We thank all three reviewers for their insightful and constructive comments. Please find our detailed response below.

Reviewer 1 We appreciate the concrete suggestions for refining the presentation further. We are definitely going to add a subsection that describes technical background and key tools such as Johnson-Lindenstrauss transforms or CountSketch matrices to make the text even friendlier for the general reader.

The main benefit of Tensorized Random Projection (TRP) compared to TensorSketch (TS) is that TRP achieves high probability bounds unlike TS, which provably does not. Section 4.1 discusses and demonstrates this. We'll emphasize it more explicitly in the introduction to make it more prominent. Minor issues and typos:

- We'll delete references to unpublished [4, 28].
- Thanks for catching next two typos, fixed in our copy.
- *"(between line 155 and 156), S^ix^i should be vector while the right-hand side seems to be scalar"*: Indeed, it should be: Note that $(S^ix^i)_\ell$, the ℓth coordinate of S^ix^i, is $(1/\sqrt{m})u_i^{\ell,1}\langle v^\ell, x^i\rangle$. The rest of the proof continues as before.

Reviewer 2 Thanks again for reading the proofs carefully and for the helpful suggestions.

Clarifying *"The scaling given in Kar and Karnick ... For each sketch coordinate it randomly picks degree t ..."*: We believe both questions refer to lines 265-267, where we only meant to summarize Algorithm 1 on page 6 of Kar and Karnick really briefly. Scaling by Maclaurin coefficients was omitted by mistake, and will be addressed in the next version. Integer t in the second sentence is denoted by N in the original Kar and Karnick paper [29]. The first occurrence of the word degree referred to the exponent of a term in the Maclaurin series and the second referred to the order of the tensor created by raising the dot product of two vectors to the tth power. We'll rewrite and expand the description of Kar and Karnick's method to avoid any ambiguity. Minor points:

- $a = (1 \pm \epsilon)b$ form is common in our experience, nevertheless we'll replace these statements with the canonical $(1-\epsilon)b \leq a \leq (1+\epsilon)b$ form.
- Thanks for spotting the 4 typos, all are fixed.

The current proof of Theorem 2.1 requires that the sketch T maps to $t = \Theta(q^3/(\epsilon^2\delta))$ intermediate sketching dimensions. $S \cdot T$ is an improvement over S if t is less than input dimension n. δ was small in our experiments and $n \leq 100$ usually, except for MNIST where $n = 784$. Thus t would be much higher than the n dimensions we started with. The running time of S is already quite fast on sparse data. Nevertheless we could try $S \cdot T$ with a heuristic dimension set for T in the next version; thanks for suggesting it.

Reviewer 3 Thank you for the detailed review. Major comments:

- S is currently defined in lines 156-149, we'll pull that into a standalone definition preceding Theorem 2.2.
- Precise mathematical definition of T given in Theorem 2.4. It can be rephrased as follows: T sketches input tensor $x^1 \otimes x^2 \otimes \cdots \otimes x^q$ by applying a CountSketch, see [13], T^i to each x^i independently and outputs tensor product of T^ix^i. CountSketch of vector $v \in \mathbb{R}^n$ is $w \in \mathbb{R}^t$ such that $w_j = \sum_{i\in[1,n]:h(i)=j} v_i \cdot r_i$, where $h : [1, n] \to [1, t]$ is a hash function and r_i are iid ± 1 random variables.
- Kar and Karnick [29] proposed the Random Maclaurin (RM) sketch, and implicitly, without naming it as such, introduced Tensorized Random Projection (TRP) as a building block of RM. The two sketches are different. Figure 2 compares TRP and RM and demonstrates that TRP is vastly more accurate. RM expands the kernel into its Maclaurin series, randomly chooses a Maclaurin term, and uses TRP as a building block to estimate that term. A simple yet important message of our paper is that for polynomial kernels one does not need the Maclaurin series expansion because for the homogeneous kernel $\langle x, y\rangle^q$ exactly one Maclaurin coefficient equals 1 and all the others are 0. When the latter are chosen by RM the sketch is the constant 0, wasting space. The inhomogeneous kernel $(1 + \langle x, y\rangle)^q$ is best handled by augmenting the input with a constant 1 feature reducing it to the homogeneous case. Prior work [29, 34] somehow overlooked these facts and ran experiments only with RM. Our main theoretical contribution is an exponential improvement in the sketching dimension of a previously considered TRP. Prior work [29, 34] claimed that TRP required many more dimensions (discussed in Section D) and thus was much less practical. Our work shows that same exact sketch in fact has a much better complexity than was known. This also enables new applications of the TRP sketch, e.g., to neural networks, that might not have previously been considered given that people thought the behavior of this sketch was worse than it actually is.
- We'll increase font size, weight and line width of Figure 1 to improve contrast and repeat in its title that error bars correspond to one standard deviation. With regard to the $m = n$ case, for $m = 100$ TensorSketch's error is always the maximum possible, 1 at $n = m$, whereas the error of TRP is much lower and increases very slowly from about 0.37 at $n = 40$, where we truncated Figure 1(a), to about 0.39 at $n = m = 100$. It's still at most 0.41 at $n = 200$. We'll explain this in the text and plot a broader range.

Minor comments:

- Line 48: We'll define dot-product preserving sketching matrices before referring to them.
- Polynomial kernel of degree 2 is one of the most popular, degrees 3 and 4 are also used in practice, higher degrees are rare. We ran experiments with degrees 3 and 4 as well, the results were qualitatively similar. We'll include these in the final version.