Task Similarity Aware Meta Learning: Theory-inspired Improvement on MAML (Supplementary File)

Pan Zhou* **Yingtian Zou†** **Xiao-Tong Yuan‡** **Jiashi Feng†** **Caiming Xiong*** **Steven Hoi***

*Salesforce Research †National University of Singapore ‡Nanjing University of Information Science & Technology
panzhou3@gmail.com zouyingt@comp.nus.edu.sg xtyuan@nuist.edu.cn
elefjia@nus.edu.sg {cxiong,shoi}@salesforce.com

This supplementary document contains the technical proofs of the results and some additional experimental results of the UAI'21 paper entitled "Task Similarity Aware Meta Learning: Theory-inspired Improvement on MAML". It is structured as follows. Appendix A first provides more experimental details and presents more experimental results, including results on miniImageNet, robustness of initialization number, and comparison between TSA-MAML and MAML using larger model. Appendix B provides more theoretical results of MAML by analyzing its convergence behaviors. Then Appendix C gives the proofs of the main results in Sec. 3.2, including Theorems 1 and 2. Finally, in Appendix D we presents the proofs of Corollary 1 in Sec. 4. Appendix E presents the proof of the convergence of the MAML algorithm in Theorem 4.

A MORE EXPERIMENTAL RESULTS

Results on miniImageNet

Dataset. MiniImageNet [Ravi and Larochelle, 2017] consists of 100 classes from ImageNet [Krizhevsky et al., 2012] and each class contains 600 images of size $84 \times 84 \times 3$. Following [Finn et al., 2017, Nichol and Schulman, 2018], we use the split proposed in [Ravi and Larochelle, 2017], which consists of 64 classes for training, 16 classes for validation and the remaining 20 classes for testing.

Results. From Table 4, 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. These results demonstrate the advantages of TSA-MAML behind which the reasons have been discussed in Sec. 5.1.

Robustness of TSA-MAML to The Number of Initializations

In Sec. 5.2 of the manuscript, we have reported the effects of the initialization number m to the testing performance of TSA-MAML on 10-way 1-shot learning tasks in CIFARFS. The experimental results show that when the value of m ranges from 3 to 11, the performance of our method on 10-way 1-shot learning tasks in CIFARFS are relatively stable. Here we further investigate the robustness of TSA-MAML to the initialization number m on 10-way 5-shot tasks in CIFARFS. From Fig. 5, one can observe that TSA-MAML is very stable when m is selected from $[3, 11]$, which further demonstrates the robustness of TSA-MAML to initialization number m. Similarly, these results also well demonstrate that using vanilla MAML to estimate the optimal model parameters of tasks and then clustering these model parameters according to their distances to the multiple group-specific initializations is valid when m is not very large. All these results are consistent with the results in Sec. 5.2.

Investigation of Losses and Accuracy of MAML in The Adaptation process

In this subsection, we investigate the loss and accuracy in MAML from the learnt initializations along with gradient descent steps. Here we evaluate on CIFARFS and miniImagenet datasets. Specifically, we randomly sample 1000 tasks and compute their losses and classification accuracy with along gradient descent steps. Then we report the average test loss and accuracy of these 1000 tasks. From Fig. 6, one can observe that at the first a few gradient descent steps, the loss of MAML decrease

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

Table 4: Few-shot Classification accuracy (%) of the compared approaches on the miniImageNet 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 [Vinyals et al., 2016]	43.56 ± 0.84	55.31 ± 0.73	17.31 ± 0.22	22.69 ± 0.20
Meta-LSTM [Ravi and Larochelle, 2017]	43.33 ± 0.77	60.60 ± 0.71	16.70 ± 0.23	26.06 ± 0.25
Reptile [Nichol and Schulman, 2018]	47.07 ± 0.26	62.74 ± 0.37	28.40 ± 0.85	45.55 ± 0.88
FOMAML [Finn et al., 2017]	45.53 ± 1.58	61.02 ± 1.12	15.21 ± 0.54	17.67 ± 0.47
MAML [Finn et al., 2017]	48.26 ± 1.76	63.11 ± 1.01	30.29 ± 0.45	45.39 ± 0.44
TSA-MAML	$\mathbf{48.44 \pm 0.91}$	$\mathbf{65.52 \pm 0.68}$	$\mathbf{31.15 \pm 0.42}$	$\mathbf{47.58 \pm 0.44}$
Reptile + Transduction [Nichol and Schulman, 2018]	$\mathbf{49.97} \pm 0.32$	65.99 ± 0.58	30.41 ± 0.83	47.08 ± 0.72
iMAML + Transduction [Rajeswaran et al., 2019]	48.76 ± 0.87	63.56 ± 0.95	24.72 ± 0.51	34.67 ± 0.56
FOMAML + Transduction [Finn et al., 2017]	48.07 ± 1.75	63.15 ± 0.91	23.16 ± 0.94	37.38 ± 0.83
MAML + Transduction [Finn et al., 2017]	48.26 ± 1.84	63.11 ± 0.92	31.31 ± 0.52	46.19 ± 0.33
TSA-MAML + Transduction	49.22 ± 0.92	$\mathbf{66.26 \pm 0.92}$	$\mathbf{32.13 \pm 0.44}$	$\mathbf{48.17 \pm 0.56}$

