Response to Reviewer 1. Thank you for your compliments on our contributions as well as our presentation!

Response to Reviewer 2. Thank you for your thoughtful comments and questions!

Q. Significance of relaxing component convexity? For practical significance, it is crucial for understanding the (local) behavior of shuffling algorithms for nonconvex problems, such as neural network training. As Reviewer 3 also mentioned, in such settings, the function F would behave like convex near a neighborhood around a local minimum, while each component function f_i could be highly nonconvex in the neighborhood.

For theoretical significance, this relaxation signifies an innovation in our proof techniques. In fact, the component convexity is heavily exploited in the previous works [10, 13]. At a high level, the previous works use this assumption to ensure that the algorithm makes a sufficient progress during *each* iteration. In contrast, our proof technique proves fast convergence rates without having to show such a per-iterate progress. We also highlight that it is thanks to our proof technique that we obtain the first optimal rate for nonconvex PL function class in Theorem 1.

Q. Constant step algorithms vs. decreasing step size algorithms are not comparable? We view the decreasing step size case as choosing a different set of hyperparameters for the *same* algorithm. In light of this, our main goal in analyzing decreasing step sizes was to see whether the common limitation in the prior works (large epoch requirement) is inherent to the algorithm. Our main results indeed show that the common limitation is a derivative of the constant step size choice. On a technical note, capturing the desired convergence rate with decreasing step sizes requires a non-trivial analysis (Section 6.3 or Appendix E), and we believe that our work develops a toolkit for analyzing decreasing step sizes for epoch-based algorithms like shuffling SGD.

Q. Guarantees without Assumption 1? We agree that Assumption 1 is a bit unsatisfactory, but please note that this assumption is also present in many prior results (see Line 144). While our Theorems 1 and 2 *do not require* Assumption 1, Theorems 3 and 4 *do* rely on Assumption 1 because they are built on existing results [10, 13] that make use of Assumption 1. We believe that removing this assumption, as done in [12] or a concurrent work "Random Reshuffling: Simple Analysis with Vast Improvements," is an important future research direction. We will add a remark on this in our revision.

Q. Comparison to GD? Our current proof techniques do not show a regime where rates for shuffling SGD can be better than that of GD. Indeed, this limitation is shared among the existing works, as their analyses rely on comparing the epoch progress of shuffling SGD to that of GD. In other words, this question is currently beyond the scope of existing proof techniques, and in particular, finding a regime where shuffling SGD outperforms both SGD and GD would be an interesting future direction to pursue. We will add this discussion in our revision.

Response to Reviewer 3. Thank you for detailing our contributions into a list. In particular, we agree that the removal of individual convexity has important consequences in practical applications. Also, as per your suggestion, we will explicitly state the definition of the constant G in Theorem 2.

Response to Reviewer 4. Thank you for appreciating the strengths of our paper! We did not spell out G in Theorem 1 because existence of G is implied by the other assumptions. We will explicitly state the constant G in the theorem statement as per your suggestion.

On a separate note, we would like to fix one technical mistake in the proof of Theorem 1. After the submission deadline, it came to our attention that the Hoeffding-Serfling (HS) inequality used in the proof of Theorem 1 can be applied only to RandomShuffle. In SingleShuffle, the first iterate $\boldsymbol{x}_0^k$ of the k-th epoch is not independent of the permutation σ as soon as $k > 1$, hence we cannot apply the HS inequality. In light of this, we note that Theorem 1 only holds for RandomShuffle. As to our claims on SingleShuffle, we will add an additional theorem in the revision. The new theorem shows that if F is μ-strongly convex and f_i's are L-smooth, SingleShuffle with number of epochs $K \geq 10\kappa^2 \log(n^{1/2}K)$, step size $\eta_i^k = \eta := 2\log(n^{1/2}K)/\mu n K$, and initialization $\boldsymbol{x}_0$ satisfies $\mathbb{E}[F(\boldsymbol{x}_0^{K+1})] - F^* \leq 2L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2/nK^2 + c \cdot \log^3(nK)/nK^2$, for some $c = O(\kappa^4)$.

Although this change slightly weakens our initial claim on SingleShuffle, we believe this new theorem is still interesting progress, because (i) it is a tight bound (up to log factors) for SingleShuffle on strongly convex functions; (ii) it does not require convexity of individual components or bounded iterates assumption (Assumption 1); (iii) it shows that the optimal rates for minimizing strongly convex functions are the same for RandomShuffle and SingleShuffle; and (iv) the proof can be easily extended to any algorithms between the spectrum of RandomShuffle and SingleShuffle. The proof of the new theorem involves an end-to-end analysis similar to the one-dimensional quadratic result [16] and a modified version of the approximate matrix AM-GM inequality in eq (6.4). For this theorem, the HS inequality is applied only to the partial sums of the individual gradients at the global minimum $\boldsymbol{x}^*$, which is always independent of the permutation σ.