General Response. We thank all the reviewers for their insightful and encouraging comments. Below we provide our point-by-point response to the main concerns raised by the reviewers.

To Reviewer #1. Per your suggestion, we will update the appendix by adding more explanations about the proof ideas.

To Reviewer #2.

1) Since in many real applications, e.g. image classification, the task number n is finite though could be large, w.l.o.g. we choose to focus on the finite setting (**FS**). But all the convergence and generalization guarantees in this work can be extended to the infinite setting (**IFS**) which will be emphasized in revision. We briefly introduce the idea of extension from FS to IFS. **For convergence**, the technical Lemmas 1 $\sim$ 4 hold for both settings as they do not involve FS and IFS. Let $\phi_{D_{T_i}}(\boldsymbol{w}) = \min_{\boldsymbol{w}_{T_i}} \mathcal{L}_{D_{T_i}}(\boldsymbol{w}_{T_i}) + \frac{\lambda}{2}\|\boldsymbol{w}_{T_i} - \boldsymbol{w}\|_2^2$ and $\boldsymbol{w}_{T_i}^* = \text{argmin}_{\boldsymbol{w}_{T_i}} \mathcal{L}_{D_{T_i}}(\boldsymbol{w}_{T_i}) + \frac{\lambda}{2}\|\boldsymbol{w}_{T_i} - \boldsymbol{w}\|_2^2$. Extending Theorem 1 from FS to IFS only needs to extend (a) $\mathbb{E}[\frac{1}{b_s}\sum_{i=1}^{b_s}\phi_{D_{T_i}}(\boldsymbol{w})] = F(\boldsymbol{w})$ and (b) $\mathbb{E}[\frac{1}{b_s}\sum_{i=1}^{b_s}\nabla\phi_{D_{T_i}}(\boldsymbol{w})] = \nabla F(\boldsymbol{w})$ with $F(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}\phi_{D_{T_i}}(\boldsymbol{w})$ under FS respectively to (a) and (b) with $F(\boldsymbol{w}) = \mathbb{E}_{T\sim\mathcal{T}}\phi_{D_T}(\boldsymbol{w})$ for IFS. By sampling mini-batch $\{T_i\}$ as $T_i \sim \mathcal{T}$, (a) and (b) hold for IFS. As tasks T_i, e.g. in image classification, are usually from a uniform distribution $\mathcal{T}$, we can uniformly sample task T_i. The remaining proofs for IFS and FS are identical. Similarly, we can extend convergence results in Theorem 4 in Appendix from FS to IFS. **For generalization**, Theorems 2 and 3 still hold for IFS without needing any changes, as they provide generalization performance guarantee of empirical solution in any task $T \sim \mathcal{T}$.

When task number n is fairly small, we agree that it is an interesting future work to explore the structure of task space, e.g. hierarchical structure. We expect that the approach developed in this paper will fuel this future investigation.

2) One advantage of MMP over MAML is that it can easily and flexibly consider the structures of solution space of tasks by designing proper $\|\boldsymbol{w}_{T_i} - \boldsymbol{w}\|_p^q$ so as to find better $\boldsymbol{w}_{T_i} = \text{argmin}_{\boldsymbol{w}_{T_i}} \mathcal{L}_{D_{T_i}}(\boldsymbol{w}_{T_i}) + \frac{\lambda}{2}\|\boldsymbol{w}_{T_i} - \boldsymbol{w}\|_p^q$ and thus better prior $\boldsymbol{w}$. In contrast, MAML uses a fixed gradient descent update rule $\boldsymbol{w}_{T_i} = \boldsymbol{w} - \eta\nabla\mathcal{L}_{D_T}(\boldsymbol{w})$ according to its model $\mathcal{L}_{D_T}(\boldsymbol{w} - \eta\nabla\mathcal{L}_{D_T}(\boldsymbol{w}))$, hampering designing more flexible relation between $\boldsymbol{w}_{T_i}$ and $\boldsymbol{w}$. For instance, assume there are a few outlier tasks $\mathcal{O} = \{T_o\}$ whose optima $\boldsymbol{w}_o$ are far away from optima $\boldsymbol{w}_s$ of normal tasks $\mathcal{S} = \{T_s\}$. To handle this case, MMP can use the robust $\ell_{2,1}$ norm, i.e. $\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{w}_{T_i} - \boldsymbol{w}\|_2$, to tolerate larger distances between $\boldsymbol{w}$ and some outlier optima $\boldsymbol{w}_{T_i}$, and the learned prior $\boldsymbol{w}$ is still close to optima $\boldsymbol{w}_s$ in $\mathcal{S}$ and only requires a few training data for adaptation to new normal tasks. In contrast, it is hard to tailor MAML to handle this case due to its fixed update rule. So being affected by outlier tasks, prior $\boldsymbol{w}$ departures away from $\boldsymbol{w}_s$ and needs more data for adaptation to new normal tasks. To verify this, let us consider an example where 5% outlier images with zero pixels are added into each class in miniImageNet to form outlier tasks. As shown in Fig. 2, our experimental results justify that the outlier tasks can be well handled by MMP+ℓ_{21} to achieve robust meta-learning. We will update this into the revision.