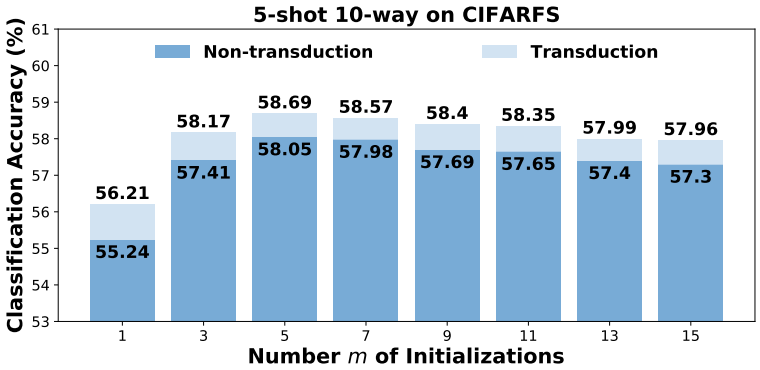


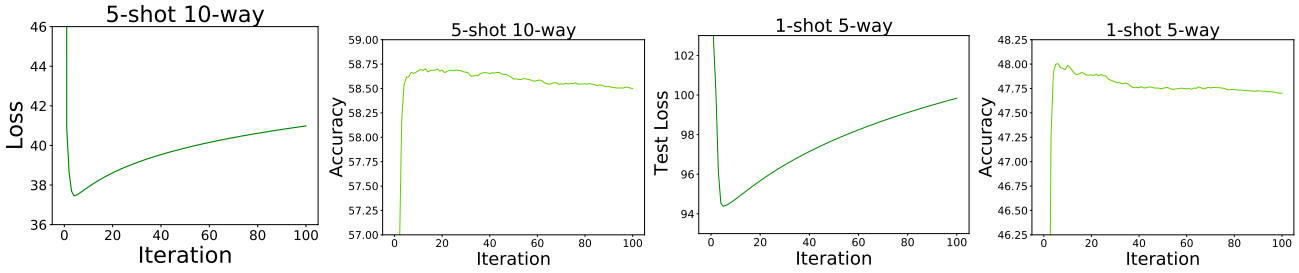
Figure 5: Effects of initialization number to TSA-MAML.

very fast, but with along more optimization gradient steps, it increases. 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. Also, for accuracy, we can also observe very similar phenomena.

Comparison between TSA-MAML and MAML Using Larger Network Model As MAML and TSA-MAML use the same network, they have same parameter dimension and thus same model complexity (data fitting capacity) [S. Ben-David, 2014] though TSA-MAML has multiple initializations. So the advantage of TSA-MAML over MAML comes from its design principle introduced above instead of higher model complexity. Indeed, we also test MAML using larger models (MAML-L). We increase its network depth from four to seven and then increase channels per layer so that new model is about $3\times$ larger than TSA-MAML. The accuracies of MAML-L are 72.68 (aircraft), 69.73 (bird) and 54.18 (fungi) on the first group-structured dataset. By comparison, TSA-MAML still outperforms MAML-L which further testifies our above conclusion. Moreover, 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 5, 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.

