Response to the reviewers

We thank all reviewers (Reviewer 1, Reviewer 2, and Reviewer 3) for their constructive feedback. Reviewer 1 and Reviewer 2 asked if the proposed analysis could be extended and compared to the state-of-the-art variance-reduced methods such as recursive methods. Additionally, Reviewer 3 raised concerns about the minibatch sizes that can achieve optimal SZO calls. We shall focus on discussing these concerns in the following and address the individual reviewer questions.

Reviewer 1

Reviewer Point P 1.1 — The authors claim that since the gradient is no longer unbiased, techniques used in [2, 3] are inapplicable. Consequently, it takes some effort to adapt techniques from ProxSVRG to the zeroth-order setting. This is true but it I am confused whether the analysis in ZO-PSVRG+ is the first to deal with biased estimators or the idea is similar to existing techniques. I believe the techniques in methods using biased estimators such as [5, 1] may still be useful. Especially using the intuition that the stochastic gradient does not have to be unbiased, as long as its second moment can be bounded then we can still achieve convergence. Therefore, the claim may need to be revised to avoid confusion.

Reply: As the reviewer pointed out, in our convergence analysis $\mathbb{E}_{I_b}[\hat{v}_{t-1}^s] = \hat{\nabla} f(x_{t-1}^s) \neq \nabla f(x_{t-1}^s)$, which shows that the zeroth-order stochastic gradient is a biased estimate of the true full-gradient. Nevertheless, the convergence studies on the first-order variance-reduction methods such as [Reddi et al., 2016, Li and Li, 2018, Fang et al., 2018] rely upon the assumption that the stochastic gradient is an unbiased estimate of the full first-order gradient, e.g., [Reddi et al., 2016, Lemma 7] or [Fang et al., 2018, Eq. (2.3)]. Thus, adopting similar ideas from first-order proximal variance-reduced methods for the nonsmooth composite optimization problem to the zeroth-order optimization is not straightforward. To address the hindrance in adopting the first-order results to the zeroth-order formulation, we analyze the upper bound on the point-wise difference of the estimation of variance-reduced gradient $\hat{v}_t^s$ and the first-order gradient over the batch of data of size $\mathcal{B}$, in Lemmas 15, 16, 17 and 18 in the appendix. We further choose the appropriate stepsize η and smoothing parameter μ in the theorems to control these error bounds. Similar techniques have been utilized in the convergence analysis presented in [Ghadimi et al., 2016, Liu et al., 2018, Ji et al., 2019, Huang et al., 2019].

Reviewer Point P 1.2 — The convergence rate of ZO-PSVRG+ has the dependence on the mini-batch size, it is not clear how better it is compared to results in [4]. I suggest the authors to provide specific choice of b to better clarify that ZO-PSVRG+ also achieves the best-known complexities. Can the best-known rate be achieve with any choice of b or only a specific order of b, this should be better clarified to emphasize the contribution of the paper.

Reply: We listed our results in Tables 1 and 2 with some recommended minibatch sizes which achieve the optimum SZO queries. The results in these tables are stated in terms of the convergence rate based on the number of iterations T and stochastic zeroth-order oracle (SZO) calls. In Figure 1 we also have shown SZO complexity in terms of minibatch sizes. From the tables it is noted that [Huang et al., 2019] and [Ji et al., 2019] achieved their optimal convergence results with $b = n^{2/3}$, which match with the best results from ZO-PSVRG+ using $b = 1/\epsilon^{2/3}$ (see Figure 1). In [Ji et al., 2019] and [Huang et al., 2019] the analysis for ZO variance-reduced approaches rely on $b \leq n^{2/3}$, while our convergence studies are valid for $b \leq n$. Our analysis also includes the cases when the full gradient is computed on a batch of size $\mathcal{B}$ from randomly selected samples

with $\mathcal{B} < n$. Therefore, by selecting $\mathcal{B} = O(\frac{1}{\epsilon}) < n$ in ZO-PSVRG+, the optimum SZO calls is obtained when $b = \epsilon^{2/3}$. Therefore, to achieve the optimal rate of convergence for ZO-PSVRG+ based on our convergence studies a moderate minibatch size $b = \epsilon^{2/3}$ versus $b = n^{2/3}$ can be chosen, which is not too small to forfeit parallelism and furthermore is not too large for the better utilization of variance-reduced approaches in practice. Note that from Figure 1 the curves of ZO-PSVRG+ overlap for $s_n = \min\{\frac{1}{\epsilon}, n\}$ for $b \geq n^{2/3}$. It is noticed from Figure 1 that our analysis for ZO-PSVRG+ yields optimal convergence for all minibatch sizes $b \leq n^{2/3}$ or $b \leq \frac{1}{\epsilon^{2/3}}$, while the convergence results in [Ji et al., 2019] only shows optimal convergence for $b = 1, \frac{1}{\epsilon^{2/3}}$ based on an involved and complicated parameter information. Similarly, for ZO-PSPIDER+, we achieve the best convergence result for $\frac{1}{\epsilon^{1/2}}$, which matches with [Ji et al., 2019] when $\mathcal{B} = n$. We derive the convergence behavior of studied variance-reduced for all the spectrum of minibatch sizes $b \in [1, n]$ as seen in Figure 1. Our analysis shows ZO-PSPIDER+ can obtain superior optimal SZO calls of $\min\{\frac{n^{1/2}}{\epsilon}, \frac{1}{\epsilon^{3/2}}\}$ for the entire $b \leq s_n^{1/2}$ where $s_n = \min\{n, \frac{1}{\epsilon}\}$, with the compensation in the order of stepsize η for the smaller minibatch sizes.

