Reviewer 1: Unclear about the evaluation for outer iterations; Does the number of aggregated tasks affect convergence: *Great question! Yes, the total complexity is proportional to the number of aggregated tasks. In addition, in terms of updating task-specific parameters, ANIL takes the same steps as MAML, and the outer-loop gradient (line* 10 *of Alg.* 1*) also depends on the inner-loop outputs* $w^i_{k,N}$ *of tasks in* $\mathcal{B}_k$*. We will clarify it in the revision.*

Add experiments to compare ANIL and MAML and w.r.t. the size B of samples: *Thanks for the suggestion! We will absolutely follow these suggestions to add experiments in the revision.*

Why sample size in inner-loop is not taken into analysis, as Fallah et al. [4] does: *Great question! In our setting, the inner-loop loss functions take a finite-sum form over pre-assigned samples. As a result, the inner-loop updates take full gradient descent without data sampling, and hence gradient estimation bias (which can introduce sample size) does not exist in convergence bound. This setting has also been considered in Rajeswaran et al. [24], Ji et al. [13]. As a comparison, Fallah et al. [4] considered a different setting, where loss functions take the form in expectation and fresh data are sampled as the algorithm runs. As a result, their analysis involves an estimation bias, which introduces the dependence on the number of samples.*

Experiments for non-convex and strongly convex cases with the same stepsize: *Great point! We have run more experiments on FC100 with the same stepsize* $0.03, 0.05, 0.1$ *for both cases, and the nature of results remain the same.*

Elaborate more for line 170: *The statement specifically refers to Theorem 1, where increasing* N *leads to larger stepsize* β_w*, which yields faster convergence rate* $\mathcal{O}(\frac{1}{K\beta_w})$*. We will clarify it in the revision.*

Reviewer 2: Dependence on κ. iMAML depends on $\sqrt{\kappa}$ in contrast to poly(κ) of this work: *Great question! High-level speaking, better dependence on* κ *for iMAML is based on an ideal solution of an inner-loop optimization problem, which can take many iterations. ANIL takes only a few inner-loop iterations (thus a lower cost), but has worse outer-loop convergence (in terms of* κ*). Technically speaking, smoothness analysis of iMAML upper-bounds the distance between two optimal points* $w^i_*(w_1)$ *and* $w^i_*(w_2)$*, each obtained by solving an inner-loop optimization problem. As a comparison, analysis of ANIL upper-bounds the distance between two inner-loop paths, which sums up the distances between all corresponding points on the two paths (see eq.* (21)*). This results in a worse dependence in* κ*.*

Add an experiment to verify the tightness: *Great point! We will definitely add such an experiment in the revision.*

Extra assumption on Lipschitzness of the objective, which is not for iMAML: *We take this assumption to ensure the meta gradient to be bounded. As a comparison, iMAML alternatively assumes the search space of parameters to be bounded (see Theorem 1 therein) so that the meta gradient (eq. (5) therein) can be bounded.*

The role of N in the theory seems to make convergence only slower: *The exponential term has a worse dependence on constants and* τ, M *than the linear term (we will add explicit forms in the revision), and hence the choice of* N *depends on how large* κ *is. For large* κ*, as the reviewer also pointed out, a small* $N = 2$ *is a better choice. However, when* κ *is not very large, e.g., in our experiments (in which increasing* N *accelerates the iteration rate), the exponential term dominates for a small* N*, and hence a larger* N *is preferred. We will clarify it in the revision.*

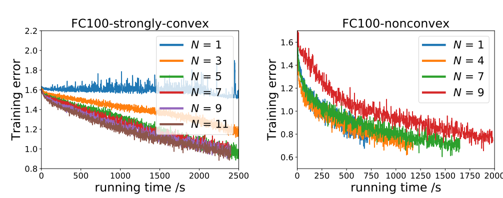
Optimality of $N = 1$ contradicts the experiments where $N = 4, 7$ are the best: *We assume the reviewer refers to the experiments in left plot of Figure 2(a). This can be due to the fact that the influence of* N *w.r.t. the number of outer-loop iterations is offset by other constant-level parameters for small* N*. Evidently, right plot of Figure 2(a) indicates that* $N = 1$ *is optimal w.r.t. the running time, which agrees with our result on computational complexity.*

Suggestions on presentation and references: *Many thanks! We will follow these suggestions to improve our paper.*

Reviewer 3: *We thank the reviewer for the positive comments!*

Reviewer 4: Comments on insight of theoretical results: *Our results theoretically characterize the order-level computational complexity for ANIL and its comparison to MAML. In addition, our analysis techniques can be useful for developing guarantee for other meta-learning and more broadly bi-level optimization algorithms.*

Convergence analysis is done with vanilla gradient descent but all experiments are done with Adam; Experiments with purely first-order methods: *Great point! We have done new experiments on FC100 dataset using mini-batch SGD with a learning rate of* 0.05*, and the results are given in the figures to the right. It can be seen that the nature of the results remains the same as those in our paper. More results will be added in the revision.*



Run experiments over different random seeds and over different hyper-parameter settings: *Many thanks! We will definitely provide these experimental results in the revision.*