Supplementary File for Task Similarity Aware Meta Learning: A Theory-improved MAML

Pan Zhou[*], **Yingtian Zou**[†], **Xiaotong Yuan**[‡], **Jiashi Feng**[†], **Caiming Xiong**[*], **Steven C.H. Hoi**[*]
[*]Salesforce Research, [†] National University of Singapore
[†] Nanjing University of Information Science & Technology
{pzhou, cxiong, shoi}@salesforce.com, zouyingt@comp.nus.edu.sg
xtyuan@nuist.edu.cn, elefjia@nus.edu.sg

Abstract

This supplementary document contains the technical proofs of the results and some additional experimental results of the NeurIPS'20 workshop submission entitled "Task Similarity Aware Meta Learning: Theory-inspired Improvement on MAML". It is structured as follows. Appendix A first provides more experimental results and details, and then presents investigation on the robustness of initialization number and curves of loss decrease along more gradient steps for adaptation. Appendix A also provides comparison between TSA-MAML and MAML using larger model. Then Appendix B gives the proofs of the main results in Sec. 3.2, including Theorem 1. Finally, in Appendix C we presents the proofs of Corollary 1 in Sec. 4.

Table 3: Few-shot classification accuracy (%) of the compared approaches on the CIFARFS dataset. The reported accuracies are averaged over 600 test episodes with 95% confidence intervals.

method	1-shot 5-way	5-shot 5-way	1-shot 10-way	5-shot 10-way
Matching Net [1]	36.64 ± 1.13	42.68 ± 0.96	15.02 ± 1.05	32.53 ± 0.93
Meta-LSTM [2]	41.93 ± 1.20	61.40 ± 1.15	31.40 ± 0.75	41.25 ± 0.66
Reptile [3]	51.26 ± 0.99	68.62 ± 0.98	35.73 ± 0.94	54.35 ± 0.91
HSML [4]	46.72 ± 0.87	68.76 ± 0.76	33.89 ± 0.55	53.94 ± 0.49
MMAML [5]	40.64 ± 0.50	49.64 ± 0.49	23.80 ± 0.28	37.19 ± 0.27
FOMAML [6]	47.03 ± 1.47	64.20 ± 1.38	34.65 ± 1.09	51.35 ± 1.16
MAML [6]	51.98 ± 0.87	68.91 ± 0.74	38.48 ± 0.55	55.24 ± 0.54
TSA-MAML	**53.07 ± 0.85**	**71.37 ± 0.74**	**39.77 ± 0.53**	**58.05 ± 0.56**
Reptile + Transduction [3]	54.03 ± 0.92	72.60 ± 0.83	38.41 ± 0.97	57.16 ± 0.87
HSML + Transduction [4]	54.71 ± 1.50	69.62 ± 1.01	38.49 ± 1.22	55.51 ± 0.68
FOMAML + Transduction [6]	49.30 ± 1.18	66.96 ± 1.27	37.83 ± 1.06	53.23 ± 1.12
MMAML+ Transduction [5]	45.16 ± 0.58	58.56 ± 0.51	27.30 ± 0.25	41.26 ± 0.26
MAML + Transduction [6]	57.46 ± 0.90	72.75 ± 0.71	39.97 ± 0.56	56.21 ± 0.55
TSA-MAML + Transduction	**58.21 ± 0.93**	**73.52 ± 0.72**	**42.18 ± 0.58**	**58.69 ± 0.56**

A More Experiments

A.1 More Evaluation Results on Group-Structured Data

Experimental setting. Following [6, 7], we use the episodic procedure for K-shot N-way few-shot learning task. We use the same 4-layered convolution network in [6, 3] for evaluation. In TSA-MAML, we set its initialization number m as three and the task number as $n = 10,000$ for clustering in k-means. For training, we use Adam [8] with learning rate 10^{-3} and total iteration number $S = 40,000$. To be more stable, we use cosine annealing in [9] to gradually decrease the learning rate. We sample 600 test tasks from each sub-dataset for evaluation. Here we test all methods on the 5-shot 5-way learning tasks under the transduction setting where test tasks share information via batch normalization [3], since the baselines are reported under this setting [6, 4].

Results. Fig. 3 further reports the usage frequency of the multiple initializations learnt by TSA-MAML when testing new tasks. After learning three initializations, we sample 1,000 test tasks from

4th Workshop on Meta-Learning at NeurIPS 2020, Vancouver, Canada.

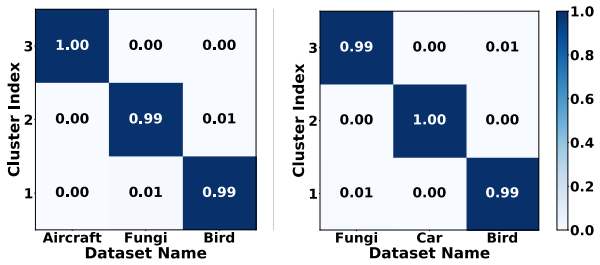


Figure 3: Usage frequency of multiple initializations in TSA-MAML on new tasks.

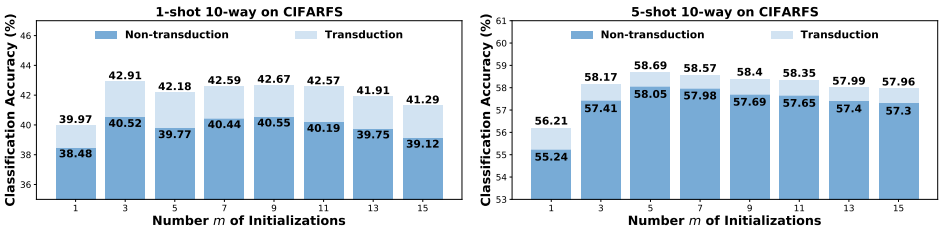


Figure 4: Effects of initialization number to TSA-MAML.

each sub-dataset of the group-structured dataset, and then assign one initialization for each test task by first using MAML to find its approximate optimal model $\boldsymbol{\theta}_T$ and selecting a learnt initialization with smallest distance to $\boldsymbol{\theta}_T$. The values in the (i, j)-th grid in Fig. 3 denotes the frequency that TSA-MAML assigns the i-th learnt initialization to the tasks from the j-th sub-dataset. From these results in Fig. 3, one can observe that in most cases, TSA-MAML assigns the same learnt initialization for the tasks from the same sub-dataset. This well demonstrates that TSA-MAML has leveraged the task similarity and thus can well learn the group structures in the tasks, explaining the superiority over state-of-the-arts.

