We thank the reviewers for their work and for the positive evaluation of our paper. Every reviewer noted that it is well-written and reproducible. There seems to be no disagreement that the paper is of practical value (noted by R1, R2, R4), relevant to the conference (R2, R4), and is generally on an interesting topic (R1, R2).

To all reviewers. We realized that we downplayed our contribution: Theorems 2 and 3 apply to Shuffle-Once without a single change in the proofs, just as in Theorem 1. Thus, we also provided guarantees for SO without strong convexity.

Reviewer 1. 1."*this work uses the fact that all component functions f_i are μ strongly convex*" This assumption is not strong because strong convexity typically comes from ℓ_2 regularization or weight decay, which is normally used in each function. Adding a small amount of regularization is also a common practice for numerical stability.
2. "*why removing some of the assumptions like bounded variance and bounded gradients is an important contribution*" This assumption does not hold for any strongly convex function, which includes any convex objective with ℓ_2 regularization. It also doesn't hold for least squares, matrix factorization and neural networks (with smooth activations). These problems are at the core of modern machine learning, and there are a lot of other similar objectives.
3. "*The quantity σ_* being finite also implies that all the gradients are finite*" This is not true, σ_* is always finite just because $\nabla f_i(x_*)$ is a finite vector for any i, but globally the gradient is often not bounded (see the previous item).

Reviewer 2. We appreciate your support of our paper. Since accepted papers will be allowed to have an extra page, if ours gets accepted, we will be happy to add discussion on the shuffling variance, which we agree will be educational.

Reviewer 3. 1. "*the main theorems (Theorems 1 and 2) need a small step size, similar to previous works. In fact Safran and Shamir (2020) show that convergence is only possible for step size $O(1/n)$*" Firstly, we disagree about Theorem 1—even with step size $\frac{1}{L}$ it guarantees convergence to a neighborhood. In practice, achieving machine precision is sometimes not important, and a neighborhood would suffice. Furthermore, a substantial part of our work is devoted to explaining why the neighborhood's size is small. Secondly, the argument of Safran and Shamir (2020) does not show convergence is not possible with large stepsizes, indeed the proofs of Propositions 1 (p. 10, 2nd bullet point) and Thm. 2 (p. 16, 4th bullet point) in their work both show that convergence proportional to n and T may be achieved with a large stepsize. We will add these clarifications.
2. "*the dependence on μ has worsened. In particular, Nagaraj et al. (2019) give an error rate of κ^2/μ*" This comparison is not fair because Nagaraj et al. (2019) bound functional gap and our bound is for the distances. If we apply $\|x - x_*\|^2 \leq \frac{1}{\mu}(f(x) - f(x_*))$, we see that the bound of Nagaraj et al. gives rate $O(\kappa^2/\mu^2)$.
3. "*The result of Theorem 3 says that RR beats SGD only after $\Omega(n)$ epochs, which seems really large. This can probably be improved.*" This claim is somewhat speculative: we are unaware of any lower bounds in this setting. We devoted a lot of time to tightening this bound and the complexity that we proved is better than any other available in the literature for the same setting, so it is unclear why a better bound should be possible.
4. "*Theorem 1 assumes that the individual functions are strongly convex. This is a really strong assumption which, in my opinion, the previous papers have deliberately avoided.*" We disagree that it is such a strong assumption (see our response to Reviewer 1). "*I also believe that this assumption makes the analysis much more easier because now the sum of every subset of the individual functions is also strongly convex.*" Our proof does not use that any sum of a subset of the individual functions is strongly convex, so we do not see how this remark is relevant.
5. "*The literature review is somewhat poor.*" We wrote a longer review but had to make it shorter because of the page limit. If the paper gets accepted, we would be happy to use part of the allowed additional page to put more discussion.
6. "*Also, all the relevant upper bounds should be compared in detail with the lower bounds given by Safran and Shamir (2019) and Rajput et al. (2020).*" We already compared the bounds with both papers, check page 6, in lines 177-185.

Reviewer 4. 1. "*Does the polynomial dependence on κ also match the lower bounds?*" To our knowledge, there is no lower bound with an explicit dependence on κ. As a sanity check, if $n = 1$, then $\sigma_* = 0$ and RR reduces to gradient descent and we obtain the standard dependence on κ, so it is tight at least in this sense.
2. "*Can the authors discuss if their results for the non strongly case can be improved or are there known lower bounds?*" We are not aware of any such lower bound. For SGD, recent work (Sebbouh et al. "On the convergence of the Stochastic Heavy Ball Method") shows that momentum makes SGD faster when there is no strong convexity, despite SGD being optimal in the strongly convex case. We think that momentum may improve the rate of RR as well.
3. "*Can the authors please comment on if a similar analysis might be possible for SO?*" We double-checked these results and they turned out to apply to SO without any change in the proof.
4. "*it seems like $\sigma^2_{\text{shuffle}}$ is much bigger than σ^2_* unless the step size is very small*" If $\kappa = 1$ and $\gamma = \frac{1}{L}$, then indeed the shuffling variance can be in theory n times bigger than σ^2_* but we didn't observe this in our experiments.
5. "*I think it would be useful to write the SGD convergence rate as in equation 4*" We can add it: the first term will be the same as in (4) and the second term would be $\frac{\sigma^2_*}{\mu^2 nT}$, where nT is the total number of steps.
6. "*For figure 2, how are the values $\sigma^2_{\text{shuffle}}$ and σ_* estimated? And, what are the values of L and μ that are used?*" We first obtained x_* by running Nesterov's acceleration until machine precision, and then we estimated expectations by randomly sampling permutations, and σ_* was computed exactly.