Reviewer Point P 1.3 — As the stochastic gradient estimator is biased anyway, I wonder if using the scheme as in [5, 1] can further improves the complexity results and in the first-order regime.

Reply: In the revised paper, we extended our analysis to the recursive type methods called ZO-PSPIDER+. Our novel analysis brings the convergence study of ZO-PSVRG+ and ZO-SPIDER+ into uniformity. In Table 2 we have shown the convergence rates along with the SZO calls. The results show that using recursive methods to estimate the gradient at the inner-loop of variance-reduced algorithms can improve the optimal SZO calls for ZO-PSVRG+ from $O(\min\{\frac{n}{\epsilon^{2/3}}, \frac{1}{\epsilon^{5/3}}\})$ to $O(\min\{\frac{n^{1/2}}{\epsilon}, \frac{1}{\epsilon^{3/2}}\})$ for ZO-PSPIDER+. In our analysis the suboptimal rate of ZO-PSVRG+ stems from the error estimation of v_t^s being bounded by $O(\frac{\left\|x_s^t - \tilde{x}_{s-1}\right\|}{b})$, while this term is controlled by $\left\|x_t^s - x_{t-1}^s\right\|^2$. Converting the expression $\left\|x_t^s - x_{t-1}^s\right\|^2$ to $\|x_t^s - \tilde{x}_{s-1}\|^2$ introduces suboptimal terms of order $O(\frac{1}{t^2})$ that has to be compensated by a smaller stepsize $\eta \sim O(\frac{\sqrt{b}}{m})$, which is suboptimal. An amendment to ZO-PSVRG+ analyses to obtain the superior optimal rate $O(\min\{\frac{n^{1/2}}{\epsilon}, \frac{1}{\epsilon^{3/2}}\})$ is to use more accurate gradient estimation v_t^s such that the estimation error of gradient decreases by $O(\frac{1}{b^2})$ or $O(\frac{1}{bm})$. Alternatively, if we assume the the deviation of x_t^s from $\tilde{x}^{s-1}$ is not smaller than the deviation from x_1^s for all $t \leq m$ as t increases at the inner-loop of variance-reduced algorithm, i.e.,

$$\mathbb{E}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 \geq \mathbb{E}\left\|x_1^s - \tilde{x}^{s-1}\right\|^2$$

due to $\sum_{t=1}^{m} \frac{1}{t^2} < \infty$ in our novel analyses, the superior rate for SZO calls for ZO-PSVRG+ could be proven.

References:

[1] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. Advances in Neural Information Processing Systems, 31, 2018.

[2] Sashank J Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. Advances in neural information processing systems, 29, 2016.

[3] Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. Advances in neural information processing systems, 31, 2018.

[4] Sijia Liu, Jie Chen, Pin-Yu Chen, and Alfred Hero. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. In International Conference on Artificial Intelligence and Statistics, pages 288–297. PMLR, 2018.

[5] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In International Conference on Machine Learning, pages 2613–2621. PMLR, 2017.

Reviewer 2

Reviewer Point P 2.1 — Note that the two-point function value zeroth-order estimators RandSGE and CoordSGE are standard and the authors directly used their corresponding properties from previous work (Liu et al, 2018; Ji et al, 2019).

Reply: We use the standard zeroth-order estimators similar to [Huang et al., 2019, Ji et al., 2019, Fang et al., 2018, Liu et al., 2018]. In [Ji et al., 2019], the authors proved a tighter error term for RandSGE gradient estimation compared to [Liu et al., 2018], which we used in our analysis to show an improved convergence rate when zeroth-order gradient at the inner-loop of variance-reduced algorithm is using a two-point random gradient estimate (RandSGE).

Reviewer Point P 2.2 — For the results, they did not compare with existing state-of-the-art (SOTA) results which used recursive gradient estimators such as SARAH (Nguyen et al., 2017) and SPIDER (Fang et al., 2018).