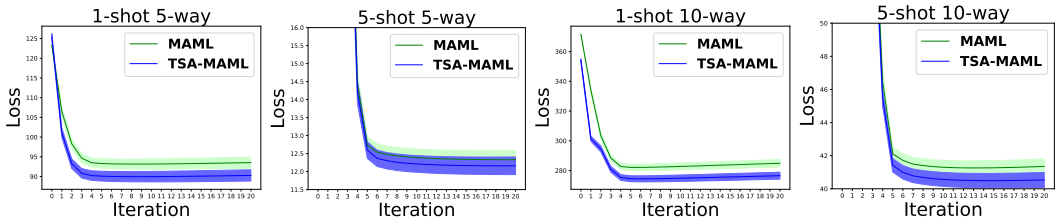
A.2 More Evaluation Results on Real Data

Datasets. We evaluate TSA-MAML on two benchmarks, CIFARFS [10] and tieredImageNet [11]. CIFARFS is a recently proposed few-shot classification benchmark. It splits the 100 classes from CIFAR-100 [12] into 64, 16 and 20 classes for training, validation, and test respectively. Each class contains 600 images of size $32 \times 32 \times 3$. TieredImageNet contains 608 classes from ILSVRC-12 dataset [13], in which each class has 600 images of size $84 \times 84 \times 3$. Moreover, it groups classes into broader hierarchy categories corresponding to higher-level nodes in the ImageNet [14]. Specifically, there are total 34 top hierarchy categories which are further split into 20 training categories (351 classes), 6 validation categories (97 classes) and 8 test categories (160 classes). So all training classes are sufficiently distinct from the test classes, giving a more challenging learning task.

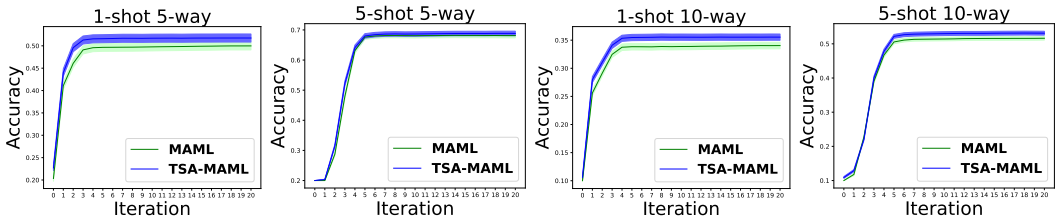
Experimental setting. We use the same network architecture, training strategy and task number n in Sec. A.1. In TSA-MAML, the training iteration number S is $40,000$ for CIFARFS and $80,000$ for tieredImageNet and the cluster number m is five for both datasets. Like [6, 3], we test all methods on 600 test episodes under (non-)transduction settings. In non-transduction, batch normalization statistics are collected from all training data and one test sample. See transduction setting in Sec. A.1.

Results. From Table 3, one can observe that TSA-MAML consistently outperforms optimization based methods, *e.g.* MAML, HSML and MMAML, and metric based method, *e.g.* Matching Net. Specifically, on CIFARFS, TSA-MAML respectively brings about 1.09%, 2.46%, 1.29% and 2.81% improvements on the four test cases (from left to right) under non-transduction setting, and under transduction setting it also makes about 0.75%, 0.77%, 2.21% and 2.48% improvements for the four cases. Similarly, on tieredImageNet, it averagely improves by about 1.68% and 1.20% on the four test cases under non-transduction and transduction cases. These results demonstrate the advantages of TSA-MAML behind which the reasons have been discussed in Sec. A.1. Besides, compared with MAML, TSA-MAML respectively makes about 1.73% and 1.44% average improvements on CIFARFS and tieredImageNet. These observations further confirm our theories in Sec. 3.2.

Robustness of TSA-MAML to The Number of Initializations. Fig. 3 shows the effects of initialization number m to the testing performance of TSA-MAML. When m ranges from 3 to 11, the performance of TSA-MAML on 1-shot 10-way learning tasks on CIFARFS are relatively stable. So



(a) Loss on the tieredImageNet dataset



(b) Classification accuracy on the tieredImageNet dataset

Figure 5: Illustration of loss decrease and classification accuracy of MAML and TSA-MAML. One can observe that at the first several optimization gradient descent steps, both MAML and TSA-MAML decrease very fast but reduce slowly along with optimization steps. Similarly, one can find that Moreover, one can find that both MAML and TSA-MAML increase their accuracy very fast but improve it slowly along with gradient descent steps. Moreover, TSA-MAML usually decreases its loss and improves its accuracy faster than MAML since its group-specific initializations are more closer to the optimal task-specific models and thus can be adapted to new tasks more quickly.

TSA-MAML is robust to m. These results testify the robustness of TSA-MAML to the choice of m. They also indicate that using MAML to estimate optimal model parameters of tasks and then clustering these model parameters according to their distances to the m group-specific initializations is valid when m is not large. This is because assigning tasks into m groups means dividing model parameter space into m regions and is not hard when m is not large, as estimating approximate location of optimal task models in the parameter space is sufficient and can be achieved by MAML.

