The authors would like to thank three reviewers for their valuable feedback. Our point-to-point response are provided below.

Reviewer 1: We thank the reviewer for the valuable feedback. Your concerns are addressed as follows:

1. Improvement compared to [20] and [21]. Compared to [21], the improvement of convergence comes from the improved bound of $2\alpha_t \sum_{k=1}^{N} p_k \left[F_k(\mathbf{w}^*) - F_k(\overline{\mathbf{w}}_t)\right]$ (as we mentioned in L489, Appendix C.1). Because of this improvement, we are able to cancel square of gradient norm ($\alpha_t^2 \|\overline{\mathbf{g}}_t\|^2$) in the upper bound of A1 (defined in L483, also see explanations in L492) while in [21] this term remains as leading term ($6\eta_t^2 L\Gamma$).

Our work is not an extension of [20], we have discussed that our assumption is a more relaxed assumption comparing to the bounded gradient diversity in [20] (see Appendix B).

2. The overparameterized case seem to be out of context here with no experimental results to back them up. The overparameterized setting is an important setting [29, 30], especially in the era of deep learning where the neural network can achieve zero training loss. We provide the geometric rate with linear speed up in this setting including a general overparameterized setting and overparameterized linear regression. Experimentally, we verify the linear speedup results of overparameterized setting in 3rd column, Figure 1 and Figure 2.

3. The novelty and main results. We would like to emphasize that the techniques used for different settings to achieve linear speedup are quite different (strongly convex see L489, L492, convex smooth see L553). The main contributions include the linear speedup results in strongly convex, convex smooth, and overparameterized setting, for both FedAvg and its accelerated variants (this is the first linear speedup results of accelerated FedAvg). We further propose FedMass, which is the first algorithm that enjoys an improved convergence rate comparing to FedAvg. We respectfully disagree that our main contribution is the improvement over [21] as the linear speedup results under other settings are not a trivial extension of the strongly convex case.

Reviewer 2: We thank the reviewer for the valuable feedback. Your concerns are addressed as follows:
1.The range of N/K.

Reviewer 3: We thank the reviewer for the valuable feedback. Your concerns are addressed as follows:
1.Comparison with SCAFFOLD.:

- The analyses in SCAFFOLD do not imply $E = \mathcal{O}(\sqrt{T})$ but in fact also require $E = O(1)$ under partial participation (sampling). To see this, it is easier to examine Theorem V in their paper, which is stated in terms of optimality gaps, but otherwise equivalent to their Theorem I. Notice that when $S < N$, the second term in M^2 gives rise to a term $O(1/R)$ in the bound for the strongly convex case (ignoring constants), and since $R = T/E$ in our notation, this term is $O(E/T)$. Therefore, in order to achieve a $O(1/T)$ convergence rate for the optimality gap, it must be the case that $E = O(1)$ as well. Similarly for the general convex case. Thus in terms of communication complexity, our results imply the same requirements for both full and partial participation as that in SCAFFOLD.
- Furthermore, the sampling procedure analyzed in our paper is different from that in SCAFFOLD, as we allow the sampling probability to scale with device specific weights and sample with replacement, whereas in SCAFFOLD the sampling is uniform without replacement.
- Under our sampling schemes, we explicitly analyze the contribution of sampling variance to the optimality gap, and in Lemma 10 of our paper in the Appendix, we provided a problem instance that lower bounds the sampling variance, showing that $E = O(1)$ cannot be improved in general when there is partial participation with sampling.

2. The convergence of Nestrov Accelerated FedAvg. As there has not been any linear speedup analysis of Nesterov accelerated FedAvg for convex problems, our result is a first in this regard and completes the picture. Our analyses of FedAvg and Nesterov FedAvg are also unified, highlighting the common elements and distinctions for the two algorithms, which has not been done by previous studies.

3. The geometrical rates of overparametrized linear regression. The geometric rates in Theorem 5 are for general overparameterized strongly convex objectives rather than just linear regression, and as SCAFFOLD is a variance reduction based algorithm, our result, which is on the FedAvg algorithm, is not directly comparable to the geometric convergence result in SCAFFOLD.