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 In my opinion, the manuscript is still in need of major revision. Please note that line numbers here correspond to those in the diff file.

1. *I still have some issue with the lack of assumptions on $u^{(h)}$ vs. $u^{(l)}$ in the "ad hoc" setting. I still don't see how you can get around needing $u^{(h)} \lesssim (u^{(l)})^2$, as otherwise you can not guarantee "exact" products as you do in your proof in lines 167-172 and thus you cannot ignore errors from computing the products as you do in (2.8).*
   In your response you said "However, we are trying to leave the analysis general so that other combinations of floating point numbers can be considered as well", but there is nothing general about the proof in lines 167-172. Also, in lines 165-166, you say here that the exact products in MP Setting 2.3 are done by using fp16 as low precision and fp32 as high precision. So is that the case (in which case the value of the disparity in precisions is clear), or are you trying to be more general (in which case I still think you need h to be at least double the precision of l)? I do not think that the assumption $u^{(h)} \lesssim (u^{(l)})^2$ will limit you technically, since this is actually the case in TensoreCore bFMAs.
   I feel like I'm rambling. I just wanted point out that the extra requirement this reviewer wants us to add is not a fix-all, but maybe it's nicer to not point this out. Feel free to cut any of the response below.
   We have added this requirement, but this requirement alone does not guarantee exact products. For example, consider Google's bfloat16, which has the same number of exponent bits as fp32, but still is a 16-bit floating point number. The unit round-off for bfloat16 is approximately `3.91e-3`, and its square is certainly larger than the unit round-off for fp32, which is approximately `5.96e-8`. However, there exist products of two bfloat16 numbers that are outside the range of numbers represented by fp32, and therefore fails the requirement in MP Setting 2.3 that products are computed exactly. This now points to a separate issue of underflow and overflow, which is the reason why we can't apply MP Setting 2.3 directly to the pair bfloat16/fp32 even when it meets the requirement $u^{(h)} \lesssim (u^{(l)})^2$.
   If we restrict the analysis to the inner products of bfloat16 vectors that avoid under/overflow when being accumulated in fp32, then we can apply our analysis. Of course, the fp16/fp32 pair can still run into issues of under/overflow at the final cast down step, and this possibility is now mentioned in the manuscript. Additionally, the final cast down step is precisely the main difference between MP Setting 2.3 and the TC bFMAs: By enforcing the cast down step after every inner product, this ad hoc mixed precision setting belongs to level-2 BLAS operations, whereas the TC bFMAs allow for inner products to return the high precision floats and allow for mixed precision algorithms work at level-3 BLAS operations. Note that the standard rounding error analysis

ignores underflow and overflow, and we follow that example in our analysis.

2. *Another major point is that looking at the plots in Figure 2, I am curious how uniform precision methods in fp16 would do compared to the "mp*2" variants. In other words, is there any benefit at all to doing "mp*2" variants, or would it be just as good numerically to do the whole thing in fp16?*

   If the whole thing is done in fp16, then the rounding errors accumulated in the summation would be of order $u_{\mathbf{fp16}}$ as opposed to the "mp2" variants which would have summation accumulation errors in the order of $u_{\mathbf{fp32}}$.
3. *Another general comment is that it seems like a large amount of text is devoted to comparing accuracy of and bounds for TSQR vs. HQR, but I'm not sure this comparison is very meaningful; the differences seem very minor (as evidenced in Figure 2).*

   We know that TSQR can be much faster when performed in parallel. The large amount of text comparing the accuracies of TSQR and HQR are needed to show that TSQR can still perform in similar accuracies, while we expect the computations to be faster.
4. *line 17: These are not my areas of expertise, but I did not think that QR factorization was used in any of the mentioned applications (clustering, ranking graph algorithms, neural network training). It therefore sounds strange to say that the investigation of mixed precision QR is a "first step" towards mixed precision implementations of such applications.*

   Not sure how to reply to this other than:

   We have removed the word "first". In the following paragraphs we discuss application that already use mixed precision computing, which is mainly restricted to training neural networks. Our goal in investigating the QR factorization is to increase the breadth of applications that can benefit from mixed precision computations.
5. *line 40-41: I still disagree with the way this sentence is written, which makes it sound like no mixed precision analyses exist. There are certainly mixed precision analyses of linear algebra routines, dating back to the 60s. You could change the sentence to something like "Many existing rounding error analyses ..."*

   Done.
6. *line 191: In (2.8), the summation upper limit should be m instead of n.*

   Done.
7. *lines 241: "swept into" rather than "swept under"; remove the word "assumption"*

   Done.
8. *line 242: "Using these two mixed precision settings". Which two? I only see one "MP Setting".*

   Clarified.
9. *line 248: "analyzed the inner product in an ad hoc mixed precision inner product". Should it be "... ad hoc mixed precision setting"?*

   Done.
10. *line 249: Don't need to spell out the HH acronym again, this is already done earlier. Actually, there is a bit of a confusion here. Earlier, e.g., line 46, Householder QR is abbreviated as "HH QR", but then in Section 3 Householder QR is abbreviated "HQR".*

    We will abbreviate Householder as HH for all instances except for the standard Householder QR algorithm, which we'll abbreviate as HQR.
11. *lines 298-299: Again, "swept under our use" sounds strange in this context. I would instead*

say “swept that minor difference into the constant defined in the $\tilde{\gamma}$ notation ...”.
Done.

