting yields the componentwise forward bound

$$648 \quad (4.5) \quad |\mathbf{Z} - \text{fl}(\mathbf{Z})| \leq \left(u^{(\text{fp16})} + \gamma_{\underline{p/4p}}^{(\text{fp32})} + u^{(\text{fp16})} \gamma_{\underline{p/4p}}^{(\text{fp32})} \right) |\mathbf{X}| |\mathbf{Y}|.$$

649 We denote BQR and TSQR computed via TensorCore bFMA's with mpBQR3 and mpTSQR3,
 650 where the 3 represents the BLAS level-3 nature of this mixed precision setting.

651 **4.1.1. BQR round down at block level: mpBQR3.** Consider the input matrix, $\mathbf{A} \in \mathbb{F}_l^{m \times n}$,
 652 partitioned into N blocks of r columns, $\mathbf{A} = [\mathbf{C}_1 \cdots \mathbf{C}_N]$ as in subsection 3.2. Algorithm 7 shows a
 653 mixed precision variant of BQR that maximizes the use of bFMAs but uses high precision arithmetic
 654 for level-1 and 2 BLAS operations which are only a $\mathcal{O}(1/N)$ fraction of the total number of FLOPs.
 655 Each block is casted up to compute a high precision HQR and to form the WY representation.
 656 The WY representation is then casted down to low precision since the bFMAs require low precision
 657 inputs for matrix products, and the \mathbf{R} factor from the high precision HQR can be casted down to
 658 return a low precision \mathbf{R} factor at the very end. Since the cast down operations for the \mathbf{R} factor
 659 and the WY representations occur at every block, we can expect columnwise error bound for alg. 7
 660 to increase by approximately $Nu^{(l)}$ from the error bound for alg. 5.

661 Since $\hat{\mathbf{W}}_k, \hat{\mathbf{Y}}_k$'s are computed with alg. 4 in high precision then cast down, the new low precision
 662 WY update is $\hat{\mathbf{X}}_k^{(l)} = \mathbf{I} - \mathbf{I}^{(l)} \hat{\mathbf{W}}_k \mathbf{I}^{(l)} \hat{\mathbf{Y}}_k^{(\top)}$. Consider applying $\hat{\mathbf{X}}_k^{(l)}$ to some matrix stored in low
 663 precision, \mathbf{B} using the TensorCore bFMAs. We analyze a single column $\mathbf{b}_j := \mathbf{B}[:, j] \in \mathbb{F}_l^{m-(k-1)r}$
 664 even though this operation is done on \mathbf{B} as a whole. Let $\mathbf{I}^{(l)} \hat{\mathbf{W}}_k = (\mathbf{I} + \mathbf{E}_W) \hat{\mathbf{W}}_k$ and $\mathbf{I}^{(l)} \hat{\mathbf{Y}}_k =$
 665 $(\mathbf{I} + \mathbf{E}_Y) \hat{\mathbf{Y}}_k$, where $\mathbf{E}_W, \mathbf{E}_Y$ are diagonal and bounded componentwise by $u^{(l)}$. Then, the Frobenius
 666 norm error of forming $\hat{\mathbf{X}}_k^{(l)}$ is,

$$667 \quad \|\hat{\mathbf{X}}_k^{(l)} - \mathbf{X}_k\|_F = \| -(\mathbf{I} + \mathbf{E}_W + \mathbf{E}_Y + \mathbf{E}_W \mathbf{E}_Y) \hat{\mathbf{W}}_k \hat{\mathbf{Y}}_k^\top + \mathbf{W}_k \mathbf{Y}_k^\top \|_F,$$

$$668 \quad \leq \left((1 + \gamma_2^{(l)} + (u^{(l)})^2) r \tilde{\gamma}_{m-(k-1)r}^{(h)} + \gamma_2^{(l)} + (u^{(l)})^2 \right) \|\mathbf{X}_k\|_F$$

$$669 \quad \leq \tilde{\gamma}_2^{(l)} + r \tilde{\gamma}_{m-(k-1)r}^{(h)} + r \tilde{\gamma}_2^{(l)} \tilde{\gamma}_{m-(k-1)r}^{(h)}.$$

Algorithm 7: $\hat{\mathbf{Q}}_{mpBQR3}, \hat{\mathbf{R}}_{mpBQR3} \leftarrow \text{mpBQR3}(\mathbf{A}, r)$: Perform a mixed precision variant of BQR of low precision \mathbf{A} with column partitions of size r . $\hat{\mathbf{Q}}_{mpBQR3}, \hat{\mathbf{R}}_{mpBQR3}$, are returned in low precision. Operations in lines 7 and 10 require low precision inputs.

Input: \mathbf{A}, r . **Output:** $\hat{\mathbf{Q}}_{mpBQR3}, \hat{\mathbf{R}}_{mpBQR3}$
 1 $N = \frac{n}{r}$ /* Let $\mathbf{A} = [\mathbf{C}_1 \cdots \mathbf{C}_N]$. */
 2 **for** $k = 1 : N - 1$ **do**
 3 $\mathbf{V}_k, \beta_k, \mathbf{C}_k \leftarrow \text{hhQR}(\text{castup}(\mathbf{C}_k))$ /* Algorithm 3 in high precision. */
 4 $\mathbf{C}_k \leftarrow \text{castdown}(\mathbf{C}_k)$ /* Builds \mathbf{R} factor in low precision. */
 5 $\mathbf{W}_k \leftarrow \text{buildWY}(\mathbf{V}_k, \beta_k)$ /* Algorithm 4 in high precision */
 6 $[\mathbf{V}_k, \mathbf{W}_k] \leftarrow \text{castdown}([\mathbf{V}_k, \mathbf{W}_k])$
 7 $[\mathbf{C}_{k+1} \cdots \mathbf{C}_N] \leftarrow \mathbf{V}_k (\mathbf{W}_k^\top [\mathbf{C}_{k+1} \cdots \mathbf{C}_N])$ /* returned in low precision */
 8 $\mathbf{Q} \leftarrow \mathbf{I}$ /* Build \mathbf{Q} : \mathbf{I}_m if full QR, and $\mathbf{I}_{m \times n}$ if thin QR. */
 9 **for** $k = N : -1 : 1$ **do**
 10 // All updates are returned in low precision.
 10 $\mathbf{Q}[(k-1)r+1 : m, (k-1)r+1 : n] \leftarrow \mathbf{W}_k (\mathbf{V}_k^\top \mathbf{Q}[(k-1)r+1 : m, (k-1)r+1 : n])$
 11 **return** \mathbf{Q}, \mathbf{A}

