Reviewer 1

1. Our focus on SVRG is largely a pragmatic one. As you mention, there are several alternative methods that have limited memory variants, although they broadly fall into categories of SVRG-like or SARAH-like. We did experiment with SARAH after submission, and we were unable to get it to perform as well as SVRG. We intend to discuss this in the final draft. We have not experimented with the other methods mentioned, so it's hard to say, but it's clear that they don't directly address any of the fundamental problems we discuss.
2. We will limit the scope of our statements as you state. Our title is certainly overly broad.
3. We did use one sample per image, over the whole dataset, which we found gave accurate enough estimates. The plot looks almost the same between runs with different seeds. We will need to rerun the experiment to provide standard error estimates for the final draft.
4. We can certainly provide more details here.
5. See 3 above.
6. We didn't produce the requested plot, we can certainly run it for the final draft.
7. Yes, we do need to add some additional experimental details here. We did use a fixed window.
8. For both SGD and SVRG we used the random permutation technique you mention, we tried to follow standard practice as much as possible. We will update the paper to reflect this.
9. We are certainly willing to commit to publishing code. We are not able to provide a GitHub link as the author response guidelines explicitly prohibit external links.

Reviewer 2

Originality section: - We do believe we are the first to provide such a comparison. We closely follow the Google Scholar feed of citations to SVRG and have not found any comparable work.

Quality section: - We will perform the requested experiment using a ResNeXt-101 model, which should meet the stated requirement. We currently believe based on the scaling shown in the paper that networks with more parameters will only show less improvement under variance reduction.

We will add the requested error bars to the CIFAR-10 plot.

We hope that Reviewer 2 will revise their rating based on the more complete empirical comparison we commit to providing.

Reviewer 3

- We are not aware of any experimental or theoretical results that show that SVRG may be useful in situations where the variance is higher than SGD at each step. This is an interesting question though. We deliberately provide a simplified view by measuring variance reduction directly.
- We used hyper-parameters that are known to be optimal for SGD. We did experiment with other values when using SVRG, but we found the same parameters to be optimal. We will update the paper to reflect this.
- Changing the learning rate or mini-batch size will not solve any of the fundamental problems we identify as roadblocks to the application of SVRG.