3) We would like to clarify that the step number 15 mentioned in the text is used in the meta-training phase, while the step number 32 mentioned in Fig. 1 (of the submission) is used for fine-tuning (meta-test).

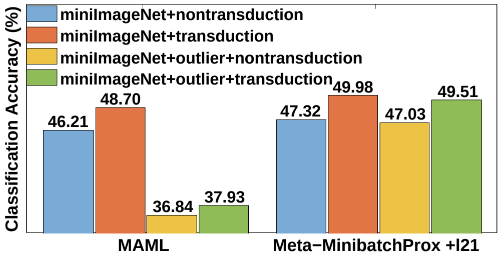

Fig. 2. Performance comparison on 1-shot 5-way outlier-corrupted data .

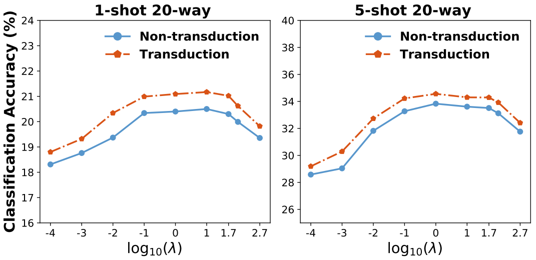

Fig. 3. Impact of λ to classification accuracy on miniImageNet.

To Reviewer #3.

1) We report the impact of λ on the testing performance of our method in Fig. 3. When the value of λ ranges from 10^{-1} to $10^{1.7}$, the performance of our method are relatively stable, demonstrating its insensitivity to the choice of λ.

2) To highlight the difference between MAML and Meta-MinibatchProx (**MMP**), MAML aims to find an initialization $\boldsymbol{w}$ such that $\boldsymbol{w}_T^* = \boldsymbol{w} - \eta\nabla\mathcal{L}_{D_T}(\boldsymbol{w}) = \text{argmin}_{\boldsymbol{w}_T} \langle\nabla\mathcal{L}_{D_T}(\boldsymbol{w}), \boldsymbol{w}_T - \boldsymbol{w}\rangle + \frac{1}{2\eta}\|\boldsymbol{w}_T - \boldsymbol{w}\|_2^2$ is close to the optimal hypothesis of task T. Differently, **MMP** is defined to find the task-specific optimal hypothesis by computing $\widetilde{\boldsymbol{w}}_T^* = \min_{\boldsymbol{w}_T}\mathcal{L}_{D_T}(\boldsymbol{w}_T) + \frac{\lambda}{2}\|\boldsymbol{w}_T - \boldsymbol{w}\|_2^2$. Essentially speaking, MAML approximates the loss $\mathcal{L}_{D_T}(\boldsymbol{w}_T)$ using its first-order taylor expansion for computing an approximate optimum $\boldsymbol{w}_T^*$; while MMP directly optimizes $\mathcal{L}_{D_T}(\boldsymbol{w}_T) = \langle\nabla\mathcal{L}_{D_T}(\boldsymbol{w}), \boldsymbol{w}_T - \boldsymbol{w}\rangle + \frac{1}{2}\langle\nabla^2\mathcal{L}_{D_T}(\boldsymbol{w})(\boldsymbol{w}_T - \boldsymbol{w}), (\boldsymbol{w}_T - \boldsymbol{w})\rangle + \frac{1}{6}\langle\nabla^3\mathcal{L}_{D_T}(\boldsymbol{w}), (\boldsymbol{w}_T - \boldsymbol{w})^{\otimes^3}\rangle + \cdots$. Therefore, **MMP** is able to make use of higher-order information of $\mathcal{L}_{D_T}$ beyond gradient to search optimal hypothesis around the prior $\boldsymbol{w}$, which could lead to better task-specific hypothesis and the prior hypothesis as well.