671 Now, we consider the backward error of applying $\hat{\mathbf{X}}_k^{(l)}$ to \mathbf{b}_j with the bFMA matrix product error
 672 bound from (4.5). The multiplication by $(\mathbf{I}^{(l)} \hat{\mathbf{Y}}_k)^\top$ yields backward error bounded by

$$673 \quad \text{fl}_{TC}((\mathbf{I}^{(l)} \hat{\mathbf{Y}}_k)^\top \mathbf{b}_j) = (\hat{\mathbf{Y}}_k + \Delta_{TC} \hat{\mathbf{Y}}_k) \mathbf{b}_j, \quad |\Delta_{TC} \hat{\mathbf{Y}}_k| \leq u^{(l)} + \gamma_{\frac{m-(k-1)}{4}}^{(h)} + u^{(l)} \gamma_{\frac{m-(k-1)}{4}}^{(h)} |\hat{\mathbf{Y}}_k| |\mathbf{b}_j|,$$

674 and the subsequent multiplication by $(\mathbf{I}^{(l)} \hat{\mathbf{W}}_k)$ and subtraction from \mathbf{b}_j result in,

$$675 \quad \text{fl}_{TC}(\hat{\mathbf{X}}_k^{(l)} \mathbf{b}_j) = (\hat{\mathbf{X}}_k^{(l)} + \Delta^{(l)} \mathbf{X}_k) \mathbf{b}_j,$$

$$676 \quad |\Delta^{(l)} \mathbf{X}_k| \leq \left(\gamma_2^{(l)} + \gamma_{1+\frac{m-(k-2)r}{4}}^{(h)} + \gamma_2^{(l)} \gamma_{1+\frac{m-(k-2)r}{4}}^{(h)} \right) \left(|\mathbf{b}_j| + |\mathbf{I}^{(l)} \hat{\mathbf{W}}_k| |\mathbf{I}^{(l)} \hat{\mathbf{Y}}_k|^\top |\mathbf{b}_j| \right).$$

677

678 Converting to a normwise error bound using the same logic from (3.9) and (3.10), we result in

$$679 \quad (4.6) \quad \|\text{fl}_{TC}(\hat{\mathbf{X}}_k^{(l)} \mathbf{b}_j) - \mathbf{X}_k \mathbf{b}_j\|_2 \leq (\tilde{\gamma}_2^{(l)} + r \tilde{\gamma}_{m-(k-1)r}^{(h)} + r \gamma_2^{(l)} \tilde{\gamma}_{m-(k-1)r}^{(h)}) \|\mathbf{b}_j\|_2,$$

680 since the rounding errors from the bFMAs are small in comparison to the errors from casting down
 681 the WY representation built in high precision. The corresponding matrix error bound is

$$682 \quad (4.7) \quad \|\text{fl}_{TC}(\hat{\mathbf{X}}_k^{(l)}) - \mathbf{X}_k\|_F \leq \tilde{\gamma}_2^{(l)} + r \tilde{\gamma}_{m-(k-1)r}^{(h)} + r \gamma_2^{(l)} \tilde{\gamma}_{m-(k-1)r}^{(h)}.$$

683 We can finally compute the forward errors from implementing alg. 7. Consider the j^{th} column of
 684 the \mathbf{Q} factor, which we denote with $\mathbf{q}_j := \hat{\mathbf{Q}}_{mpBQR3}[:, j]$, and let $k = \lfloor j/r \rfloor$. Invoking Lemma 3.2
 685 with error bounds for $\text{fl}_{TC}(\hat{\mathbf{X}}_k^{(l)})$'s in (4.7) results in columnwise error,

$$686 \quad (4.8) \quad \|\Delta \mathbf{q}_j\|_2 \leq -1 + \prod_{k'=1}^k (1 + \tilde{\gamma}_2^{(l)}) (1 + r \tilde{\gamma}_{m-(k'-1)r}^{(h)})$$

$$687 \quad (4.9) \quad \leq k \tilde{\gamma}_2^{(l)} + k r \tilde{\gamma}_m^{(h)} + k^2 r \tilde{\gamma}_2^{(l)} \tilde{\gamma}_m^{(h)},$$

where $\Delta \mathbf{q}_j = (\text{fl}_{TC}(\hat{\mathbf{X}}_1^{(l)}) \cdots \text{fl}_{TC}(\hat{\mathbf{X}}_k^{(l)}) - \mathbf{X}_1 \cdots \mathbf{X}_k) \hat{\mathbf{e}}_j$. Summing over the columns to find a matrix norm error bound yields

$$(4.10) \quad \|\hat{\mathbf{Q}}_{mpBQR} - \mathbf{Q}\|_F \leq n^{1/2} \left(\tilde{\gamma}_N^{(l)} + n \tilde{\gamma}_m^{(h)} \right),$$

where the summation of the third term in (4.9) is swept under the tilde notation in $n^{1/2} \tilde{\gamma}_N^{(l)}$. This bound shows that [alg. 7](#) only adds $n^{1/2} \tilde{\gamma}_N^{(l)}$ order errors to the bounds in (3.22). Using that $u^{(l)} = M_{l,h} u^{(h)}$, this increase corresponds to a multiplicative factor shown below,

$$(4.11) \quad n^{1/2} \tilde{\gamma}_N^{(l)} + n^{(3/2)} \tilde{\gamma}_m^{(h)} \approx \left(1 + \frac{M_{l,h}}{rm} \right) n^{(3/2)} \tilde{\gamma}_m^{(h)}.$$

~~Therefore, In practice, we expect hardware specifications to restrict r between $\mathcal{O}(10)$ and $\mathcal{O}(100)$ and fix finite options for u_l , u_h , and therefore $M_{l,h}$. This helps narrow our analysis. In general, the loss in accuracy due to mixed precision computing is relatively small when the disparity in precision ($M_{l,h}$) is small in comparison to the block size, mr . However, as we can assume r to be fixed at $\mathcal{O}(100)$ at the largest, the loss of accuracy attributed to mixed precision computing grows small as m grows large, $N = n/r$ decreases which then reduces the portion of indicating that mpBQR3 performed using level-3 BLAS operations can yield accuracy comparable to its high, uniform precision variant while still benefiting from speed-ups. Next, note that $r = 1$ and increases the size of high precision HQR being performed at each block. Whether this loss in accuracy in the worst-case scenario is worth the $r = n$ both revert to HQR with only level-2 BLAS operations, and other values of r determine the proportion of level-2 and level-3 BLAS operations. The optimal blocksize only needs to be searched within the permissible ranges for r and $M_{l,h}$. Overall, the actual trade-off between speed-ups and accuracy from using mixed precision hardware for the QR factorization is an open question that can be tackled more comprehensively in future research. Our analysis shows that the block-size-worst-case bounds depend on block width r , the dimension of the input matrix m, n , and hardware specificities will be contributing factors as well as hardware specificities.~~