Reply: In the revised paper we extended our analysis to the recursive-type zeroth-order variance-reduced algorithms. Our analysis provides a uniform convergence analysis for ZO-PSVRG+ and ZO-PSPIDER+. Differently from ZO-PSVRG+, in ZO-PSPIDER+ at inner-loop iterations, we recursively utilize the latest gradients to update the gradient estimator. Similar to [Ji et al., 2019], our convergence analysis removes the Gaussian random generation requirement in [Fang et al., 2018] and improves the stepsize from $O(\epsilon)$ to $O(1)$ while achieving the same superior optimal SZO calls as listed in Table 2. Our analysis shows an improved SZO query for ZO-PSPIDER+ compared to ZO-PSVRG+ for all $b \leq n^{1/2}$, while in [Ji et al., 2019] the improvement for SPIDER-type algorithms is only shown for $b = 1, n^{1/2}$. We also show that the optimal convergence rate can be achieved for moderate minibatch sizes of $b \leq \frac{1}{\epsilon^{1/2}}$ when $s_n = \frac{1}{\epsilon} < n$ (see Figure 1). We deferred the analysis for ZO-PSPIDER+ (RandSGE) to the appendix.

Reviewer Point P 2.3 — According to their Table 1 or theorems, the best results for their ZO-PSVRG+ is $O(\min\{n^{2/3}d/eps, d/eps^{5/3}\})$ which is much worse than SOTA result $O(\min\{n^{1/2}d/eps, d/eps^{3/2}\})$ (see e.g., SPIDER-SZO (Fang et al., 2018), ZO-SPIDER-Coord (Ji et al., 2019)). The gap in this zeroth-order case is the same as SVRG vs. SPIDER in the first-order situation.

Reply: Now our analysis in Table 2 for ZO-PSPIDER+ shows the superior optimal SZO calls $O(\min\{\frac{1}{\epsilon^{3/2}}, \frac{n^{1/2}}{\epsilon}\})$ for $b \leq \min\{\frac{1}{\epsilon^{1/2}}, n^{1/2}\}$. Our analysis also characterizes several sufficient conditions for ZO-PSVRG+ to achieve the optimal SZO calls similar to ZO-PSPIDER+. For example, if we assume the the deviation of x_t^s from $\tilde{x}^{s-1}$ is not smaller than the deviation from x_1^s for all $t \leq m$ as t increases at the inner-loop of variance-reduced algorithm, i.e.,

$$\mathbb{E}\left\|x_t^s - \tilde{x}^{s-1}\right\|^2 \geq \mathbb{E}\left\|x_1^s - \tilde{x}^{s-1}\right\|^2$$

due to $\sum_{t=1}^{m} \frac{1}{t^2} < \infty$ in our analysis, the superior rate for SZO calls for ZO-PSVRG+ could be proven.

Reviewer Point P 2.4 — This paper claimed Lemmas 9 and 10 as their results, not pointing to previous work.

Reply: The referred lemmas are classical results in the proximal gradient descent analysis and the proofs are similar to [Reddi et al., 2016, Lemma 1, Lemma 2], [Ghadimi and Lan, 2013, Lemma 1, Proposition 1] and [Li and Li, 2018, Lemma 1, Lemma 2]. We added the proofs of these lemmas for completeness and we clarified it in the revised paper.

Reviewer Point P 2.5 — In the proof of Lemma 12, the Eq (S27) should be obtained using Lemma 16 rather than Lemma 15, since here Eq (S27) is the RandSGE case not like Eq (S17) in the CoordSGE case.

Reply: The equation (S27) is related to the gradient estimation at the outer-loop of variance-reduced algorithm, i.e., full-gradient over all the samples, which is computed coordinate-wisely for the studied ZO algorithms in the paper.

Reviewer Point P 2.6 — Regarding the properties of CoordSGE, i.e., Lemma 15. It comes from (Liu et al., NeurIPS'18) not (Liu et al., AISTATS'18). The citation is wrong. They are two different papers. Liu et al. (2018b) cited in this submission is "Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. In AISTATS'18". However, it should be "Zeroth-order stochastic variance reduction for nonconvex optimization. In NeurIPS'18".

Reply: Fixed.

Reviewer Point P 2.7 — Typo: Item 2 of Lemma 16, the second term $\|f_\mu(x) - f(x)\|$ should be $\|\nabla f_\mu(x) - \nabla f(x)\|$.

Reply: Fixed. This result corresponds to [Gao et al., 2018, Eq. 71] which provides a tighter bound than that of the Gaussian smoothing scheme in [Nesterov and Spokoiny, 2017].

Reviewer 3

Reviewer Point P 3.1 — The authors need some specific choices of parameters such as batch size b, epoch length m. The developed analysis here seems to be more flexible. This is good.

