We thank all reviewers for their detailed comments and acknowledgment of our contribution. We first address the common question on the relation to prior work and then respond to each reviewer.

Related Work: Below we briefly point out the key differences from related work. As reviewers suggested, we will dedicate a section for related work and discuss in greater details in our revision.

- **Prior work on SCO**: [37, 38] and [Yang et al, 2019] (mentioned by Reviewer 3) considered SCO problem: $\min_{x\in\mathcal{X}} \mathbf{f}\circ\mathbf{g}(x) := \mathbb{E}_\xi\big[f_\xi\big(\mathbb{E}_\eta[g_\eta(x)]\big)\big]$ and its T-level composition extension. They assumed that there exists a *deterministic function* $\mathbf{g}(x)$, independent of ξ, such that $\mathbf{g}(x) := \mathbb{E}_\eta[g_\eta(x)|\xi]$. This does not hold for general CSO, and their algorithms and analysis do not work for CSO.
- **Prior work on estimating conditional expectations:** [Giles et al., 2019 (mentioned by Reviewer 2), Hong and Sandeep, 2009] focuses on pure estimation, whereas we deal with optimization, which is more challenging.
- **Prior work on biased gradient methods:** [23, Karimi et al., 2019 (mentioned by Reviewer 4), Hu et al., 2020, Chen and Ronny, 2018]. These work focuses on general biased oracles, whereas in our work, the bias comes from estimating the conditional expectation and is controllable by the inner sample size m and the smoothness condition.

Response to Reviewer 1: Thanks for the careful reading. We will add the result in [12] in Table 1 and fix other issues.

Response to Reviewer 2: Thanks for the useful comments. In our humble opinion, *all issues brought up by the reviewer are relatively minor and can be easily clarified.*We hope that our reply addresses the reviewer's concerns.

Justification of assumptions: We impose the assumption on F to keep it general and will present conditions on f and g such that the assumption is satisfied. For instance, F is convex when (i) f_ξ is convex and g_η is linear; or (ii) f_ξ and g_η are convex and f_ξ is non-decreasing. Strong convexity of F holds when l_2-regularization is added to convex objectives. Weak convexity of F holds when (i) f_ξ and g_η are both Lipschitz continuous and smooth, or (ii) f_ξ is convex and g_η is smooth [14]. Restricting to specific conditions on f_ξ and g_η may limit the generality of our results.

Assumption 1(b): The reviewer is correct that the boundedness of gradient estimator can be derived (instead of assumed) under certain conditions. For instance, (i) when f_ξ and g_η are Lipschitz continuous, Assumption 1(b) holds with $\beta = 0$, $M^2 = L_f^2 L_g^2$; (ii) when F and $\hat{F}$ are μ-strongly convex and S-smooth, $\mathcal{X} = \mathbb{R}^d$, Assumption 1(b) holds with $\beta = 2S^2/\mu^2$, $M^2 = 2\mathbb{E}\|\nabla\hat{F}(x^*)\|_2^2$ [Nguyen et al., 2018]. This is because $\mathbb{E}\|\nabla\hat{F}(x)\|_2^2 \le 2\mathbb{E}\|\nabla\hat{F}(x) - \nabla\hat{F}(x^*)\|_2^2 + 2\mathbb{E}\|\nabla\hat{F}(x^*)\|_2^2 \le 2S^2\|x - x^*\|_2^2 + 2\mathbb{E}\|\nabla\hat{F}(x^*)\|_2^2 \le 2S^2/\mu^2\|\nabla F(x)\|_2^2 + 2\mathbb{E}\|\nabla\hat{F}(x^*)\|_2^2$. where the second inequality modifies Thm 1 in [Johnson and Zhang, 2013]. However, beyond these situations, the boundedness of gradient estimator may not automatically hold. We will clarify this assumption in the paper.

Notation on derivative and typo in Line 101: We will clarify that ∇ is abusively used to denote the Jacobian matrix, gradient vector, and derivative for simplicity. We apologize for the typo in Line 101. The correct equation is $\nabla F(x) = \mathbb{E}_\xi\big[(\mathbb{E}_{\eta|\xi}\nabla g_\eta(x,\xi)])^\top \nabla f_\xi(\mathbb{E}_{\eta|\xi} g_\eta(x,\xi))\big]$. Both issues do not affect our analysis and contribution.

Effects of m in Theorems: We explicitly define the bias term $\Delta_f(m)$ in Eq. (3) between Lines 130 and 131 and use it in Theorems 2, 3, and 4 to show the dependence of the upper bounds on the inner sample size m. The bias term is $\mathcal{O}(1/m)$ when f_ξ is Lipschitz smooth and $\mathcal{O}(1/\sqrt{m})$ if f_ξ is Lipschitz continuous.

Response to Reviewer 3: CSO is not a special case of [Yang et al., 2019] when $T = 2$. Please refer to the above discussion on Related Work. We compare with SAA [24] in Table 1 as both results of BSGD and SAA characterize the sample complexity of finding an ϵ-optimal solution of the original objective F. However, they require different computation complexities since SAA requires solving the empirical problem.

Additional comments: (2) As marked in the footnote of Table 1, there is a difference in the lower bound when $\hat{F}$ is Lipschitz continuous or Lipschitz smooth. (3) Both lower bound definitions are valid (see [31] for discussion on their relationship); note that the first definition fixes the iteration T while the second fixes the inaccuracy ϵ. (5): The solution to SAA is obtained by CVXPY solver.

Response to Reviewer 4: We will discuss the technical novelty of our lower bound proof and fix other issues.
Results not surprising: While the algorithmic idea of BSGD and analysis seem intuitive, these have never been studied before in the context of CSO problems, which have various important applications in emerging fields like MAML. We provide a comprehensive analysis of stochastic first-order methods for CSO in convex and nonconvex regimes, with both upper and lower bounds. We do find several results quite interesting (to some extent, surprising):

- The smoothness of the outer function plays an important role in the total sample complexity of CSO.
- Simple algorithm such as the BSGD already attains near-optimal complexities for solving CSO, which is quite different from recent results on nested stochastic optimization (requires sophisticated algorithms for different settings).
- Our lower bounds in nonconvex setting build on a novel construction of a worst-case biased gradient estimator by perturbing the last coordinate; this may look straightforward in hindsight. Nonetheless, these results seem new.
- Albeit its simplicity, our numerical study shows that BSGD performs quite well on MAML tasks for meta-learning.