4.1.2. TSQR round down at block level: mpTSQR3. Unlike BQR which is rich in level-3 BLAS operations, the variant of TSQR in [alg. 6](#) uses none. Therefore, we modify [alg. 6](#) by replacing all instances of `hh_mult` with level-3 BLAS operations. We omit presenting the exact algorithm for mixed precision variant of TSQR in this paper, but consider computing the HQR of each block in high precision and build and store the WY representation of the HH transformations in low precision as we did in lines (3-6) of [alg. 7](#). The low precision WY representation is then applied with TensorCore bFMAs when building the \mathbf{Q} factor (lines 11-16 of [alg. 6](#)).

Rounding Error analysis. The analysis in [21] shows that each column of \mathbf{Q} is transformed by n HH transformations of length $2n$ from levels $L : -1 : 1$, and another set of n HH transformations of length $\underline{m2^{-L}2^{-L}m}$ at level 0. Let us represent the WY representation at the j^{th} block of level i and its bFMA counterpart as $\mathbf{X}_j^{(i)}$ and $\text{fl}_{TC}(\hat{\mathbf{X}}_j^{(i)})$. Then, we can use (4.7) to form backward error

$$(4.12) \quad \|\text{fl}_{TC}(\hat{\mathbf{X}}_j^{(i)}) - \mathbf{X}_j^{(i)}\|_F \leq \tilde{\gamma}_2^{(l)} + n \tilde{\gamma}_{m'}^{(h)} + n \tilde{\gamma}_2^{(l)} \tilde{\gamma}_{m'}^{(h)}, \quad m' = \begin{cases} 2^{-L}m, & i = 0 \\ 2n, & i = 1 : L \end{cases}.$$

We can now modify the analysis in [21] by replacing ~~$n \tilde{\gamma}_{m2^{-L}2^{-L}m}$~~ and $n \tilde{\gamma}_{2n}$ with

$$(1 + \tilde{\gamma}_2^{(l)})(1 + n \tilde{\gamma}_{\underline{m2^{-L}2^{-L}m}}^{(h)}) - 1, \quad \text{and} \quad (1 + \tilde{\gamma}_2^{(l)})(1 + n \tilde{\gamma}_{2n}^{(h)}) - 1,$$

and apply [Lemma 3.2](#). Then, the factors formed by `mpTSQR3` are denoted by $\hat{\mathbf{R}}_{mpTSQR3}$, $\hat{\mathbf{Q}}_{mpTSQR3}$ and the error bounds for the j^{th} column of the triangular factor and the orthogonal factor are

$$\|(\hat{\mathbf{R}}_{mpTSQR3} - \mathbf{R})[:, j]\|_2 \leq \tilde{\gamma}_{L+1}^{(l)} + n \left(L\tilde{\gamma}_{2n}^{(h)} + \tilde{\gamma}_{\underline{m2-L} \underline{2-Lm}}^{(h)} \right) \|\mathbf{A}[:, j]\|_2,$$

$$\|\hat{\mathbf{Q}}_{mpTSQR3} - \mathbf{Q}\|_F \leq n^{1/2} \tilde{\gamma}_{L+1}^{(l)} + n^{3/2} \left(L\tilde{\gamma}_{2n}^{(h)} + \tilde{\gamma}_{\underline{m2-L} \underline{2-Lm}}^{(h)} \right).$$

Converting the low precision rounding errors as a fraction of the TSQR error bound in [\(3.24\)](#) to quantify the impact of modifying [alg. 6](#) to utilize bFMAs yields

$$n^{1/2} \tilde{\gamma}_{L+1}^{(l)} + n^{3/2} \left(L\tilde{\gamma}_{2n}^{(h)} + \tilde{\gamma}_{\underline{m2-L} \underline{2-Lm}}^{(h)} \right) = \left(1 + \frac{M_{l,h}(L+1)}{\underline{n(2nL+m2-L)}} \frac{M_{l,h}(L+1)}{\underline{n(2-Lm+2nL)}} \right) n^{3/2} \left(L\tilde{\gamma}_{2n}^{(h)} + \tilde{\gamma}_{\underline{m2-L} \underline{2-Lm}}^{(h)} \right)$$

Like in [\(4.11\)](#), the disparity in the two precisions, $M_{l,h}$ is compared against the original matrix size m, n and the block size specifications derived from L . Let us consider the shallowest, middle, and the deepest levels of TSQR that are possible given some matrix in $\mathbb{R}^{m \times n}$. All three cases in [Table 4](#) show that `mpTSQR3` on sufficiently large matrices may yield errors closer to the high precision implementation, and the optimal choice for L depends on m, n .

| Number of levels, L | 1 | $\frac{1}{2} \log_2(m/n)$ | $-1 + \log_2(m/n)$ |
|---|-----------------|--|--------------------|
| $\frac{(L+1)}{\underline{n(2nL+m2-L)}} \frac{(L+1)}{\underline{n(2-Lm+2nL)}}$ | $1/(n^2 + m/4)$ | $1/\left(2n^2 + \frac{m^{1/2}n^{3/2}}{\log_2(m/n)}\right)$ | $1/(2n^2)$ |

TABLE 4

Error bounds for $\|\Delta \mathbf{Q}_{mpTSQR3}\|_F$ for varying L 's.

