We thank all the reviewers for their valuable feedback. We appreciate the typos, minor errors, stylistic suggestions and unclear math steps pointed out by reviewers, and we will update our paper accordingly.We address the main comments from reviewers (abbreviated)

Reviewer_1

Q1: Unfair comparison with Wang 2017a because of smoothness assumption A1: For earlier works (such as Wang 2017a) dealing with non-smooth cases, the algorithm using its techniques share the same rate with the smooth case itself. That being said, we will add columns to Table 1 to highlight the assumption difference.

Q2: mixing ∂g with ∇g A2: Thanks for pointing that out. We decide to keep this notation ∂g for two reasons: (1) the notation has been adopted by a few earlier works (Lin et al 2018, Liu et al 2017) for smooth cases, where the "partial" notation is simply adopted to highlight that it is a (Jacobian) matrix; (2) our work is not addressing the non-smooth case, so there is no confusion with sub-gradient in the non-smooth case.

Q3: Assumptions not satisfied for equation (4) A3: First of all, (4) has a minor typo: the second sum should have its bracket squared. Although the derivative is not bounded, the differentiable functions f and g does admit bounded Lipschitz derivatives in the domain of optimization (a domain that contains the initializer, optimizer, and the entire path), and their compositional function also has Lipschitz derivatives. Our Assumptions 2, 3, 4 and 9 of Lipschitz class of functions are very similar to those in standard literature on compositional optimization (Lin et al 2018, Liu et al 2017).

Q4: Mini-batch assumption of order $\mathcal{O}(1/(\epsilon^2))$ is impractical A4: Different from $A_2 = B_2 = C_2 = 1$, the large-batch step A_1, B_1, C_1 occurs only once every ϵ^{-2} steps, so the IFO complexity remains unchanged. These large-batch steps helps reduce the variance of gradient estimation to $\mathcal{O}(\epsilon)$, and is very crucial for SARAH-type variance reduction of not accumulating noises.It is a selection which would allow the optimal theoretical guarantees (by central limit theorem, a $\mathcal{O}(1/(\epsilon^2))$ mini-batch every q steps would give ϵ-accurate estimator in the beginning).

Q5: High number of assumptions and thus no new insight to help researchers A5: The theoretical guarantees are valuable since they provide a state-of-the-art, and is believed to be optimal (although there is no lower bound result yet). As mentioned earlier, all assumptions are in standard literature and they see valuable applications in portfolio management, reinforcement learning, dimension reduction, etc.

Reviewer_2

The reviewer is extremely careful in checking our proof. We appreciate that a lot. And we have fixed all the typos accordingly.

Q1: Misleading use of the $\|\cdot\|$ as the Frobenius norm A1: Thanks for your comment. Indeed, all norms involving the SARAH estimator for Jacobian matrices adopts a Frobenius norm. The complexity results still hold after a simple fix. Previous literatures also admit this "caveat" because there is potentially a gap between the Frobenius and (its lower-bounded) operator norms, this would not lead to any disadvantage of complexities.

Q2: Experimental parts should be discribed in the body. A2: We've done some experiments on this problem to validate our theory before the review period. We did more careful experimental setting and testing on the three applications mentioned: portfolio management, reinforcement learning, and t-SNE after then. We would position part of the experiments in the body part instead of the proof in our next submission.

Q3: Inconsistent notation A3: Thanks for checking in details. F_ξ and f_ξ, M_g, B_g and B_G are indeed the same thing. We seperately used f_i and F_ξ before, which leads to duplicated constants. We already fixed them in our follow-up version.

Q4: Optimality claims are vague A4: We appreciate your preciseness. Our IFO complexity is state-of-the-art compared with previous literature. We will state our contribution in a more rigorous fashion.

Q5: L_Φ is not necessary. Step size eta can be simplified A5: We agree and already made this fix in our revision. Thank you.

Q6: The proof of Thm 10 only relies on the formula for the variance of the average of iid vector random variable A6: Yes, we corrected the statement of Theorem 10. Thanks for pointing that out

Reviewer_3

Q1: Novelty is limited A1: We would like to argue our paper is not an incremental one. We believe that using variance reduced gradient methods (SVRG, SPIDER, among others) can be a potential alternative to existing stochastic compositional optimization methods, which includes and extends the current framework of SGD and SCGD. We aimed to provide an "optimal" nonconvex analysis and hence augment the current theoretical framework of this problem. We will try to polish more of this work in the next round of submission.

Q2: Lacking numerical results A2: Our work mainly forcuses on proving that the convergence rates are sharper than all other existing convergence rates currently available for the compositional optimization problem. This demonstrates the power of using the SPIDER estimator to trace quantities needed. This work is mainly a theoretical extension along the directions pointed by the SPIDER paper. We did several experiments on three traditional applications in the field of compositional optimization problems. Our main body of the paper mainly focuses on proposing an optimal theoretical guarantee. See also A2 of **Reviewer_2**