We thank all reviewers for their insightful reviews which have helped us improve our work.

To Reviewer #4. We thank R4 for pointing out our contributions as well as some confusions. This review summarizes the highlights in our work, including the advantage of independence of dimensionality, analysis of optimal quantization parameter setting, and the extension to non-linear quantization.

The dimension dependent bound is tight so why does the re-parameterization help? Our paper is intended to provide insights for scaling real-life low-precision applications based on theoretical analysis. Because the dimension-dependent bound is tight under the standard assumptions, this means that if we want to understand when low-precision SGD may be dimension-independent, we need to look at it under different assumptions. The reparameterization in terms of L_1, σ_1 gives us these different assumptions, allowing us to show that (under our new assumptions) the outcome of low-precision training on large-scale problems does not necessarily go worse as the dimension scales. And based on this, we made analyzed non-linear quantization schemes such as logarithmic and floating-point quantization to show that, given the problem we are faced with, there is a way to adjust our quantization setting to optimize the performance. We made a comparison between our results and previous results in Table 1. As we stated in the analysis after Theorems 1 and 2, our result only reduces to the error level of prior analyses in the worst case, where parameter L_1, σ_1 are unfixed and can only be bounded by $L_1 \leq \sqrt{d}L, \sigma_1 \leq \sqrt{d}\sigma$; otherwise, our result is an improvement.

To Reviewer #5. We thank R5 for the encouraging feedback. This review perfectly summarizes our work and points out the advantages of our results and possible applications.

Presentation. We presented Table 1 with comparisons of our results and previous results to point out the novelty and improvement of our work. But as you suggested, we have moved it to appear after the assumptions. Parameter κ, κ_1 are introduced between Assumptions 3 and 4 on page four, as condition numbers. σ and σ_1 are bounds for loss gradients in different norms, σ_0 is the bound for the gradient variance, which is normally defined as σ in other works, so we added subscripts to distinguish them.

Analysis of other algorithms. We have applied our analysis to two other algorithms which use low-precision models: low-precision SVRG [13] and HALP [8]. We achieved similar dimension-independence conclusion and explored the application of non-linear quantization schemes for these algorithms—however, since the analysis was essentially identical and merely repeated our other claims about SGD, we did not include it. Other low-precision algorithms (e.g. [22,23]) use low-precision arithmetic in different ways (such as to store intermediate values used during gradient computation) to which our theory does not directly apply: we plan to explore theory for these algorithms in future work.

Parameter values. The parameter values mentioned by R5 were measured for the MNIST dataset: these are the smallest values for L, L_1, σ and σ_1 for which Assumptions 1–4 hold for multiclass logistic regression on MNIST. If larger (i.e. loose) values were used here instead of the reported ones, this would just result in a looser theortical bound.

To Reviewer #6. This review helps us understand what parts of our work are not explained well enough and may cause confusion. We thank R6 for the useful constructive feedback.

R6 points out that the parameters L_1 and σ_1 may depend on d, and is concerned about our overall bounds being dependent on d in this case. We have two responses to this. First, as shown in Fig 1(a) the standard dimension-dependent bound is in some sense tight, so we should expect to see classes of problems for which the performance depends strongly on d. For these classes of problems, our parameters L_1 and σ_1 will also increase strongly with d. However, there are classes of problems for which this does not happen, and for the class we study in Figure 2(a), the performance does not depend on d either, which is what our theory predicts. Second, even in the worst-case scenario when the parameters L_1 and σ_1 do depend strongly on dimension, our results in Table 1 show that, by using non-linear quantization, we can actually put those terms inside double $\log$ and get a $\mathcal{O}(\log\log d)$ upper bound when it comes to the number of bits required. This is better than the $\mathcal{O}(\log d)$ bound from previous work on linear quantization.

In the experiment, we choose the number of non-zero entries to be a fixed number $s = 16$ to guarantee that the parameters in Assumptions 1–4 are fixed, so that we can validate the dimension-free bound from our theorems. The results showed no dependence on the dimension d, as we expected. For denser cases with an increasing s, which would result in non-fixed model parameters, we do expect the performance to change, like what happened in Figure 2(b). But this does not contradict the results presented in our work, and is is concordance with what our theory predicts.

Other issues. • $\mathrm{dom}(\delta, b)$ denotes the domain of low-precision numbers that can be represented based on parameter δ and b, i.e. $\{-2^{b-1}, -2^{b-1}+1, \cdots, -1, 0, 1, 2, \cdots, 2^{b-1}-1\}$ times δ. • As we showed in Table 1 and analysis after Theorem 4, non-linear quantization is better than linear quantization by a $\sqrt{d}$ factor when it comes to the number of bits needed. • Proving convergence with a constant step size is actually a stronger result than using decreasing step size. • We included the different setting to show that our work can be applied to various problem classes in real-life applications. Though the 8-page length limits our discussion to some extent, we were able to present the highlights of our results and include detailed analysis in the appendix.