A.3 Fast Decrease of Losses of MAML and TSA-MAML at The First Several Gradient Descent Steps

In this subsection, we investigate the loss decreases in MAML and TSA-MAML from the learnt initializations along with gradient descent steps. Here we evaluate on TieredImageNet dataset. Specifically, we randomly sample 600 tasks and compute their losses and classification accuracy with along gradient descent steps. Then we report the average loss and accuracy of these 600 tasks. From Fig. 5, one can observe that at the first several gradient descent steps (*e.g.* 7), both the losses $\mathcal{L}_{D_T}(\boldsymbol{\theta}^t)$ of MAML and TSA-MAML decrease very fast, but with along more optimization gradient steps, they reduce very slowly. This is because the training dataset is very small and thus a few gradient descent steps are sufficient to fit these data. Note for MAML, the first term in the upper bound in Theorem 1 always increases exponentially along with the gradient steps. In this way, to achieve smaller excess risk, we should not run many gradient steps. This well explains why MAML usually adapts the learnt initialization $\boldsymbol{\theta}^*$ to new tasks by taking only a few gradient descent steps. Similarly for TSA-MAML, our Corollary 1 also suggests us to use a few gradient descent steps for fast adaptation since we also need to balance the two terms in the upper bound of the excess risk. Also, for accuracy, we can also observe very similar phenomena. Besides, by comparison, we also observe that the loss of TSA-MAML is always much smaller than MAML which means that compared the common initialization learnt by MAML for all tasks, the learnt group-special initializations by TSA-MAML are closer to the optimal model parameters of the testing tasks. We can also observe similar results on the accuracy metric. These results well demonstrate the advantages of TSA-MAML over MAML.

A.4 Comparison between TSA-MAML and MAML Using Larger Network Model

In Sec. A.1 of the manuscript, we have explained that the advantage of TSA-MAML over MAML comes from its design principle introduced above instead of higher model complexity. This is because MAML and TSA-MAML use the same network and have same parameter dimension and thus same model complexity (data fitting capacity) [15]. We also conduct experiments to further investigate this view. On the group-structured dataset, we test MAML whose network is $3\times$ larger than that used in TSA-MAML. For brevity, we call the MAML using large network MAML-L. By comparison, TSA-MAML still outperforms MAML-L. Here we test MAML-L and TSA-MAML on CIFARFS. We directly increase the depth of MAML-L from 4 to 12 and then test its performance. From the results in Table 4, one can observe that TSA-MAML performs better than MAML-L. Indeed, MAML-L faces over-fitting issue for few-shot learning, which can be observed from the comparison between MAML-L and MAML. So these experimental results further show the superiority of TSA-MAML over MAML comes from its design principle instead of higher model complexity. Unlike MAML learning one initialization for all tasks, TSA-MAML clusters similar tasks into the same group and learns group-specific initialization which can faster and better adapt itself to tasks in the same group.

Table 4: Few-shot classification accuracy (%) of the compared approaches on the CIFAR-FS dataset. The reported accuracies are averaged over 600 test episodes with 95% confidence intervals.

CIFAR-FS (transduction)	1-shot 5-way	5-shot 5-way	1-shot 10-way	5-shot 10-way
MAML(4 layers)	57.46 ± 0.90	72.75 ± 0.71	39.97 ± 0.56	56.21 ± 0.55
MAML-L(12 layers)	55.66 ± 1.04	68.29 ± 0.78	40.77 ± 0.64	53.74 ± 0.53
TSA-MAML (4 layers)	$\mathbf{58.21 \pm 0.93}$	$\mathbf{73.52 \pm 0.72}$	$\mathbf{42.18 \pm 0.58}$	$\mathbf{58.69 \pm 0.56}$

B Proof of The Results in Sec. 3.2

B.1 Auxiliary Lemmas

In this section, we introduce auxiliary lemmas which will be used for proving the results in Sec. 3.2.

Lemma 1. *Assume that $\ell(f(\boldsymbol{\theta}_T, \boldsymbol{x}), \boldsymbol{y})$ is L_s-smooth in $\boldsymbol{\theta}_T$. If $\alpha \leq \frac{1}{L_s}$, then it holds for any task T and any parameter $\boldsymbol{\theta}_T$ that*

$$\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^1) - \mathcal{L}_{D_T}(\boldsymbol{\theta}_T) \leq \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2,$$

where $\boldsymbol{\theta}^*$ denotes the learned initialization and $\boldsymbol{\theta}_T^1 = \boldsymbol{\theta}^* - \alpha\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*)$.

Proof. Let $h_{D_T}(\boldsymbol{\theta}_T) = \langle\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*), \boldsymbol{\theta}_T - \boldsymbol{\theta}^*\rangle + \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|_2^2$. Then we know that $\boldsymbol{\theta}_T^1 = \text{argmin}_{\boldsymbol{\theta}_T} h_{D_T}(\boldsymbol{\theta}_T) = \boldsymbol{\theta}^* - \alpha\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*)$. By using Taylor expansion, for $\boldsymbol{\theta}^*$ and any $\boldsymbol{\theta}_T$, there exists a constant $\lambda \in (0, L_s]$ such that

$$\mathcal{L}_{D_T}(\boldsymbol{\theta}_T) = \mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + h_{D_T}(\boldsymbol{\theta}_T) + \frac{1}{2}\left(\lambda - \frac{1}{\alpha}\right)\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2. \tag{1}$$

Then respectively replacing $\boldsymbol{\theta}_T$ and λ with $\boldsymbol{\theta}_T^1$ and $\lambda^* \in (0, L_s]$, conducting subtraction on the two equations, we can obtain

$$\begin{aligned}\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^1) - \mathcal{L}_{D_T}(\boldsymbol{\theta}_T) =& h_{D_T}(\boldsymbol{\theta}_T^1) - h_{D_T}(\boldsymbol{\theta}_T) + \frac{\lambda^* - \frac{1}{\alpha}}{2}\|\boldsymbol{\theta}_T^1 - \boldsymbol{\theta}^*\|^2 - \frac{\lambda - \frac{1}{\alpha}}{2}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2 \\ \overset{①}{\leq}& \frac{\lambda^* - \frac{1}{\alpha}}{2}\|\boldsymbol{\theta}_T^1 - \boldsymbol{\theta}^*\|^2 - \frac{\lambda - \frac{1}{\alpha}}{2}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2 \\ \overset{②}{\leq}& \frac{\frac{1}{\alpha} - \lambda}{2}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2 \leq \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2,\end{aligned}$$

where ① uses the fact that $\boldsymbol{\theta}_T^1$ is the optimum of $h_{D_T}(\boldsymbol{\theta}_T)$ giving $h(\boldsymbol{\theta}_T^1) \leq h(\boldsymbol{\theta}_T)$; in ②, we set $\alpha \leq \frac{1}{L_s}$ giving $\lambda - 1/\alpha \leq 0$ and $\lambda^* - 1/\alpha \leq 0$ due to $\lambda, \lambda^* \in (0, L_s]$. The proof is completed. □

