We thank all reviewers for reading our paper and appreciating our results. We respond to each reviewer's comments below.

Response for Reviewer #2.

1. In a classical stochastic approximation framework it is well known that choosing too small step sizes can lead to slow convergence while larger step sizes improve convergence but suppress noise poorly [see e.g. Nemirovski et al. 2009]. In order to suppress, to some extent, noisy trajectories while keeping large step sizes and thus fast convergence, a common approach is to take appropriate averages of the iterates. The intuition behind tail-averaging is that an appropriate choice of the tail-length realizes a trade-off between robustness (less noise) and fast convergence. In a nutshell, averaging allows larger step-size since it reduces the variance of SGD. Tail averaging with sufficiently "long" tail preserve this benefit. Characterizing the exact length which allows this effect is one of the contribution of our analysis.
2. By the saturation effect and the improvements provided by tail-averaging being "purely deterministic", we mean that both already manifest themselves when considering (deterministic) population gradient descent rather than stochastic gradient descent. Specifically, we show in Section 3 that population gradient descent with full averaging suffers from a saturation effect, which can be remedied by tail-averaging. Our detailed analysis of SGD shows that the same observation can be used to obtain improved bounds on the approximation error.
3. In line 219, we refer to the contributions of Jain et al. [19], listed in their Section 3. We will rephrase this sentence to avoid potential confusion in the final version.
4. Correct, thanks for pointing this out!
5. Also correct, thanks!
6. Indeed, the first experiment shows the results after a single pass over the data. We will make this clear in the final version.

Response for Reviewer #4. You are correct to point out that the bulk of our analysis uses tools that have been known before, but nevertheless the observation that tail-averaging can eliminate the saturation effect associated with uniform averaging is novel. Also non trivial work was needed to extend previous results to encompass both tail averaging and mini-batching, specifically to control the variance of SGD. Regarding the experiments, we feel that they are on par with those included in most theoretical papers on the same topic at venues like NeurIPS, COLT, ICML, or AISTATS. Their purpose is to illustrate theoretical findings, rather than providing a thorough empirical analysis. Finally, let us respectfully point out that the second set of experiments illustrated on Figure 1.b-c actually illustrates the interplay between 3 parameters: tail-averaging, step size, and batch size. We opted to focus on the effects of tail-averaging since we felt that this emphasizes the main novelty of our theoretical analysis.

Response for Reviewer #5. Regarding the experiments, we feel that they are on par with those included in most theoretical papers on the same topic at venues like NeurIPS, COLT, ICML, or AISTATS. Their purpose is to illustrate theoretical findings, rather than providing a thorough empirical analysis. We expect the conclusions of the experimental section to hold for large real datasets as well, since such data tends to verify our Assumption 2 for large r. Regarding other convex and smooth losses such as the cross-entropy or the logistic loss, we expect most results to hold but proofs will be considerably different and details might change.