We thank the reviewers for their comments. We first discuss some overall concerns raised by multiple reviewers, then proceed to more specific points.

- *Results for non-convex*: Some reviewers mention the lack of results for non-convex settings as a weakness of the paper. We highlight that to the best of our knowledge, our work is the first to provide precise theoretical regimes in which data-echoing could lead to significant advantage. Indeed, if mere convergence results are expected, it is easy to show that data-echoing (with an appropriate learning rate) converges in the non-convex case. Providing the precise characterization of the benefits of data-echoing in non-convex settings is indeed very exciting future work. We believe that the convex results provided in our paper will eventually provide the foundation for the non-convex results, as they have in any other field of optimization.
- *Paper structure*: Reviewer 4 raised the concern that "the paper and the math seem unconnected at times". We disagree with this remark. To clarify the general structure and role of the proofs (as is already laid out in the paper):
  - 3 algorithms are analyzed (Thms. 7,10,13). The proof for each has the same structure viz. proving potential bounded regret (Def. 3) and stability (Def. 1). Both properties are needed. This common structure is laid out in Thm. 4.
  - Lemmas 5, 8, 11 prove the potential bounded property. Due to similarity the proofs are bundled in Appendix A.
  - Lemmas 6, 9, 12 prove the stability. Due to similarity the proofs are bundled in Appendix B.
  - Finally the proofs of the main theorems are bundled in Appendix C. There are no other theorems/lemmas in the paper.

Reviewer 2: We appreciate the positive feedback, and the pointer to confusing notation. As you note, the notations are not incorrect but could be confusing, and this will be addressed promptly in a revision.

- Average iterate as opposed to last iterate is indeed done to produce a simple and generalizable analysis. We believe that akin to SGD (using techniques as in [Shamir & Zhang '12]), last iterate guarantees can be obtained here.

Reviewer 3:

- *Resampling batches*: Data-echoing is not being proposed as a general alternative to stochastic optimization methods. If batches can be sampled at a rate compatible with the computation of gradients, then one should resample at every iteration. **Data-echoing is relevant when batches cannot be sampled as fast** and we highlight the regime when it could be advantageous over the current practice (of doing nothing).
- *Stochastic AGD will not converge*: The reference given in the table (to Lan's AC-SA) provides an accelerated method for smooth stochastic optimization with the rate mentioned in the table. We disagree with the "well-known result in first-order optimization community" that contradicts this. If your objection is that Nesterov's acceleration does not directly work for SGD, as noted by Reviewer 4, then this will be clarified.
- *Comparisons with Adam, etc.*: This is moot with respect to the scope of the paper. Such comparison requires developing principled data-echoed variants of adaptive methods. A first attempt at this comparison was made in [Choi et al. '19]. Here we focus on providing a theoretical foundation for data-echoing. Data-echoed adaptive methods are interesting for future investigation.

Reviewer 4: We thank the reviewer for reading closely and pointing out typos. We stress that the issues pointed out are mere typos, as we highlight below. (We provided a detailed discussion regarding the point of "unconnected math" earlier.) We request the reviewer to revisit these and consider increasing their score, post clarifications.

- *Nesterov Acc vs Lan's AC-SA*: Thanks for pointing out this nuance, which might confuse other readers. We will clarify in the final version.
- *Unrelated appendix*: Appendix B contains stability proofs (for Lem 6,9,12). Lemmas are at lines 168,179,204 in the main paper and are necessary for the main proofs (Thms 7,10,13). It is clearly stated that proofs appear in the appendix.
- *Related work*: Resampling of batches as a trick to *reduce variance* is of course ubiquitous in literature (equivalent to increasing batch size). Performing *multiple gradient steps on the same batch* has not been analyzed in stochastic optimization as far as we know. Nevertheless, we request that the reviewer provide precise references, and we will include and discuss them.
- *Typos*:
  - line 322-323: These are misplaced references: (4) and (6) should read (2) and (3).
  - Lemma 8: The only typo we spot here is Line 324 should say γ strong convexity instead of λ. The step-size is clearly defined in the description of the algorithm in Line 137.
  - Lemma 11: There is no indexing error as far as we can see. There is however a minor typo below Line 336, that might be the source of confusion. Corrected version below:

$$\lambda_t \|\eta \nabla f(x_{t+1})\|^2 - 2\eta \nabla f(x_{t+1})^\top (w_t + \lambda d_t - w) = \lambda_t^{-1} \left(\|w_t + \lambda_t d_t - w - \lambda_t \eta \nabla f(x_{t+1})\|^2 - \|w_t + \lambda_t d_t - w\|^2 \right).$$

  - Line 351: Indeed, RHS is the difference between gradients. Thanks.