Finally, the error bounds for the \mathbf{Q} matrix formed from implementing all of the algorithms discussed in [sections 3 and 4](#) are summarized below in [Table 5](#).

| Algorithm | Uniform | Mixed BLAS-2 (§4.2) | Mixed BLAS-3 (§4.1) |
|--|---|---|---|
| HQR (§3.1) | $\underline{n^{3/2} \tilde{\gamma}_m}$ | $\underline{n^{1/2}(\tilde{\gamma}_{10n}^{(l)} + n\tilde{\gamma}_m^{(h)})}$ | $\underline{n/a}$ |
| BQR | $\underline{n^{3/2} \tilde{\gamma}_m}$ | $\underline{n^{1/2} \tilde{\gamma}_{10N}^{(l)} + n^{3/2} \tilde{\gamma}_m^{(h)}}$ | $\underline{n^{1/2}(\tilde{\gamma}_N^{(l)} + n\tilde{\gamma}_m^{(h)})}$ |
| TSQR | $\underline{n^{3/2}(\tilde{\gamma}_{m2-L} + L\tilde{\gamma}_{2n})}$ | $\underline{n^{1/2} \tilde{\gamma}_{10(L+1)}^{(l)} + n^{3/2} (L\tilde{\gamma}_{2n}^{(h)} + \tilde{\gamma}_{m2-L}^{(h)})}$ | $\underline{n^{1/2} \tilde{\gamma}_{L+1}^{(l)} + n^{3/2} (L\tilde{\gamma}_{2n}^{(h)} + \tilde{\gamma}_{2-Lm}^{(h)})}$ |

TABLE 5

Error bounds for $\|\Delta \mathbf{Q}\|_F$ for all algorithms in [sections 3 and 4](#).

4.2. Round down at inner product: level-2 BLAS mixed precision setting. While the previous section discussed blocked variants of HQR that can be easily adapted for the mixed precision setting specific to TensorCore bFMA's, we want to provide a more general mixed precision environment in this section. Recall that HQR, BQR, and TSQR all rely on HH transformations in one way or another, and implementations of HH transformations are expressed by [\(3.8\)](#). This implementation capitalizes on the rank-1 update structure of HH transformations where the predominant share of FLOPs is spent on an inner product, and computing the HH vector and constant

also rely heavily on inner products. Therefore, nearly all of the computational tasks for algs. 3, 5 and 6 are attributed to the inner product, which is important in other linear algebra tools such as projections, matrix-vector, and matrix-matrix multiply. Consequently, we return to MP Setting 2.3, where every inner product is cast down to the lower precision as shown in (2.10). We denote HQR, BQR, and TSQR computed with MP Setting 2.3 with mpHQR2, mpBQR2, and mpTSQR2, where the 2 represents the mixed precision procedure computed at a level-2 BLAS operation.

4.2.1. HQR round down at inner product: mpHQR2. Consider forming an HH transformation that zeros out $\mathbf{x} \in \mathbb{R}^m$ below the i^{th} element. We need to compute σ , β , $\hat{\mathbf{v}}_1$, and \mathbf{v} as defined in subsection 3.1,

$$(4.14) \quad \text{fl}(\sigma) = \text{fl}(-\text{sign}(\mathbf{x}[1])\|\mathbf{x}\|_2) = \sigma + \Delta\sigma, \quad |\Delta\sigma| \leq \left(\gamma_2^{(l)} + \gamma_m^{(h)} + \gamma_2^{(l)}\gamma_m^{(h)}\right)|\sigma|,$$

$$(4.15) \quad \text{fl}(\mathbf{v}'[1]) = \mathbf{v}'[1] + \Delta\mathbf{v}'[1] = (1 + \delta^{(l)})(\mathbf{x}[1] - \sigma - \Delta\sigma), \quad |\Delta\mathbf{v}'[1]| \leq (\gamma_3^{(l)} + \tilde{\gamma}_m^{(h)})|\mathbf{v}'[1]|$$

$$(4.16) \quad \text{fl}(\beta) = \beta + \Delta\beta = (1 + \delta^{(l)})(-\mathbf{v}'[1]/\hat{\sigma}), \quad |\Delta\beta| \leq (\gamma_5^{(l)} + \tilde{\gamma}_m^{(h)})|\beta|,$$

$$(4.17) \quad \text{fl}(\mathbf{v}[j]) = \mathbf{v}[j] + \Delta\mathbf{v}[j] \text{ where } |\Delta\mathbf{v}[j]| \leq (\gamma_7^{(l)} + \tilde{\gamma}_m^{(h)})|\mathbf{v}_j|, j = 2 : m - i + 1.$$

These bounds on $\Delta\sigma$, $\Delta\mathbf{v}'[1]$, $\Delta\beta$, and $\Delta\mathbf{v}[j]$ are computed by using the rules from Lemma 2.4 on the analysis shown in subsection 3.1. Using these, we can formulate the mixed precision version of (3.9) where $\hat{\mathbf{y}} = \text{fl}(\mathbf{P}_\mathbf{v}\mathbf{x}) \in \mathbb{R}^m$ is implemented via (3.8). Note that the inner product $\hat{\mathbf{v}}^\top \mathbf{x}$ via MP Setting 2.3, and all other operations are done in the lower precision. Then, the transformed vector is bounded by

$$(4.18) \quad \hat{\mathbf{y}} = \mathbf{y} + \Delta\mathbf{y}, \quad \|\Delta\mathbf{y}\|_2 \leq (\gamma_{25}^{(l)} + \tilde{\gamma}_m^{(h)})\|\mathbf{y}\|_2.$$

Thus, a backward error can be formed using $\Delta\mathbf{P}_\mathbf{v} = \Delta\mathbf{y}\mathbf{x}^\top / \|\mathbf{x}\|_2^2$, and by introducing the $\tilde{\gamma}$ notation for the low precision term,

$$(4.19) \quad \hat{\mathbf{y}} = (\mathbf{P}_\mathbf{v} + \Delta\mathbf{P}_\mathbf{v})\mathbf{x}, \quad \|\Delta\mathbf{P}_\mathbf{v}\|_F \leq (\gamma_{25}^{(l)} + \tilde{\gamma}_m^{(h)}).$$

Now, we form the error bounds for applying n HH transformations to \mathbf{x} using Lemma 3.2,

$$(4.20) \quad \hat{\mathbf{z}} = \text{fl}(\mathbf{P}_1 \cdots \mathbf{P}_n \mathbf{x}) = \mathbf{Q}(\mathbf{x} + \Delta\mathbf{x}) = (\mathbf{Q} + \Delta\mathbf{Q})\mathbf{x},$$

$$(4.21) \quad \|\Delta\mathbf{y}\|_2 \leq (\tilde{\gamma}_{10n}^{(l)} + n\tilde{\gamma}_m^{(h)})\|\mathbf{x}\|_2, \quad \|\Delta\mathbf{Q}\|_F \leq (\tilde{\gamma}_{10n}^{(l)} + n\tilde{\gamma}_m^{(h)}).$$

~~Note that we use the $\tilde{\gamma}^{(l)}$ notation, where the small integer e is now required to be $\mathcal{O}(25)$.~~ The analogous mixed precision QR factorization error bounds are shown in Theorem 4.1.