Lemma 2. *Assume that $\ell(f(\boldsymbol{\theta}_T, \boldsymbol{x}), \boldsymbol{y})$ is L_s-smooth in $\boldsymbol{\theta}_T$. If $\alpha \leq \frac{1}{L_s}$, then it holds for any task T and parameter $\boldsymbol{\theta}_T$ that*

$$\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) - \mathcal{L}_{D_T}(\boldsymbol{\theta}_T) \leq \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2 - \alpha\left(1 - \frac{\alpha L_s}{2}\right)\sum_{t=1}^{q-1}\|\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t)\|_2^2,$$

where $\boldsymbol{\theta}^*$ denotes the learned initialization and $\boldsymbol{\theta}_T^q = \boldsymbol{\theta}^* - \alpha\left(\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1}\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t)\right)$. Here $\boldsymbol{\theta}_T^t = \boldsymbol{\theta}^* - \alpha\left(\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + \sum_{s=1}^{t-1}\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^s)\right)$ with $\boldsymbol{\theta}_T^1 = \boldsymbol{\theta}^* - \alpha\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*)$ denotes the adapted parameter after the t-th iteration.

Proof. Let $h_{D_T}(\boldsymbol{\theta}_T) = \langle\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*), \boldsymbol{\theta}_T - \boldsymbol{\theta}^*\rangle + \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|_2^2$. Then we know that $\boldsymbol{\theta}_T^1 = \operatorname{argmin}_{\boldsymbol{\theta}_T} h_{D_T}(\boldsymbol{\theta}_T) = \boldsymbol{\theta}^* - \alpha\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*)$. Then by using Lemma 1, we can obtain the following results. If $\alpha \leq \frac{1}{L_s}$, then it holds for any task T and parameter $\boldsymbol{\theta}_T$ that

$$\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^1) - \mathcal{L}_{D_T}(\boldsymbol{\theta}_T) \leq \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2.$$

At the same time, we have $\boldsymbol{\theta}_T^q = \boldsymbol{\theta}^* - \alpha\left(\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1}\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t)\right) = \boldsymbol{\theta}_T^1 - \alpha\sum_{t=1}^{q-1}\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t)$. This actually means that we want to minimize the loss $\mathcal{L}_{D_T}(\boldsymbol{\theta})$ from the initialization $\boldsymbol{\theta}_T^1$. Specifically, here we only run $(q-1)$ gradient steps. In this way, we can upper bound the loss at each iteration as follows:

$$\begin{aligned}\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^{t+1}) \leq& \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t) + \langle\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t), \boldsymbol{\theta}_T^{t+1} - \boldsymbol{\theta}_T^t\rangle + \frac{L_s}{2}\|\boldsymbol{\theta}_T^{t+1} - \boldsymbol{\theta}_T^t\|^2\\ \overset{①}{=}& \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t) - \alpha\left(1 - \frac{\alpha L_s}{2}\right)\|\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t)\|_2^2,\end{aligned}$$

where ① uses $\boldsymbol{\theta}_T^{t+1} - \boldsymbol{\theta}_T^t = -\alpha\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t)$. In this way, summing up from $t = 1$ to $q-1$, we have

$$\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) \leq \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^1) - \alpha\left(1 - \frac{\alpha L_s}{2}\right)\sum_{t=1}^{q-1}\|\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t)\|_2^2.$$

Therefore, we have

$$\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) - \mathcal{L}_{D_T}(\boldsymbol{\theta}_T) \leq \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2 - \alpha\left(1 - \frac{\alpha L_s}{2}\right)\sum_{t=1}^{q-1}\|\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t)\|_2^2.$$

The proof is completed. □

Lemma 3. *Assume that $\ell(f(\boldsymbol{\theta}, \boldsymbol{x}), \boldsymbol{y})$ is G-Lipschitz continuous and L_s-smooth with respect to $\boldsymbol{\theta}$. Given a learning task T, let $\mathcal{L}(\boldsymbol{\theta}_T) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim T}\left[\ell(f(\boldsymbol{\theta}_T, \boldsymbol{x}), \boldsymbol{y})\right]$ and $\mathcal{L}_{D_T}(\boldsymbol{\theta}_T) = \frac{1}{K}\sum_{(\boldsymbol{x},\boldsymbol{y})\in D_T}\ell(f(\boldsymbol{\theta}_T, \boldsymbol{x}), \boldsymbol{y})$ respectively denote the expected and empirical losses on $D_T = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^K \sim T$. Consider the following empirical minimization problem:*

$$\boldsymbol{\theta}_T^1 = \operatorname*{argmin}_{\boldsymbol{\theta}_T}\left\{h_{D_T}(\boldsymbol{\theta}_T) = \langle\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*), \boldsymbol{\theta}_T - \boldsymbol{\theta}^*\rangle + \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|_2^2\right\} = \boldsymbol{\theta}^* - \alpha\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*).$$

Assume $D_T^{(i)}$ is identical to D_T except that one of the $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is replaced by another random sample $(\boldsymbol{x}_i', \boldsymbol{y}_i')$. We then denote

$$\boldsymbol{\theta}_{T,i}^1 = \operatorname*{argmin}_{\boldsymbol{\theta}_T} h_{D_T^{(i)}}(\boldsymbol{\theta}_T) = \boldsymbol{\theta}^* - \alpha\nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}^*),$$

where $h_{D_T^{(i)}}(\boldsymbol{\theta}_T) := \frac{1}{K}\left(\left\langle\sum_{j\neq i}\nabla\ell(f(\boldsymbol{\theta}_T, \boldsymbol{x}_j), \boldsymbol{y}_j) + \nabla\ell(f(\boldsymbol{\theta}_T, \boldsymbol{x}_i'), \boldsymbol{y}_i'), \boldsymbol{\theta}_T - \boldsymbol{\theta}^*\right\rangle\right) + \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|_2^2$. Then the following bound holds that

$$\|\boldsymbol{\theta}_T^1 - \boldsymbol{\theta}_{T,i}^1\| \leq \frac{4\alpha G}{K}.$$

