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.

Reviewer #1 *I find the reorganization and reframing of the paper to be an improvement. I believe that the work has merit and deserves to be published. However, in my opinion further revision is required before this can be considered. Some of the main issues I find are the use of the term "backward error" and the treatment of the "disparity in precisions". Namely, there must be some constraint on the disparity in precisions in order for the assumption of "full precision" to apply, but this is not stated clearly.*

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"?*
   Not sure what to say.
5. *line 25: "4x" versus what?*
   Not sure what to say.
6. *line 36: "computed up to 16x than that of fp64"— do you mean "16x faster"?*
   Done.
7. *lines 36-37: "The existing rounding error analyses are built within ..." Existing rounding error analyses of what?*
   Done (added "of linear algebra routines").
8. *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. Is this what the reviewer wants?
9. *lines 43-44: "... its block variants that partition the columns ... and those that partition the columns" Is there some typo here?*
   Fixed.
10. *line 58: What does "sufficient in performance" mean?*
    This is now clarified.
11. *line 137: The "(j,k)" in the max should not be a subscript.*
    Done.
12. *line 144: State here what algorithms 5 and 6 are, since they have not been introduced yet.*
    Done.
13. *line 145: Define what is meant by "ad hoc".*
    Not sure what to say.
14. *line 155: "is in" → "in"*

Done.

15. *line 157: There is no "Assumption 2.3". I assume you meant "MP Setting 2.3".*
    Done.
16. *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.
17. *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 contrains 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 have had troubles defining "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 Settin 2.3 to reflect this and to get rid of ambiguities that follow from writing "full precision".
18. *line 288: extra plus sign in "2(m+2)+".*
    Done.
19. *line 288-289: "swept that minor difference between under" Some typo/extra word here.*
    Done.
20. *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)
21. *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.**
22. *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 Q. The backward error may be bounded in terms of the loss of orthogonality and the residual.*
    I disagree that $\|\mathbf{A} - \hat{\mathbf{Q}}\hat{\mathbf{R}}\|$ should be understood as the optimal error from having the "best" QR factorization. How do we respond to this?
23. *line 399: "negligble" spelling*
    Fixed.
24. *line 416: Where is* $\Delta X_k^{(j)}$ *defined?*
    This is now explicitly defined.
25. *line 441: missing period after displayed equation*
    Added.
26. *line 494:* $2^{(l)}$ *should be* 2^L*, correct?*
    Yes! Thanks for catching that.
27. *line 536 (and elsewhere):* $m2^{-L}$ *would look nicer as* $2^{-L}m$
    Done.
28. *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.
29. *lines 552-553: This case ..." Ok, but do we really expect this in practice? Say something*

about this.
I'm considering adding "Note that this likely is not the typical situation in practice." Any suggestions?

30. *line 556: '2Ln" typo*
We got $2Ln$ from the approximation $L\gamma_{2n} = L\frac{2nu}{1-2nu} \approx 2Lnu$.
31. *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?*
If we address this, we should go back to defining MP Setting 2.3 with full precision and making a comment about how that affects $M_{l,h}$. Should we do this? Or we could say some blanket statement earlier on about how the disparity we assumed to be large and therefore $M_{l,h}$ is sufficiently large?
32. *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 has been explicitly stated in description and in algorithm.) Is this response ok?
33. *lines 708-710: Are there some factors dropped from the displayed equations here?*
We believe these are the correct order of coefficients for the $\tilde{\gamma}$ terms. Right?
34. *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.
35. *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. Should we reword this sentence in the manuscript?
36. *line 870: "mpTSQR2 can outperform mpHQR2" - "outperform" in terms of what?*
We have added that this is with respect to accuracy.
37. *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 *I would like to start by thanking the authors for clearly having taken the review process very seriously. I have found the revised article to be much stronger and clearer. The authors have taken into account all the major points of my first review, and in some cases have completely exceeded my expectations. I'm notably referring to the fact that the revised article now analyzes both BLAS-2 and BLAS-3 (blocked) QR algorithms, thereby covering the use of block FMA units such as GPU tensor cores. This is a valuable addition that makes the paper very useful for practitioners interested in exploiting this type of hardware to accelerate QR factorization. After this second round of review, I have several other new comments, but I consider them to be minor and*

hopefully easy to address. I have no doubt that this paper will go on to be a reference on the error analysis of mixed precision QR.

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.*
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, thatis, 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.*
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* $mu^{(h)}$*.*
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.*
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.*
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).*
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 seems 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 significand 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.

8. I suggest adding a Table at the end of section 4 or beginning of section 5 sum- marizing 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 ar- ticles to refer to, and as reference when comparing with the numerical experiments of section 5.
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.
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) ?

Stylistic suggestions, typos, and other very minor things:

1. *Everywhere: a HH → an HH.*
   Done.
2. *L44: columns → 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 → 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 → zeroes out.*
8. *L288: remove "between".*
   Done.
9. *L294: (3.7) show → (3.7) to show ?*
10. *L295: calculating of $\|x\|$ → remove "of".*
11. *L316, L317, L319: I suggest rewriting $v^\top xv$ either as $vv^\top x$ or $(v^\top x)v$.*
12. *(3.13): $\Delta y \to \Delta x$.*
13. *L375: constants, β → missing space.*
14. *L385: $k + 1^{st} \to (k+1)^{th}$, idem at L420.*
15. *L397, L617: TensorCore → TensorCores ?*
16. *L399: negligble → negligible.*
17. *L761: "the inner product mixed precision setting yields higher error bounds": specify than what (add a reference to Thm. 3.4 ?)*
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?*

19. *L798: what are d1 and d2 here?*
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).*
21. *L885: analysis that accurately bound* $\rightarrow$ *change to "analyses" or "bounds".*
22. *Reference [5] has now appeared: cf. [1].*
    The referenc 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.