THEOREM 4.1. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m \geq n$ have full rank, n . Let $\hat{\mathbf{Q}}_{\text{mpHQR2}} \in \mathbb{R}^{m \times n}$ and $\hat{\mathbf{R}} \in \mathbb{R}_{\text{mpHQR2}}^{n \times n}$ be the thin QR factors of \mathbf{A} obtained via alg. 3 with mixed precision FLOPs where inner products are computed in precision h then cast down. All other operations are carried out in precision l . Then,

$$(4.22)$$

$$\|\Delta\mathbf{R}_{\text{mpHQR2}}[:, j]\|_2 \leq (\tilde{\gamma}_{10n}^{(l)} + n\tilde{\gamma}_m^{(h)})\|\mathbf{A}[:, j]\|_2, \quad \|\Delta\mathbf{R}_{\text{mpHQR2}}\|_F \leq (\tilde{\gamma}_{10n}^{(l)} + n\tilde{\gamma}_m^{(h)})\|\mathbf{A}\|_F$$

$$(4.23)$$

$$\|\Delta\mathbf{Q}[:, j]_{\text{mpHQR2}}\|_2 \leq (\tilde{\gamma}_{10n}^{(l)} + n\tilde{\gamma}_m^{(h)}), \quad \|\Delta\mathbf{Q}_{\text{mpHQR2}}\|_F \leq n^{1/2}(\tilde{\gamma}_{10n}^{(l)} + n\tilde{\gamma}_m^{(h)}).$$

Unsurprisingly, the inner product mixed precision setting yields higher error bounds (see Theorem 3.4) as it uses more low precision arithmetic than the settings described in subsection 4.1. Before commenting on the significance of Theorem 4.1, we show that the same bounds hold for the BQR variant. In the next sections we analyze using mpHQR2 instead of HQR within algs. 5 and 6.

4.2.2. BQR round down at inner product: mpBQR2. Now, we analyze alg. 5 implemented with MP Setting 2.3. At the k^{th} block, we first apply the mixed precision HQR summarized in Theorem 4.1. Next, we construct the WY representation, where we can now use (4.18) and (4.19) and Lemma 3.2 to form

$$(4.24) \quad \|\hat{\mathbf{X}}_k^{(l)} - \mathbf{X}_k\|_F = \|(\hat{\mathbf{P}}_k^{(1)} \cdots \hat{\mathbf{P}}_k^{(r)}) - (\mathbf{P}_k^{(1)} \cdots \mathbf{P}_k^{(r)})\|_F \leq \tilde{\gamma}_{\underline{r}10r}^{(l)} + r\tilde{\gamma}_m^{(h)}.$$

Then, the 2-norm bound for the j^{th} column of the \mathbf{R} factor and the Frobenius norm bound for the orthogonal factor resulting from mpBQR2 are

$$(4.25) \quad \|\hat{\mathbf{R}}_{mpBQR2}[:, j]\|_2 = \|\hat{\mathbf{X}}_1 \cdots \hat{\mathbf{X}}_N \mathbf{A}[:, j]\|_2 \leq \left(N\tilde{\gamma}_{\underline{r}10r}^{(l)} + n\tilde{\gamma}_m^{(h)} \right) \|\mathbf{A}[:, j]\|_2,$$

$$(4.26) \quad \|\hat{\mathbf{Q}}_{mpBQR2}\|_F \leq n^{1/2} \left(N\tilde{\gamma}_{\underline{r}10r}^{(l)} + n\tilde{\gamma}_m^{(h)} \right) \approx \left(1 + \frac{M_{l,h}}{m} \frac{10M_{l,h}}{m} \right) n^{3/2} \tilde{\gamma}_m^{(h)}.$$

Note that this error bound is of the same order as the error bound for mpHQR2, shown in (4.23). The corresponding error bound for mpBQR3 of subsection 4.1.1 yielded low precision errors r times smaller than that from using MP Setting 2.3 inner products, an unsurprising result as intermediate results are cast down more often in mpBQR2. Furthermore, the $\tilde{\gamma}^{(l)}$ in this section requires $c = \mathcal{O}(25)$, whereas the same notation in subsection 4.1.1 assumes c to be a small positive integer. Therefore, the ~~Therefore, guarantees of numerical stability of mpBQR2 is guaranteed at are limited to smaller matrix sizes than the numerical stability when compared to those of mpBQR3 and BQR in high precision.~~ While it is technically possible that the low precision errors introduced from utilizing MP Setting 2.3 do not dominate the errors incurred in mpBQR2 and mpHQR2 when $m \gg M_{l,h}$ and can result in accuracy comparable to that of mpBQR3 and high precision BQR, our numerical results in section 5 show that mpHQR2 is already unstable at $m \approx M_{l,h}$.

4.2.3. TSQR round down at inner product: mpTSQR2. Finally, we consider using MP Setting 2.3 in alg. 6. This corresponds to replacing every instance of $n\tilde{\gamma}_{m'}$ for $m' \in \{2n, m2^{-L}\}$ with $\tilde{\gamma}_m^{(l)} + n\tilde{\gamma}_{m'}^{(h)}\tilde{\gamma}_{10n}^{(l)} + n\tilde{\gamma}_{m'}^{(h)}$. We first consider the norm errors for the j^{th} column of the \mathbf{Q} factor computed by this mixed precision variant of alg. 6,

$$(4.27) \quad \|\hat{\mathbf{Q}}_{mpTSQR2}[:, j] - \mathbf{Q}[:, j]\|_2 \leq (L+1)\tilde{\gamma}_{\underline{n}10n}^{(l)} + n(\tilde{\gamma}_{\underline{m}2^{-L}2^{-L}m}^{(h)} + L\tilde{\gamma}_{2n}^{(h)}).$$