Proof. The result can be proved by stability argument. For brevity, let $r(\boldsymbol{\theta}_T) = \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2$ is an $\frac{1}{\alpha}$-strongly convex regularization function. Then we can show that

$$
\begin{aligned}
&h_{D_T}(\boldsymbol{\theta}^1_{T,i}) - h_{D_T}(\boldsymbol{\theta}^1_T) \\
=&\left\langle \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^*), \boldsymbol{\theta}^1_{T,i} - \boldsymbol{\theta}^1_T \right\rangle + r(\boldsymbol{\theta}^1_{T,i}) - r(\boldsymbol{\theta}^1_T) \\
=&\frac{1}{K}\sum_{j\neq i}\left(\langle \nabla \ell(f(\boldsymbol{\theta}^*, \boldsymbol{x}_j), \boldsymbol{y}_j), \boldsymbol{\theta}^1_{T,i} - \boldsymbol{\theta}^1_T\rangle\right) + \frac{1}{K}\langle \nabla \ell(f(\boldsymbol{\theta}^*, \boldsymbol{x}_i), \boldsymbol{y}_i), \boldsymbol{\theta}^1_{T,i} - \boldsymbol{\theta}^1_T\rangle + r(\boldsymbol{\theta}^1_{T,i}) - r(\boldsymbol{\theta}^1_T) \\
=&h_{D_T^{(i)}}(\boldsymbol{\theta}^1_{T,i}) - h_{D_T^{(i)}}(\boldsymbol{\theta}^1_T) + \frac{1}{K}\langle \nabla \ell(f(\boldsymbol{\theta}^*, \boldsymbol{x}_i), \boldsymbol{y}_i), \boldsymbol{\theta}^1_{T,i} - \boldsymbol{\theta}^1_T\rangle - \frac{1}{K}\langle \nabla \ell(f(\boldsymbol{\theta}^*, \boldsymbol{x}'_i), \boldsymbol{y}'_i), \boldsymbol{\theta}^1_{T,i} - \boldsymbol{\theta}^1_T\rangle \\
\overset{①}{\leq}&\frac{1}{K}\left[\|\nabla \ell(f(\boldsymbol{\theta}^*, \boldsymbol{x}_i), \boldsymbol{y}_i)\| + \|\nabla \ell(f(\boldsymbol{\theta}^*, \boldsymbol{x}'_i), \boldsymbol{y}'_i)\|\right] \cdot \left\|\boldsymbol{\theta}^1_{T,i} - \boldsymbol{\theta}^1_T\right\| \\
\overset{②}{\leq}&\frac{2G}{K}\|\boldsymbol{\theta}^1_{T,i} - \boldsymbol{\theta}^1_T\|,
\end{aligned}
$$

where in ① we have used the optimality of $\boldsymbol{\theta}^1_{T,i}$ with respect to $h_{D_T^{(i)}}(\boldsymbol{\theta}_T)$, and in ② we use the Lipschitz continuity of the loss function $\ell(\cdot)$. Since $h_{D_T^{(i)}}(\boldsymbol{\theta}_T)$ is $\frac{1}{\alpha}$-strongly-convex, it is easily to verify that

$$
h_{D_T}(\boldsymbol{\theta}^1_{T,i}) \geq h_{D_T}(\boldsymbol{\theta}^1_T) + \frac{1}{2\alpha}\|\boldsymbol{\theta}^1_{T,i} - \boldsymbol{\theta}^1_T\|^2.
$$

Then combining the above two inequalities we arrive at

$$
\|\boldsymbol{\theta}^1_{T,i} - \boldsymbol{\theta}^1_T\| \leq \frac{4\alpha G}{K}.
$$

The proof is concluded. □

Lemma 4. *Assume that $\ell(f(\boldsymbol{\theta}, \boldsymbol{x}), \boldsymbol{y})$ is G-Lipschitz continuous and L_s-smooth with respect to $\boldsymbol{\theta}$. Given a learning task T, let $\mathcal{L}(\boldsymbol{\theta}_T) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim T}\left[\ell(f(\boldsymbol{\theta}_T, \boldsymbol{x}), \boldsymbol{y})\right]$ and $\mathcal{L}_{D_T}(\boldsymbol{\theta}_T) = \frac{1}{K}\sum_{(\boldsymbol{x},\boldsymbol{y})\in D_T}\ell(f(\boldsymbol{\theta}_T, \boldsymbol{x}), \boldsymbol{y})$ respectively denote the expected and empirical losses on $D_T = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^K \sim T$. Consider the following empirical minimization problem:*

$$
\begin{aligned}
\boldsymbol{\theta}^q_T =&\operatorname*{argmin}_{\boldsymbol{\theta}_T}\left\{h_{D_T}(\boldsymbol{\theta}_T) = \left\langle \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1}\nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^t_T), \boldsymbol{\theta}_T - \boldsymbol{\theta}^*\right\rangle + \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|_2^2\right\} \\
=&\boldsymbol{\theta}^* - \alpha\left(\nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1}\nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^t_T)\right)
\end{aligned}
$$

where $\boldsymbol{\theta}^t_T = \boldsymbol{\theta}^* - \alpha\left(\nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + \sum_{s=1}^{t-1}\nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^s_T)\right)$ with $\boldsymbol{\theta}^1_{T,i} = \boldsymbol{\theta}^* - \alpha\nabla \mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}^*)$ denotes the adapted parameter after the t-th iteration. Then for any q, the following bound holds that

$$
\left|\mathbb{E}_{D_T\sim T}\left[\mathcal{L}(\boldsymbol{\theta}^q_T) - \mathcal{L}_{D_T}(\boldsymbol{\theta}^q_T)\right]\right| \leq \frac{2G^2\left[(1+2\alpha L_s)^q - 1\right]}{L_s K}
$$

and

$$
\left\|\mathbb{E}_{D_T\sim T}\left[\nabla \mathcal{L}(\boldsymbol{\theta}^q_T) - \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^q_T)\right]\right\| \leq \frac{2G\left[(1+2\alpha L_s)^q - 1\right]}{K}.
$$

