Since reviewer 3 raised a most critical issue, we respond to this reviewer first.

Reviewer 3: **Q:** *The three papers are extremely similar.*

A: We are so shocked about the reviewer's judgement. The contributions of three papers are fully orthogonal, and completely under three different research lines/topics: **paper 1410 (this paper)** on proximal and momentum algorithms, **paper 2460** on batch size adaptation, and **paper 1495** on reinforcement learning (RL) algorithms. In the past, there have been extensive studies under each research line, and many top conference papers contribute only to one line.

Q: *All three papers focus on developing a practical variant of SPIDER algorithm which achieves same theoretical bounds as SPIDER.*

A: First, not all three papers focus on SPIDER algorithms. Only paper 1410 focuses on proximal and acceleration of SPIDER. The other two papers study other aspects of stochastic algorithms (paper 2460 on batch size adaptation, and paper 1495 on reinforcement learning) and both analyze three representative stochastic algorithms SGD, SVRG and SPIDER as examples. Inclusion of SPIDER in these two papers is mainly because it is the state-of-the-art stochastic algorithm. Even for SPIDER-type algorithms, analyzing them under three directions still requires significantly different treatments. They do not achieve the same theoretical bound either.

Specifically for this paper 1410, we emphasize its difference from the other two papers below.

- This paper makes two major contributions: develop and prove that **proximal** SPIDER algorithm has better complexity order than existing art; and develop **momentum** SPIDER algorithm and prove its performance guarantee. Neither of the other two papers even touch the topics of proximal and momentum algorithms.
- The performance guarantee for both algorithms in this paper are nontrivial, and require considerable new technical developments specifically for proximal and momentum algorithms, which were absolutely not in the other papers' proofs. In fact, the proof of vanilla SPIDER does not extend to Proximal SPIDER.

Q: *To increase the score: (1) At the very least each of your papers should devote a paragraph to explaining the distinct contributions of each paper and ideally how they form a cohesive body. (2) Personally, I'd suggest reducing the number of papers. For example, creating one long, high quality paper.*

A: Regarding (1), both papers 2460 and 1495 have already discussed their differences from paper 1410. Paper 1410 did not discuss the other two papers because they have not been released publicly yet.

Regarding (2), the three papers are developed under different research lines and should naturally be written separately. **Specifically for this paper 1410, we have already made the best efforts into including as many relevant results as possible to supplementary materials, which we believe has made a comprehensive body of work by itself.** Even if we try to combine them, it is unrealistic for NeurIPS. Clearly, the reviewer also realizes combining all three papers would be LONG, and 8-page limit of NeurIPS submission will not allow all major results to be presented within the main text.

Q: *Detailed comments*

A: It is disappointing that the reviewer's detailed comments posted here are not for this paper 1410, but for paper 2460.

Reviewer 1: **Q1:** *Can similar analysis be applied to handle sampling without replacement and obtain better complexity?*

A: Yes, the analysis here can be applied to sampling without replacement, but need to incorporate different concentration bounds due to the sampling without replacement. We expect the resulting bounds to be tighter.

Q2: *The final output is randomly selected among all past iterations. Does this require to store all generated variables?*

A: The implementation of the algorithm does not need to store all variables. Our theory suggests that we can randomly pick $i \in 1, \cdots K$ first, and then stop at the i-th iteration. In practice, we can store only the variable with minimum batch gradient norm. Alternatively, outputting the last iteration also performs well in practice.

Reviewer 4: **Q1:** *Step size 1/2L is enabled due to big minibatch $\sqrt{n}$, which could be a problem for its practical usage.*

A: We agree. This is also the reason why we study online version (see appendix), which requires the minibatch $\mathcal{O}(\epsilon^{-1})$ and total sample complexity of $\mathcal{O}(\epsilon^{-3})$ to achieve the same accuracy. This helps for large n regime.

Q2: *Experimental results are not appealing since the parameters are chosen neither based on theory nor any tuning. Moreover, there is no comparison to SAGA algorithm, which would be also a natural choice.*

A: Thanks for the suggestion! We will add more experiments as suggested.

Q3, Q4, Q5: *presentation suggestions*

A: Thanks for the suggestions! We will revise the paper accordingly.