Then, the matrix norm error bound is

$$(4.28) \quad \|\hat{\mathbf{Q}}_{mpTSQR2} - \mathbf{Q}\|_F \leq n^{1/2}(L+1)\tilde{\gamma}_{\underline{n}10n}^{(l)} + n^{3/2}(\tilde{\gamma}_{\underline{m}2^{-L}2^{-L}m}^{(h)} + L\tilde{\gamma}_{2n}^{(h)})$$

$$(4.29) \quad \approx \left(1 + \frac{M_{l,h}L}{m2^{-L} + 2Ln} \frac{10M_{l,h}L}{2^{-L}m + 2Ln} \right) n^{3/2}(\tilde{\gamma}_{\underline{m}2^{-L}2^{-L}m}^{(h)} + L\tilde{\gamma}_{2n}^{(h)}),$$

and contributes larger low precision rounding errors than in (4.13). If the mpTSQR2 error bound were to outperform that of mpHQR2, we now need integers $m, n > 0$, and $L \geq 0$ that satisfy

$$1 \gg n^{1/2} \left(\tilde{\gamma}_{\underline{n}10\underline{n}}^{(l)} + n\tilde{\gamma}_m^{(h)} \right) \gg n^{1/2} \left((L+1)\tilde{\gamma}_{\underline{n}10\underline{n}}^{(l)} + n(\tilde{\gamma}_{\underline{m}2^{-L}\underline{2}^{-L}\underline{m}}^{(h)} + L\tilde{\gamma}_{2n}^{(h)}) \right).$$

In contrast to the analysis for uniform precision settings, large L values do not necessarily reduce the error bounds of TSQR. While large L can imply $\underline{m} \gg \underline{m}2^{-L} + 2Ln$, it does not always lead to $\underline{d} \gg \underline{d}_1 + L\underline{d}_2 \underline{m} \gg 2^{-L}\underline{m} + 2Ln$, it is not always the case. Although the theoretical error bounds do not give a clear indication of the worst-case performances of HQR and TSQR in mixed-precision settings, TSQR outperformed HQR on ill-conditioned matrices within our numerical simulations. These experiments are discussed in detail in the next section.

5. Numerical Experiments. We conducted several numerical experiments to confirm the validity of the error bounds formed in section 4 by varying size for all algorithms, block sizes in mpBQR3, and comparing mpHQR2 against mpTSQR2 with varying condition numbers. We used Julia, a programming language which allows fp16 storage and `castup` and `castdown` operations between types in fp16, fp32, fp64, but no built-in fp16 arithmetic. Therefore, we relied on using `alg. 1` for $f \in \text{OP} \cup \{\text{dot_product}\}$ to simulate MP Setting 2.3 and TensorCore bFMAs.

In sections 3 and 4, we gave the forward error bounds for \mathbf{R} and \mathbf{Q} separately. Since our numerical experiments instead measure a backward error, $\|\hat{\mathbf{Q}}\hat{\mathbf{R}} - \mathbf{A}\|_F$, and an orthogonal error, $\|\hat{\mathbf{Q}}^\top \hat{\mathbf{Q}} - \mathbf{I}\|_2$, we show how to convert general forward errors into those computed quantities. Given $\|(\hat{\mathbf{R}} - \mathbf{R})[:, j]\|_2 \leq \epsilon_R \|\mathbf{A}[:, j]\|_2$ and $\|\hat{\mathbf{Q}} - \mathbf{Q}\|_F \leq \epsilon_Q$,

$$(5.1) \quad \|(\hat{\mathbf{Q}}\hat{\mathbf{R}} - \mathbf{A})[:, j]\|_2 \leq (\epsilon_R + \epsilon_Q + \epsilon_R \epsilon_Q) \|\mathbf{A}[:, j]\|_2, \quad j = 1 : n, \quad \text{see [14]},$$

$$(5.2) \quad \|\hat{\mathbf{Q}}\hat{\mathbf{R}} - \mathbf{A}\|_F \leq n^{1/2}(\epsilon_R + \epsilon_Q + \epsilon_R \epsilon_Q) \|\mathbf{A}\|_F,$$

$$(5.3) \quad \|\hat{\mathbf{Q}}^\top \hat{\mathbf{Q}} - \mathbf{I}\|_2 \leq \|\hat{\mathbf{Q}}^\top \hat{\mathbf{Q}} - \mathbf{I}\|_F \simeq 2\epsilon_Q, \quad \text{see [21]}.$$

First, we tested algs. 3 and 5 to 7, mpHQR2, mpBQR2, and mpTSQR2 for varying matrix sizes. We increased the number of rows m from 1000 to 13949, while keeping $\underline{n} = \underline{m}/4$, $\underline{r} = \underline{n}/4$, $\underline{n} = 250$, $\underline{r} = 63$, and $L = 2$ and the test matrices were sampled from the standard normal distribution. On the left plot of Figure 2, we see three clusters which each correspond to: top, MP Setting 2.3; middle, TensorCore bFMAs; and bottom, uniform precision implementations in fp32. The high precision and bFMA implementations scale similarly to each other when increasing the matrix size, whereas the MP Setting 2.3 variants grow unstable more quickly. In addition, while HQR, BQR, and TSQR perform similarly in high precision and when using bFMAs, mpTSQR2 is less accurate by a quarter to a half order of magnitude in comparison to mpBQR2 and mpHQR2. The specifications for $m, n, L, M_{l,h}$ for this experiment derive the upper bound for $\|\Delta \mathbf{Q}_{mpTSQR2}\|_F$, (4.29), to be larger than that of $\|\Delta \mathbf{Q}_{mpHQR2}\|_F$, (4.23). However, a more careful comparison of mpHQR2 and mpTSQR2 show that there exists a regime where mpTSQR2 can outperform mpHQR2.

Next, we varied the block sizes for performing fp32 BQR and mpBQR3 on 2048-by-256 sized matrices, which were chosen to yield error bounds below 1 for both algorithms. The right plot of Figure 2 shows the error bounds and the computed value for the backward error for the two algorithms where the block size r varies from 2 to 256. The test matrices were generated following example from [?] [5] by setting $\mathbf{A} = \text{castdown}(\mathbf{Q}_1 \mathbf{D} \mathbf{Q}_2)$ where $\mathbf{Q}_1 \in \mathbb{F}_h^{m \times n}$, $\mathbf{Q}_2 \in \mathbb{F}_h^{n \times n}$ are orthogonal and $\mathbf{D} = \text{Diagonal}(\{\log_{10}(0), \dots, \log_{10}(-3)\}) \in \mathbb{F}_h^{n \times n}$. The high precision implementation yields backward error close to $u^{(fp32)}$ and mpBQR3 yields errors near $u^{(fp16)}$ that follows the