12. *lines 305-306 and 308: “At iteration” → “In iteration”*
Done.
13. *Lemma 3.2: I still don’t understand where the “n” in the final product is coming from. Ditto in line 351.*
Fixed.
14. *line 352: “have 2-norm, 1” → “have unit 2-norm”*
Done.
15. *Algorithm 4: Capitalize ”householder” in caption.*
Done.
16. *line 395: Remove comma after ”matrix”.*
Done.
17. *line 407: Use “bFMA” acronym already defined.*
Done.
18. *line 409: “negligible” appears twice? Also, “does not” → “do not”*
Done.
19. *line 412: “result a” → “result in a”*
Done.
20. *line 417: “consist of” → “consists of”; spacing after “1)” and “2)”.*
Done.
21. *line 419: Consider changing “exact variant” to “particular variant” to avoid any confusion. Also, change “denote ... to be the outputs” to “let ... denote the outputs”.*
Done.
22. *line 430: “$(k-1)r+1^{st}$” → “$((k-1)r+1)^{th}$”, “kr^{th}” → “$(kr)^{th}$”. Or better yet, reword to ”... corresponds to columns (k-1)r+1 through kr of ...”*
Done.
23. *line 489: “outperform” → ”achieve higher accuracy than” ?*
Done.
24. *Section 3.3: Include some definition of what “Tall and skinny” means? (Doesn’t come until line 509)*
Added.
25. *line 504: Why are there parentheses around L in “$2^{(L)}$”?*
Removed parentheses.
26. *line 517: Capitalize beginning of sentence.*
Done.
27. *line 529: Should ”$2^{(\ell)}$” be ”2^{L}” here? Ditto 2x in the following line.*
Yes.
28. *line 538: Remove comma after ”matrix”.*
Removed.
29. *line 564: “stability is not guaranteed in HQR”...“the method is stable when using TQSR”. This needs a more careful definition of what you mean by “stability”. HQR is indeed normwise backward stable (we know this from a result of Wilkinson), and indeed, your example here obeys the known bounds on the normwise relative backward error if you plug in the parameters you used. It is not that the method isn’t backward stable, it is simply that the choice of parameters here means that the computed quantites are not very good.*

Changed to: "This case exemplifies a situation in which high accuracy is not guaranteed in HQR, but it is guaranteed when using TSQR. "

30. *line 565: Clarify this sentence. Something like "Note that these worst-case bounds are likely overestimates in practice".*
    Done.
31. *Figure 1: Make font bigger. Also, in the last paragraph in Section 3, you should maybe comment on the bounds on ΔQ, since this is what is shown in the figure (rather than the relative backward error).*
    Changes made.
32. *line 656: "casted up" → "cast up"*
    Done.
33. *line 657: "casted down" → "cast down"*
    Done.
34. *line 662: "Since $\hat{W}_k$, $\hat{Y}_k$'s are computed with alg. 4 in high precision then cast down" → "Since the $\hat{W}_k$, $\hat{Y}_k$'s in alg. 4 are computed in high precision and then cast down"*
    Done.
35. *line 663: "to some matrix stored in low precision, B" → "to some matrix B stored in low precision"*
    Done.
36. *Algorithm 7 caption: "of low precision A" → "on low precision A"*
    Done.
37. *line 679: ", we result in" → " results in"*
    Done.
38. *line 626: The latest A100 GPUs are not restricted to fp16/fp32 in the TensorCore instructions, which should be mentioned.*
    Added.
39. *lines 830: "varying size" → "varying matrix size"*
    Done.
40. *Figure 2: Font needs to be much bigger. What does the "h" mean here in the uniform precision setting, e.g., what is "HQRh" in the legend? The "h" in the label makes me think it is HQR in half precision, but the text says that the uniform precision is single.* **We have increased the font size the best we could with the legend holding all of the labels, and "h" here refers to high precision which is single precision in this experiment.**
41. Figure 3: I am very confused as to why the condition number has any effect. Overall, it is entirely unclear what the right plots are showing. What do the colors mean, and what does each line represent? What are the horizontal red lines? I would recommend just omitting this. Is this response okay? **In this experiment, we use the condition number to reach the "worst-case" error bounds. We have added explanations of the colors in the caption as well as the text.**

Reviewer #2 The authors have satisfactorily responded to all my comments and suggestions. I recommend the paper to be accepted and provide a few more comments and suggestions that do not require another review round.

1. *L61-63: the sentence defining the backward error that has been added is not so easy to understand , for clarity I would explicitly add that in the context of this paper the backward error is defined as $\|A - \hat{Q}\hat{R}\|$. Regarding the other reviewer's comment, while it is true*

that $\|A - Q\hat{R}\|$ may more commonly be called backward error, $\|A - \hat{Q}\hat{R}\|$ is certainly also a valid definition.
Clarified.

2. *L388–389: negligible is repeated.*
Removed.
3. *L388, 392, 653: compared to → compared with*
Done.
4. *L635–636: this sentence refers to a result from subsection 4.2 which comes after. I would move this comment to subsection 4.2.*
Moved.
5. *L650–657: again, this would be better placed after section 4.2.*
Moved.
6. *Table 5: I suggest adding in the caption the definition of every variable: m, n, N, etc. The objective being to make this table self-contained for other articles to refer to it.*
Added.
7. *Table 5, mpBQR2,3: be consistent in whether you factorize the $n^{1/2}$.*
Fixed.
8. *L901: imporved (typo)*
Fixed.