Reply: We updated Tables 1 and 2 where we provided some recommended minibatch sizes that achieve optimal SZO calls for each scheme. Our results are consistent and more general compared with the optimal SZO queries in [Liu et al., 2018, Ji et al., 2019, Huang et al., 2019] with improved order of stepsizes. Figure 1 summarizes our results on SZO queries in terms of minibatch sizes b for $b \in [1, n]$. Please note that the analysis for ZO-SVRG in the aforesaid research works is limited to $b \leq n^{2/3}$. Our convergence study also covers $s_n \neq n$ when $\mathcal{B} = \frac{1}{\epsilon} < n$ which advocates moderate minibatch sizes in the optimal case, i.e., $b = \frac{1}{\epsilon^{2/3}}$, $b = \frac{1}{\epsilon^{1/2}}$ versus $b = n^{2/3}$, $b = n^{1/2}$ for ZO-PSVRG+ and ZO-PSPIDER+, respectively.

Reviewer Point P 3.2 — Based on my reading, it is not very clear to me that what kinds of new analysis is developed to achieve the generality of the hyperparameter selection. I suggest that the authors can add some sentences or a paragraph to emphasize such technical novelties.

Reply: Our convergence analysis for ZO-PSVRG+ is remarkably different from the convergence studies in [Liu et al., 2018, Ji et al., 2019, Huang et al., 2019]. In particular, in the proof of [Ji et al., 2019, Theorem 1], inspired from [Reddi et al., 2016], a Lyapunov function $\Psi_t^s = \mathbb{E}[F(x_t^s) + c_t \left\|x_t^s - \tilde{x}^{s-1}\right\|]$ is defined and it was shown that Ψ_t^s is a descending sequence at each inner-loop of the variance-reduced algorithm by the sum of gradient mapping $\sum_{t=1}^m \|g_\eta(x_k^s)\|$. However, in our analysis, we directly show the descent equation (S34) for $F(x_t^s)$ which is exploited in the subsequent theorems to prove the convergence of gradient mapping $\mathbb{E}[\|g_\eta(\hat{x})\|]$ for nonconvex functions. For this purpose, we compute the error estimate of the point-wise difference of the first-order full-gradient (the gradient over the entire batch of samples) and the zeroth-order estimation of the gradient at each iteration of the inner-loop of the variance-reduced algorithm, i.e., $\mathbb{E}\left[\left\|\nabla f(x_{t-1}^s) - \hat{v}_{t-1}^s\right\|^2\right]$ in Lemmas 15, 16, 17, and 18. In this manner, we provide a uniform convergence analysis for all the studied variance-reduced algorithms. While our analysis is valid for any minibatch sizes $b \in [1, n]$, the analysis in [Liu et al., 2018, Huang et al., 2019, Ji et al., 2019] rely on the minibatch sizes $b \leq n^{2/3}$. In our proposed convergence results, the parameters of the variance-reduced algorithms are specified explicitly with straightforward equations, however, the parameter expressions and parameter selection information in the referred papers listed in Table 1 are involved and are dependent on the implicit set of equations (for example see [Ji et al., 2019, Theorem 1]).

Reviewer Point P 3.3 — It would be better if more experiments, e.g., in some adversarial attack applications, can be provided.

Reply: We extended our experiments to ZO-PSPIDER+ and provided an additional empirical study on generating universal black-box attacks over CIFAR-10 dataset. We also visualized the perturbation and eventual black-box attacks to compare the patterns of adversarial distortions across different ZO algorithms.

References

[Fang et al., 2018] Fang, C., Li, C. J., Lin, Z., and Zhang, T. (2018). Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699.

[Gao et al., 2018] Gao, X., Jiang, B., and Zhang, S. (2018). On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing*, 76(1):327–363.

[Ghadimi and Lan, 2013] Ghadimi, S. and Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

[Ghadimi et al., 2016] Ghadimi, S., Lan, G., and Zhang, H. (2016). Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305.

[Huang et al., 2019] Huang, F., Gu, B., Huo, Z., Chen, S., and Huang, H. (2019). Faster gradient-free proximal stochastic methods for nonconvex nonsmooth optimization. In *AAAI*.

[Ji et al., 2019] Ji, K., Wang, Z., Zhou, Y., and Liang, Y. (2019). Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International Conference on Machine Learning*, pages 3100–3109.

[Li and Li, 2018] Li, Z. and Li, J. (2018). A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 5564–5574.

[Liu et al., 2018] Liu, S., Kailkhura, B., Chen, P.-Y., Ting, P., Chang, S., and Amini, L. (2018). Zeroth-order stochastic variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3727–3737.

[Nesterov and Spokoiny, 2017] Nesterov, Y. and Spokoiny, V. (2017). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566.

[Reddi et al., 2016] Reddi, S. J., Sra, S., Póczos, B., and Smola, A. J. (2016). Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, pages 1145–1153.