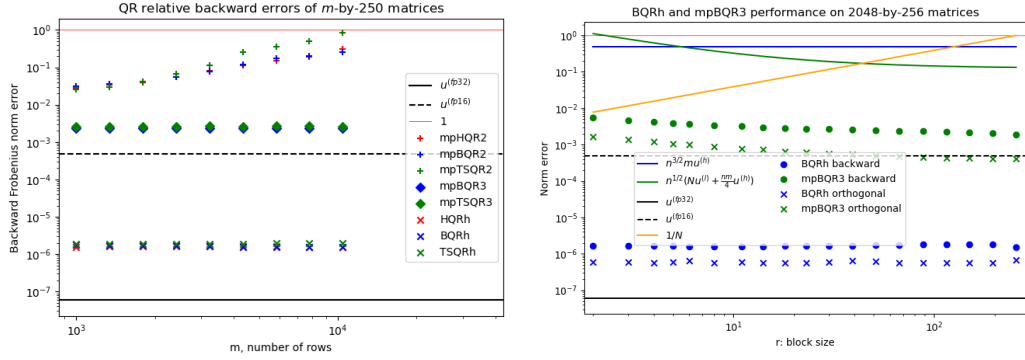


FIG. 2. Left plot: Backward errors of HH QR factorization algorithms in sections 3 and 4 with varying matrix sizes. Right plot: Norm errors of fp32 BQR and mpBQR3 for 2048-by-256 matrices for varying block sizes.

downward trend suggested by (4.11). As block sizes increase, mpBQR3 grows more accurate. This trend correlates to $1/N$, the approximate fraction of FLOPs in mpBQR3 performed in high precision, marked in orange. However, the rightmost data for mpBQR3 (corresponds to $r = n$), is still between 3 and 4 orders of magnitude less accurate than its high precision variant. Further studies that directly test speed-ups from bFMAs against the accuracy of mpBQR3 are needed to fully understand the potential uses for mixed precision QR algorithms.

Lastly, we compared a mixed precision variant of a communication-avoiding QR algorithm (mpTSQR2) against a mixed precision variant of the HH QR algorithm (mpHQR2) on a set of fixed-size matrices with varying condition numbers. Note that an empirical comparison of the two algorithms implemented in fp64 arithmetic were reported in [21], and we omit the comparison against mpBQR2 since it performs very similarly to mpHQR2. Following example from [21], we used m -by- n random matrices, $\mathbf{A}_\alpha = \mathbf{Q}'(\alpha\mathbf{E} + \mathbf{I})/\|\mathbf{Q}'(\alpha\mathbf{E} + \mathbf{I})\|_F$, where $\mathbf{Q}' \in \mathbb{R}^{m \times n}$ is orthogonal and $\mathbf{E} \in \mathbb{R}^{n \times n}$ is the matrix of 1's. We constructed \mathbf{Q}' by computing the default QR factorization of matrix $\mathbf{\Omega} \in \mathbb{F}_{fp64}^{4000 \times 100}$ in Julia, which performs BQR with $r = 36$ entirely in fp64 arithmetic, and elements of the random matrix $\mathbf{\Omega}$ were sampled from the uniform distribution over $[0, 1]$. By construction, \mathbf{A}_α has 2-norm condition number $n\alpha + 1$. By varying α from $1\text{e-}4$ to 1, we varied the condition number from 1.1 to 101, and we generated 10 samples for each value of α . Even though the condition number is not a part of our rounding error analysis, we use it in this experiment in an attempt to reach the “worst-case” scenario described by our deterministic error bounds in section 4. The relative backward error, $\|\hat{\mathbf{Q}}\hat{\mathbf{R}} - \mathbf{A}\|_F/\|\mathbf{A}\|_F$, was computed by casting up $\hat{\mathbf{Q}}$, $\hat{\mathbf{R}}$, and \mathbf{A} to fp64 to compute the Frobenius norms. Plugging in $m = 4000$, $n = 100$, $u^{(l)} = u^{(fp16)}$, $u^{(h)} = u^{(fp32)}$, and $c = 1$ (for $\tilde{\gamma}$) into the error bounds for mpHQR2 combined with (5.2) and (5.3) are approximately 1.179 and 1.146. These error bounds are *relative* and these worst-case bounds do not guarantee errors below 100%. The TSQR bounds for the same parameters for $L = 1 : 5$ are even larger, which indicates that stability is not guaranteed. The leftmost plot of Figure 3 shows the backward errors of mpHQR2 increasing as the theoretical condition numbers of the generated random matrices increase, and these errors correspond to the error data on the vertical axis, $L = 0$, of the middle plot. In addition to the errors from mpHQR2, Figure 3 shows the errors from mpTSQR2s of levels varying from $L = 1$ to $L = 5$, where each line represents the errors of HQR and variants of TSQR calculated from the same random test matrix. Figure 3 reveals two different trends for the errors as we deepen the complexity of the QR algorithm from mpHQR2 to mpTSQR2 with $L = 5$. One trend occurs for matrices with smaller condition numbers, where mpHQR2 is stable, but mpTSQR2

with higher levels yield larger errors. Another trend occurs for matrices with higher condition numbers, where single-level and 2-level mpTSQR2 yield smaller errors than mpHQR2. In these cases, errors from mpTSQR2 with 3 or more levels are similar to or worse than their 2-level variants, but generally do not exceed those of mpHQR2 most of the times. These results suggests that TSQR can outperform HQR ~~even in mixed precision settings, and particularly when HQR is unstable due to larger condition numbers~~ suffers from accumulated rounding errors.

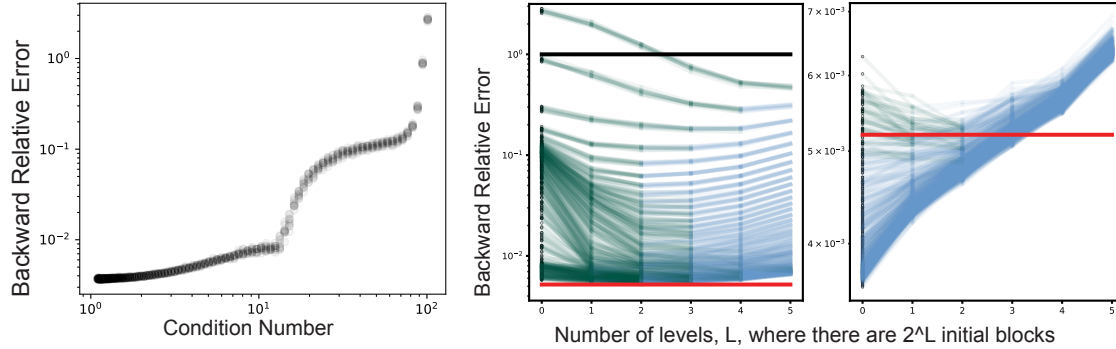


FIG. 3. All plots show the backward relative error for 4000-by-100 sized test matrices. Left: mpHQR2 on condition numbers ranging from 1.1 to 101; Middle: mpTSQR2 on condition numbers ranging from 5.3 to 101; Right: mpTSQR2 on condition numbers ranging from 1.1 to 5.3.