Proof. For brevity, let $r(\boldsymbol{\theta}_T) = \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2$ is an $\frac{1}{\alpha}$-strongly convex regularization function. Let us consider $D_T^{(i)}$ which is identical to D_T except that one of the $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is replaced by another random sample $(\boldsymbol{x}'_i, \boldsymbol{y}'_i)$. We then denote

$$
\boldsymbol{\theta}^q_{T,i} = \operatorname*{argmin}_{\boldsymbol{\theta}_T}\left\{h_{D_T^{(i)}}(\boldsymbol{\theta}_T) := \left\langle \nabla \mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1}\nabla \mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}^t_{T,i}), \boldsymbol{\theta}_T - \boldsymbol{\theta}^*\right\rangle + \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|_2^2\right\},
$$

where $\boldsymbol{\theta}_{T,i}^t = \boldsymbol{\theta}^* - \alpha\left(\nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}^*) + \sum_{s=1}^{t-1}\nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_{T,i}^s)\right)$ with $\boldsymbol{\theta}_{T,i}^1 = \boldsymbol{\theta}^* - \alpha\nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}^*)$ denotes the adapted parameter after the t-th iteration on the dataset $D_T^{(i)}$. Then we can show that

$$h_{D_T}(\boldsymbol{\theta}_{T,i}^q) - h_{D_T}(\boldsymbol{\theta}_T^q) = \left\langle \nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1}\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t), \boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q \right\rangle + r(\boldsymbol{\theta}_{T,i}^q) - r(\boldsymbol{\theta}_T^q)$$

$$= h_{D_T^{(i)}}(\boldsymbol{\theta}_{T,i}^q) - h_{D_T^{(i)}}(\boldsymbol{\theta}_T^q) + \left\langle \nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*) - \nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1}\left[\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t) - \nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_{T,i}^t)\right], \boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q \right\rangle.$$

Now we bound the term $\left\langle \nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}), \boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q \right\rangle$ with $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ or $\boldsymbol{\theta} = \boldsymbol{\theta}_T^t$ ($t = 1, \cdots, q-1$) in the above equation as follows:

$$\begin{aligned}
\left\langle \nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}), \boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q \right\rangle &= \frac{1}{K}\langle \nabla\ell(f(\boldsymbol{\theta}, \boldsymbol{x}_i), \boldsymbol{y}_i), \boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\rangle - \frac{1}{K}\langle \nabla\ell(f(\boldsymbol{\theta}, \boldsymbol{x}_i'), \boldsymbol{y}_i'), \boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\rangle \\
&\leq \frac{1}{K}\left[\|\nabla\ell(f(\boldsymbol{\theta}, \boldsymbol{x}_i), \boldsymbol{y}_i)\| + \|\nabla\ell(f(\boldsymbol{\theta}, \boldsymbol{x}_i'), \boldsymbol{y}_i')\|\right] \cdot \left\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\right\| \\
&\overset{①}{\leq} \frac{2G}{K}\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\|,
\end{aligned}$$

where in ① we use the Lipschitz continuity of the loss function G. At the same time, by using the optimality of $\boldsymbol{\theta}_{T,i}^q$ with respect to $h_{D_T^{(i)}}(\boldsymbol{\theta}_T)$ which means $h_{D_T^{(i)}}(\boldsymbol{\theta}_{T,i}^q) \leq h_{D_T^{(i)}}(\boldsymbol{\theta}_T^q)$, we can further obtain

$$h_{D_T}(\boldsymbol{\theta}_{T,i}^q) - h_{D_T}(\boldsymbol{\theta}_T^q) \leq \left\langle \nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*) - \nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1}\left[\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t) - \nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_{T,i}^t)\right], \boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q \right\rangle$$

$$= \left\langle \nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*) - \nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1}\left[\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t) - \nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_T^t) + \nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_T^t) - \nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_{T,i}^t)\right], \boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q \right\rangle$$

$$\leq \frac{2qG}{K}\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\| + \sum_{t=1}^{q-1}\left\langle \nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_T^t) - \nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_{T,i}^t), \boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q \right\rangle$$

$$\leq \frac{2qG}{K}\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\| + \sum_{t=1}^{q-1}\left\|\nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_T^t) - \nabla\mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_{T,i}^t)\right\| \cdot \left\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\right\|$$

$$\leq \frac{2qG}{K}\left\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\right\| + L_s\left\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\right\| \sum_{t=1}^{q-1}\left\|\boldsymbol{\theta}_T^t - \boldsymbol{\theta}_{T,i}^t\right\|.$$

Since $h_{D_T^{(i)}}(\boldsymbol{\theta}_T)$ is $\frac{1}{\alpha}$-strongly-convex, it is easily to verify that

$$h_{D_T}(\boldsymbol{\theta}_{T,i}^q) \geq h_{D_T}(\boldsymbol{\theta}_T^q) + \frac{1}{2\alpha}\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\|^2.$$

Then combining the above two inequalities we arrive at

$$\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\| \leq \frac{4\alpha qG}{K} + 2\alpha L_s\sum_{t=1}^{q-1}\left\|\boldsymbol{\theta}_T^t - \boldsymbol{\theta}_{T,i}^t\right\|.$$

Note that $\|\boldsymbol{\theta}_{T,i}^1 - \boldsymbol{\theta}_T^1\| \leq \frac{4\alpha G}{K}$ in Lemma 3. Then we can easily obtain

$$\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\| \leq \frac{2G\left[(1+2\alpha L_s)^q - 1\right]}{L_s K}.$$

It then follows consequently from the Lipschitz continuity of ℓ that for any sample $(\boldsymbol{x}, \boldsymbol{y}) \sim T$

$$|\ell(f(\boldsymbol{\theta}_{T,i}^q, \boldsymbol{x}), \boldsymbol{y}) - \ell(f(\boldsymbol{\theta}_T^q, \boldsymbol{x}), \boldsymbol{y})| \leq G\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\| \leq \frac{2G^2\left[(1+2\alpha L_s)^q - 1\right]}{L_s K}. \tag{2}$$

Note that D_T and $D_T^{(i)}$ are both i.i.d. samples of the task T. It follows that

$$\mathbb{E}_{D_T}\left[\mathcal{L}(\boldsymbol{\theta}_T^q)\right] = \mathbb{E}_{D_T^{(i)}}\left[\mathcal{L}(\boldsymbol{\theta}_{T,i}^q)\right] = \mathbb{E}_{D_T^{(i)}\cup\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}}\left[\ell(f(\boldsymbol{\theta}_{T,i}^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right].$$

