We thank all four reviewers for their insightful comments and helpful feedback. Most of the reviewers describe the paper as very well written and clear, and most identified its contribution as being important to the ML community, as understanding scan orders is a fundamental problem in this space. We now address the reviewers' concerns individually.

R1. Why is it necessary to use a parallel SGD example? Good point! The parallelism itself is not necessary; what is necessary is the *averaging*. The averaging is necessary (even for large n) because it models the expected value in the original Recht and Ré inequality: without it the convergence rate may be effected by higher-order moments (not just the expected value). We study the non-averaged case in Section 5. We use parallel SGD as an example because it is a "real" algorithm that uses averaging (one that has been previously proposed in [9,26]). We will clarify this in the text.

R2. What intuition is given about the structure which causes the behavior? Could a class of problems be constructed? Part of our goal in Section 2 was to construct a class of multiple counterexamples of arbitrarily high dimension, to give better intuition. For example, in Section 2.1, our construction produces a counterexample whenever

$$|\lambda| = \left| \left(1 + \frac{1}{n-1}\right)^{n/2} \cdot \cos\left(n \cdot \arcsin\left(\frac{1}{\sqrt{n}}\right)\right) \right| > 1, \qquad \text{for example, when } 5 \leq n \leq 16 \text{ or } 29 \leq n \leq 51 \text{ etc.}$$

This family of counterexamples is inherited by our other counterexamples in 2.2 and 2.3, which depend on it. As far as we can tell, the "structure" responsible for this phenomenon (which allows us to get at the counterexamples) is the permutation-matrix-symmetry that we discuss at the beginning of Section 2.

R2. Another weakness is the requirement that the number of matrices n to be equal to the dimension d. In linear regression, n is typically much larger than d. This can be avoided in the counterexample by simply adding additional I matrices to the ensemble, since (1) this does not change the arithmetic mean, which is already I, and (2) this does not change the random-reshuffled product, which is not affected by multiplying by I. This approach can generate counterexamples with n arbitrarily large compared to d. We will add text to clarify this.

R2. Citations. We thank R2 for the additional very helpful citations, which we will add to the paper.

R4. We thank Reviewer 4 for a clear and insightful review. We share R4's hopes that "counterexamples will provide new intuition to the community about what makes these different sampling strategies work (or not work), and this new intuition may lead to new, deeper understanding of these sampling methods in SGD," and we view this as a great summary of the impact we hope our results will have on the community.

R3. Writing. R3 says that they find the paper poorly written and difficult to follow. This writing issue seems to permeate and color the rest of R3's review: for example, they could not identify any strengths at all in the work, and say the contribution is unclear. We apologize for the confusion. This is perplexing to us because we are unsure what makes the writing unclear, and all the other reviewers said that the paper was clear, saying things like "very clear" and even "exceptionally clear." Perhaps our writing was insufficiently clear that main point of the paper is to present *negative results*, concrete counterexamples to longstanding conjectures in the space for which no previous constructive disproofs were known. We will try to improve this as we revise our manuscript.

R3. Note that the convergence rate of $O(1/t^2)$ for random shuffling should also depend on "n" while the one of $O(1/t)$ for regular one does not. Recent work ("SGD without Replacement: Sharper Rates for General Smooth Convex Functions," which Reviewer 2 just made us aware of) gives a convergence rate of $O(\frac{1}{nK^2})$ for K epochs of random-reshuffling on n examples, while the best result with replacement is $O(\frac{1}{nK})$. So while n is involved, in this setting RR does indeed have an asymptotically better class upper bound. The fact that this problem-class upper bound does *not* transfer to RR being asymptotically better than with-replacement SGD for individual problems is interesting, and this is what we are trying to get at in this paper.

R3. The problem in Definition 1 is not really practical. Could you provide some specific ML problems satisfying all the conditions? This class roughly corresponds to the setting of the Randomized Kaczmarz method originally studied by Recht and Ré [17]. Here, the task is to solve a system of linear equations $Ax - b = 0$ (which for simplicity we assume has a unique solution) by minimizing the objective $\frac{1}{2n}\sum_{i=1}^{n}(a_i^T x - b_i)^2$, where a_i are the rows of $A \in \mathbb{R}^{n \times d}$ and b_i the entries of $b \in \mathbb{R}^n$. Since the problem has a unqiue solution x^*, the gradient of each component of this loss is 0 at that point, and they are of course obviously convex quadratics. It is fairly easy to choose a_i such that the problem would be non-trivial. (See also our statement in Footnote 2 about why we did not try to generalize this class further.)

Also note that Definition 1 is only used to prove Corollary 1, which is introduced to give some context/interpretation for Theorem 1, which itself is the main technical result of the section—Definition 1 and the problem class it defines are not central to the claims of the paper.

R3. Reproducibility and Broader Impact. The code to reproduce all the figures in the paper is given in the supplemental material, as are all the proofs. As such, it is not clear to us why R3 evaluates the paper as not reproducible. We are equally confused about the "no" on broader impact, and would appreciate some feedback to improve that section.