In conclusion, most of the experiments display the trends that error bounds in sections 3 and 4 suggest, and bFMA variants perform in between the high precision and MP Setting 2.3 variants as expected. Also, a special case is shown that demonstrate mpTSQR2 can outperform mpHQR2 in accuracy despite having higher error bounds. All of the experiments showed that the actual errors were many orders of magnitude lower than the error bounds even when ill-conditioned, but this discrepancy varied for different mixed precision settings. For example, backward and forward errors of mpBQR3 were *only* 2-3 orders of magnitude below the error bounds, whereas the fp32 implementation of BQR yielded errors up to 6 orders of magnitude below the error bounds. Although further studies with larger problem sizes and timings would be beneficial in developing an mpBQR3 with the optimal block size, r , our experiments confirm the intuition built from the error analysis in section 4.

6. Conclusion. The development of GPUs that optimize low precision floating point arithmetic have accelerated the interest in half and mixed precision algorithms that naturally reduces the bandwidth and storage needs. Loss in precision, stability, and representable range offset for those advantages, but these shortcomings may have little to no impact in some applications. It may even be possible to navigate around those drawbacks with algorithmic design.

We present the algorithm and standard error analysis of HQR and its blocked variants (BQR and TSQR), modify the algorithms to support two mixed precision settings, and performed error ~~analysis that accurately analyses that~~ bound the mixed precision versions. One mixed precision setting is that of NVIDIA’s TensorCore bFMAs, and the other is an ad hoc setting that mimics the bFMAs at the level of inner products. These two are presented to offer mixed precision arithmetic at both level-2 and 3 BLAS operations and can be applied to other linear algebra tools as well. The new error bounds more accurately describe how rounding errors are accumulated in mixed precision settings. For a given problem, available hardware, and some error tolerance, these bounds can be

used to first narrow down which QR factorization algorithms are feasible. Then, the speed-ups from the hardware specifications can be considered next to choose the most appropriate settings within the algorithms (i.e. block size r in BQR or number of levels, L , in TSQR). We found that TSQR can outperform HQR under [MP Setting 2.3](#) for ill-conditioned, extremely overdetermined cases even when the error bounds imply the opposite. While an optimistic interpretation of this result would be that algorithms like TSQR are more robust against lower precision arithmetic, further research is needed to explore other divide-and-conquer methods that can harness parallel capabilities. Meanwhile, we should rely on the error bounds formed in [section 4](#).

Acknowledgements. The authors are grateful to the anonymous reviewers for their constructive criticisms, which improved the clarity of the presentation.

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. SORENSEN, *LAPACK Users' Guide (Third Ed.)*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999; also available online from <http://www.netlib.org>.
- [3] G. BALLARD, J. W. DEMMEL, L. GRIGORI, M. JACQUELIN, H. DIEP NGUYEN, AND E. SOLOMONIK, *Reconstructing Householder vectors from tall-skinny QR*, vol. 85, 05 2014, pp. 1159–1170, <https://doi.org/10.1109/IPDPS.2014.120>.
- [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] P. BLANCHARD, N. J. HIGHAM, F. LOPEZ, T. MARY, AND S. PRANESH, *Mixed Precision Block Fused Multiply-Add: Error Analysis and Application to GPU Tensor Cores*, SIAM Journal on Scientific Computing, 42 (2020), pp. C124–C141, <https://doi.org/10.1137/19M1289546>, <https://epubs.siam.org/doi/10.1137/19M1289546>.
- [6] M. COURBARIAUX, Y. BENGIO, AND J.-P. DAVID, *Training deep neural networks with low precision multiplications*, arXiv preprint, arXiv:1412.7024, (2014).
- [7] M. COURBARIAUX, J.-P. DAVID, AND Y. BENGIO, *Low precision storage for deep learning*, arXiv preprint arXiv:1412.7024, (2014).
- [8] 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>.
- [9] 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>.
- [10] 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>.
- [11] G. H. GOLUB AND C. F. VAN LOAN, *Matrix computations*, JHU press, 4 ed., 2013.
- [12] 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.
- [13] 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>.
- [14] N. J. HIGHAM, *Accuracy and Stability of Numerical Methods*, 2002, <https://doi.org/10.2307/2669725>.
- [15] 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>.

- siam.org/doi/10.1137/18M1226312.
- [16] 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>.
- [17] A. S. HOUSEHOLDER, *Unitary triangularization of a nonsymmetric matrix*, Journal of the ACM (JACM), 5 (1958), pp. 339–342.
- [18] 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>.
- [19] S. MARKIDIS, S. W. D. CHIEN, E. LAURE, I. B. PENG, AND J. S. VETTER, *NVIDIA tensor core programmability, performance & precision*, Proceedings - 2018 IEEE 32nd International Parallel and Distributed Processing Symposium Workshops, IPDPSW 2018, (2018), pp. 522–531, <https://doi.org/10.1109/IPDPSW.2018.00091>, <https://arxiv.org/abs/1803.04014>.
- [20] P. MICIKEVICIUS, S. NARANG, J. ALBEN, G. DIAMOS, E. ELSEEN, D. GARCIA, B. GINSBURG, M. HOUSTON, O. KUCHARIEV, G. VENKATESH, AND H. WU, *Mixed precision training*, in International Conference on Learning Representations, 2018, <https://openreview.net/forum?id=r1gs9JgRZ>.
- [21] D. MORI, Y. YAMAMOTO, AND S. L. ZHANG, *Backward error analysis of the AllReduce algorithm for householder QR decomposition*, Japan Journal of Industrial and Applied Mathematics, 29 (2012), pp. 111–130, <https://doi.org/10.1007/s13160-011-0053-x>.
- [22] 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>.
- [23] 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>.
- [24] 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>.