Since the above holds for all $i = 1, ..., K$, we can show that

$$\mathbb{E}_{D_T}\left[\mathcal{L}(\boldsymbol{\theta}_T^q)\right] = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{D_T^{(i)}\cup\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}}\left[\ell(f(\boldsymbol{\theta}_{T,i}^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right] = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{D_T\cup\{(\boldsymbol{x}_i',\boldsymbol{y}_i')\}}\left[\ell(f(\boldsymbol{\theta}_{T,i}^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right].$$

Concerning the empirical case, we can see that

$$\mathbb{E}_{D_T}\left[\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q)\right] = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{D_T}\left[\ell(f(\boldsymbol{\theta}_T^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right] = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{D_T\cup\{(\boldsymbol{x}_i',\boldsymbol{y}_i')\}}\left[\ell(f(\boldsymbol{\theta}_T^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right].$$

By combining the above two inequalities we get

$$\begin{aligned}\left|\mathbb{E}_{D_T}\left[\mathcal{L}(\boldsymbol{\theta}_T^q) - \mathcal{L}_{D_T}(\boldsymbol{\theta}_{T,i}^q)\right]\right| &= \left|\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{D_T\cup\{(\boldsymbol{x}_i',\boldsymbol{y}_i')\}}\left[\ell(f(\boldsymbol{\theta}_T^q,\boldsymbol{x}_i),\boldsymbol{y}_i) - \ell(f(\boldsymbol{\theta}_{T,i}^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right]\right| \\ &\leq \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{D_T\cup\{(\boldsymbol{x}_i',\boldsymbol{y}_i')\}}\left[\left|\ell(f(\boldsymbol{\theta}_T^q,\boldsymbol{x}_i),\boldsymbol{y}_i) - \ell(f(\boldsymbol{\theta}_{T,i}^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right|\right] \\ &\leq \frac{2G^2\left[(1+2\alpha L_s)^q - 1\right]}{L_s K},\end{aligned}$$

where in the last inequality we have used (2). This proves the objective function inequality in the first part of the lemma. To prove the gradient norm inequality, we note from the smoothness assumption that

$$\|\nabla\ell(f(\boldsymbol{\theta}_T^q,\boldsymbol{x}),\boldsymbol{y}) - \nabla\ell(f(\boldsymbol{\theta}_{T,i}^q,\boldsymbol{x}),\boldsymbol{y})\| \leq L_s\|\boldsymbol{\theta}_T^q - \boldsymbol{\theta}_{T,i}^q\| \leq \frac{2G\left[(1+2\alpha L_s)^q - 1\right]}{K}. \quad (3)$$

The rest of the argument mimics that for the objective value case. Here we provide the details for the sake of completeness. Again, note that D_T and $D_T^{(i)}$ are both i.i.d. samples of the task distribution T. It follows that

$$\mathbb{E}_{D_T}\left[\nabla\mathcal{L}(\boldsymbol{\theta}_T^q)\right] = \mathbb{E}_{D_T^{(i)}}\left[\nabla\mathcal{L}(\boldsymbol{\theta}_{T,i}^q)\right] = \mathbb{E}_{D_T^{(i)}\cup\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}}\left[\nabla\ell(f(\boldsymbol{\theta}_{T,i}^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right].$$

Since the above holds for all $i = 1, ..., K$, we can show that

$$\begin{aligned}\mathbb{E}_{D_T}\left[\nabla\mathcal{L}(\boldsymbol{\theta}_T^q)\right] &= \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{D_T^{(i)}\cup\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}}\left[\nabla\ell(f(\boldsymbol{\theta}_{T,i}^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right] \\ &= \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{D_T\cup\{(\boldsymbol{x}_i',\boldsymbol{y}_i')\}}\left[\nabla\ell(f(\boldsymbol{\theta}_{T,i}^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right].\end{aligned}$$

Concerning the empirical version, we can see that

$$\mathbb{E}_{D_T}\left[\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q)\right] = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{D_T}\left[\nabla\ell(f(\boldsymbol{\theta}_T^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right] = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{D_T\cup\{(\boldsymbol{x}_i',\boldsymbol{y}_i')\}}\left[\nabla\ell(f(\boldsymbol{\theta}_T^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right].$$

By combining the above two inequalities we get

$$\begin{aligned}\left\|\mathbb{E}_{D_T}\left[\nabla\mathcal{L}(\boldsymbol{\theta}_T^q) - \nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_{T,i}^q)\right]\right\| &= \left\|\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{D_T\cup\{(\boldsymbol{x}_i',\boldsymbol{y}_i')\}}\left[\nabla\ell(f(\boldsymbol{\theta}_T^q,\boldsymbol{x}_i),\boldsymbol{y}_i) - \nabla\ell(f(\boldsymbol{\theta}_{T,i}^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right]\right\| \\ &\leq \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{D_T\cup\{(\boldsymbol{x}_i',\boldsymbol{y}_i')\}}\left[\left\|\nabla\ell(f(\boldsymbol{\theta}_T^q,\boldsymbol{x}_i),\boldsymbol{y}_i) - \nabla\ell(f(\boldsymbol{\theta}_{T,i}^q,\boldsymbol{x}_i),\boldsymbol{y}_i)\right\|\right] \\ &\leq \frac{2G\left[(1+2\alpha L_s)^q - 1\right]}{K}.\end{aligned}$$

where in the last inequality we have used (3). The proof is concluded. □

B.2 Proof of Theorem 1

Proof. Consider a fixed task $T \sim \mathcal{T}$ and its associated random sample $D_T \sim T$ of size K. We denote $\mathcal{L}_{D_T}(\boldsymbol{\theta}) = \frac{1}{K}\sum_{(\boldsymbol{x},\boldsymbol{y})\in D_T} \ell(f(\boldsymbol{\theta}_T, \boldsymbol{x}), \boldsymbol{y})$. From Lemma 4, we know that when we adapt q gradient steps to new task, we have

$$\left|\mathbb{E}_{D_T\sim T}\left[\mathcal{L}(\boldsymbol{\theta}_T^q) - \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q)\right]\right| \leq \frac{2G^2\left[(1+2\alpha L_s)^q - 1\right]}{L_s K}. \tag{4}$$

From Lemma 2, when $\alpha \leq \frac{1}{L_s}$, for any $\boldsymbol{\theta}_T$ we have

$$\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) - \mathcal{L}_{D_T}(\boldsymbol{\theta}_T) \leq \frac{1}{2\alpha}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2,$$

where $\boldsymbol{\theta}^*$ denotes the learnt prior.

