We thank the reviewers for their helpful feedback.

Related Work: We have actually already revised our paper to include a more thorough discussion of the SCAFFOLD paper. It is important to note that SCAFFOLD is presented in a related, but different, FL setting where only a subset $S < M$ of the machines are available in each round of communication. Specialized to our setting ($S = N$ in their notation), the SCAFFOLD analysis actually does not show any improvement at all over MBSGD (compare their Thm III to Table 1 in our paper). As we will describe in the final version, SCAFFOLD is like Local SGD with variance-reduction for the inter-machine variance; this helps in the FL setting when some machines aren't available in each round; but it does not in our setting. In addition, for the stepsizes analyzed in their theorems (specifically, very small η_l), SCAFFOLD is actually little different MBSGD.

We compare to the Khaled et al 2020 paper mentioned by Rev #2 as ref [9], under the name of an earlier version of that paper. We will update the citation (the relevant content is unchanged).

Regarding the homogenous case (requested by Rev #3): the paper [22] studies the homogenous case in detail and includes most of the relevant references–we will add a comment directing to [22] as well as a brief mention of the references. Our focus here is the difference between the homogeneous and heterogeneous settings.

Relationship to consensus optimization: As Rev #2 points out and as highlighted by our results, communication between the machines is often the bottleneck in heterogeneous optimization, and consequently, (Acc) MBSGD will often significantly outperform Local SGD.

This may be intuitive in the context of consensus optimization, but it is our experience (eg based on papers on federated/local SGD, talks on distributed learning, and comments from other researchers) that this is far from clear to everyone working on distributed learning. E.g., following demonstration that Local SGD can be worse than MBSGD in the homogeneous case [22], a recurring sentiment is "well that's just the homogeneous case, in the harder heterogeneous setting, you'll see more of an advantage for Local SGD." But we show (as Reviewer #2's intuition correctly indicates) that this is backwards!

For this reason, we feel there is significant value in understanding and highlighting the relationship between Local SGD, MBSGD, and other algorithms in the heterogeneous setting, and in carefully considering how the level of heterogeneity affects the comparison. It's also important to test the limit of this intuition. E.g., we do show in Theorem 3 that in some heterogeneous regime, additional computation as in Local SGD DOES improve over MBSGD (this is not captured nor hinted by work we are aware of on consensus optimization).

$\bar{\zeta}$: Indeed, eq (12) is a strong assumption and the tightest bound on $\bar{\zeta}$ may be large (or infinite). However, in cases where this is bounded and smaller than $1/R$, our analysis shows that local SGD can outperform minibatch SGD, which we feel is useful information. As an example of where this could arise, consider the following: data is shuffled and randomly partitioned across the M machines and the local distributions are the empirical distribution over the local sample. These local distributions ARE heterogeneous, nevertheless, as long as there are enough samples on each machine and, for example, the loss is Lipschitz, $\bar{\zeta}$ will be bounded and small, and we can conclude that local SGD might be advantageous over MBSGD in finding the over empirical minimizer.

σ_* **vs** σ: As Rev #1 mentioned, we used σ_* for the MBSGD analysis and sigma for the Accelerated MBSGD analysis. We agree that analyzing Acc MBSGD in terms of σ_* is interesting and valuable, but the known analysis (due to Lan) is in terms of σ and quite delicate, and as also acknowledged by Lan, extending it to σ_* is a significant challenge.

Variance definitions: We thank Rev #2 for the suggestion that the variances in eq (6)/(7) be the average of the local distributions' variances, this indeed gives stronger bounds and easily fits into our analysis.

Quadratics: Rev #3 raises an interesting question about whether the Local SGD analysis can be improved in the special case of quadratic objectives. This is definitely possible in the homogeneous case, where Local SGD strictly dominates MBSGD in all regimes for quadratic objectives. In the heterogeneous case, the argument does not go through in the same way and in addition, the ζ_* term in our lower bound for Local SGD comes from a "quadratic part" of the hard instance construction. Therefore, Local SGD won't benefit significantly from the local objectives being quadratic, at least not in terms of the ζ_* dependence.