By taking expectation over the random sample set D_T at $\boldsymbol{\theta}_T = \boldsymbol{\theta}_T^*$ we obtain

$$\mathbb{E}_{D_T}[\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) - \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^*)] \leq \frac{1}{2\alpha}\mathbb{E}_{D_T}\left[\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_T^*\|^2\right]. \tag{5}$$

Then we can show the following

$$\begin{aligned}
\mathbb{E}_{D_T}\left[\mathcal{L}(\boldsymbol{\theta}_T^q) - \mathcal{L}(\boldsymbol{\theta}_T^*)\right] =& \mathbb{E}_{D_T}\left[\mathcal{L}(\boldsymbol{\theta}_T^q) - \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q)\right] + \mathbb{E}_{D_T}\left[\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) - \mathcal{L}(\boldsymbol{\theta}_T^*)\right] \\
\leq& \left|\mathbb{E}_{D_T}\left[\mathcal{L}(\boldsymbol{\theta}_T^q) - \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q)\right]\right| + \mathbb{E}_{D_T}\left[\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) - \mathcal{L}(\boldsymbol{\theta}_T^*)\right] \\
\overset{①}{\leq}& \frac{2G^2\left[(1+2\alpha L_s)^q - 1\right]}{L_s K} + \mathbb{E}_{D_T}\left[\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) - \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^*)\right] \\
\overset{②}{\leq}& \frac{2G^2\left[(1+2\alpha L_s)^q - 1\right]}{L_s K} + \frac{1}{2\alpha}\mathbb{E}_{D_T}\left[\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_T^*\|^2\right],
\end{aligned}$$

where ① uses Eqn. (4) and ② employs inequality (5). Note that $\mathbb{E}_{D_T}[\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^*)] = \mathbb{E}_{D_T}[\mathcal{L}(\boldsymbol{\theta}_T^*)]$. Then we can take expectation of both sides of the above over $T \sim \mathcal{T}$ to obtain

$$\begin{aligned}
\mathbb{E}_{T\sim\mathcal{T}}\mathbb{E}_{D_T}\left[\mathcal{L}(\boldsymbol{\theta}_T^q) - \mathcal{L}(\boldsymbol{\theta}_T^*)\right] \leq& \frac{2G^2\left[(1+2\alpha L_s)^q - 1\right]}{L_s K} + \mathbb{E}_{T\sim\mathcal{T}}\mathbb{E}_{D_T}\left[\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) - \mathcal{L}(\boldsymbol{\theta}_T^*)\right] \\
\leq& \frac{2G^2\left[(1+2\alpha L_s)^q - 1\right]}{L_s K} + \frac{1}{2\alpha}\mathbb{E}_{T\sim\mathcal{T}}\left[\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_T^*\|^2\right].
\end{aligned}$$

This proves the results in the theorem. □

C Proofs of The Results in Sec. 4

C.1 Proof of Corollary 1

Proof. For the results in Corollary 1, we can easily follow the proof sketch of Theorems 1 and 2 to obtain this kind of results. Specifically, we can replace the one common initialization $\boldsymbol{\theta}^*$ by the learned $\{\boldsymbol{\theta}_i^*\}_{i=1}^m$. For each task T, we also replace its adapted model parameter $\boldsymbol{\theta}_T^q = \boldsymbol{\theta}^* - \alpha[\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1}\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t)]$ in MAML as the adapted parameter $\boldsymbol{\theta}_T^q = \boldsymbol{\mathcal{A}}(\{\boldsymbol{\theta}_i^*\}_{i=1}^m, T) - \alpha\big(\nabla\mathcal{L}_{D_T}(\boldsymbol{\mathcal{A}}(\{\boldsymbol{\theta}_i^*\}_{i=1}^m, T)) + \sum_{t=1}^{q-1}\nabla\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t)\big)$ in TSA-MAML. In this way, by following the proof steps of Theorems 1 and 2, we can prove the desired results in Corollary 1. The proof is completed. □

References

[1] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra. Matching networks for one shot learning. In *Proc. Conf. Neural Information Processing Systems*, pages 3630–3638, 2016. 1

[2] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *Int'l Conf. Learning Representations*, 2017. 1

[3] A. Nichol and J. Schulman. Reptile: a scalable meta-learning algorithm. *arXiv preprint arXiv:1803.02999*, 2, 2018. 1, 2

[4] H. Yao, Y. Wei, J. Huang, and Z. Li. Hierarchically structured meta-learning. In *Proc. Int'l Conf. Machine Learning*, 2019. 1

[5] R. Vuorio, S. Sun, H. Hu, and J. Lim. Multimodal model-agnostic meta-learning via task-aware modulation. In *Proc. Conf. Neural Information Processing Systems*, pages 1–12, 2019. 1

[6] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. Int'l Conf. Machine Learning*, pages 1126–1135, 2017. 1, 2

[7] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *Proc. Conf. Neural Information Processing Systems*, pages 4077–4087, 2017. 1

[8] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *Int'l Conf. Learning Representations*, 2014. 1

[9] I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. In *Int'l Conf. Learning Representations*, 2017. 1

[10] L. Bertinetto, J. Henriques, P. Torr, and A. Vedaldi. Meta-learning with differentiable closed-form solvers. In *Int'l Conf. Learning Representations*, 2019. 2

[11] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. Tenenbaum, H. Larochelle, and R. Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018. 2

[12] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 2

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. Imagenet large scale visual recognition challenge. *Int'l. J. Computer Vision*, 115(3):211–252, 2015. 2

[14] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[15] S. Shalev-Shwartz S. Ben-David. Understanding machine learning: From theory to algorithms. *Cambridge University Press*, 2014. 4