Details of Two Group-Structured Datasets

In Sec. 5.1 in the manuscript, we use two group-structured datasets to evaluate the compared meta learning methods. Here



(a) CIFARFS (b) miniImageNet

Figure 6: Illustration of loss decrease and classification accuracy of MAML. One can observe that at the first a few optimization gradient descent steps, the loss of MAML decreases very fast but increases along with optimization steps. Similarly, one can find that MAML increase their accuracy very fast but also suffers from performance degradation along with gradient descent steps due to the over-fitting issue.

Table 5: 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)	$\mathbf{58.21 \pm 0.93}$	$\mathbf{73.52 \pm 0.72}$	$\mathbf{42.18 \pm 0.58}$	$\mathbf{58.69 \pm 0.56}$

we give more details about the splitting of meta-training, meta-validation and meta-testing classes of each sub-dataset.

For the **first group-structured dataset** which consists of the Aircraft dataset [Maji et al., 2013], CUB Birds [Wah et al., 2011] and the FGVCx-Fungi dataset [Maji et al., 2018], we follow [Yao et al., 2019] to split the Aircraft dataset, CUB Birds and the FGVCx-Fungi dataset as follows.

Aircraft dataset. For this dataset, it is fine-grained visual categorization of aircraft. It consists of 102 different kinds of aircraft. Here following [Yao et al., 2019] we select 100 species with 100 images in each species. We split the meta-training, meta-validation and meta-testing classes as 64, 16 and 20 species respectively. In our experiments, all image size is resized to $84 \times 84 \times 3$.

meta-training classes: MD-90, 737-600, A310, An-12, DR-400, Falcon-900, DC-3, Challenger-600, Fokker-70, Cessna-172, 747-400, ERJ-145, Dornier-328, A330-300, A319, Model-B200, E-170, A340-500, BAE-125, Metroliner, 747-300, C-130, DH-82, HawkT1, 727-200, 767-300, DC-10, Spitfire, E-195, BAE-146-300, F-16A-B, Beechcraft-1900, 747-200, Boeing-717, Falcon-2000, 777-300, Cessna-560, DHC-8-100, Cessna-525, 737-200, DC-8, Global-Express, DHC-1, CRJ-200, A340-300, DC-9-30, CRJ900, A320, 737-300, Eurofighter-Typhoon, SR-20, E-190, Saab-340, C-47, Il-76, MD-87, 757-300, DHC-6, Tu-154, 777-200, 767-200, A318, 757-200, A300B4.

meta-validation classes: 737-900, A340-600, 737-800, 737-400, L-1011, A330-200, Gulfstream-V, 737-500, A340-200, ATR-72, MD-11, CRJ-700, EMB-120, Fokker-100, DC-6, 737-700.

meta-testing classes: 707-320, PA-28, Cessna-208, F-A-18, DHC-8-300, ERJ-135, Tornado, BAE-146-200, A321, ATR-42, Saab-2000, Tu-134, Fokker-50, A380, MD-80, Gulfstream-IV, Yak-42, 747-100, 767-400, Embraer-Legacy-600.

CUB Birds dataset. It is a bird image dataset. Specifically, it consists of 11,788 images of 200 bird species. Here following [Yao et al., 2019] we select 100 species with 60 images in each species. We split the meta-training, meta-validation and meta-testing classes as 64, 16 and 20 species respectively. In our experiments, all image size is resized to $84 \times 84 \times 3$.

-meta-training classes : Savannah Sparrow, Dark eyed Junco, Black footed Albatross, Henslow Sparrow, Cape Glossy Starling, Black throated Sparrow, Northern Waterthrush, Hooded Warbler, Baltimore Oriole, Scarlet Tanager, Cerulean Warbler, Downy Woodpecker, Black and white Warbler, Tropical Kingbird, Canada Warbler, Blue Jay, Elegant Tern, Groove billed Ani, Mallard, European Goldfinch, Red breasted Merganser, Geococcyx, Red winged Blackbird, Ringed Kingfisher, Prairie Warbler, Florida Jay, Hooded Oriole, American Redstart, Western Wood Pewee, Sayornis, Myrtle Warbler, Yellow Warbler, Tree Swallow, Rufous Hummingbird, Fish Crow, Bewick Wren, Seaside Sparrow, Vesper Sparrow, American Crow, Eared Grebe, Blue headed Vireo, White necked Raven, Frigatebird, Horned Lark, Tree Sparrow, Red bellied Woodpecker, Pacific Loon, Caspian Tern, Anna Hummingbird, Olive sided Flycatcher, Common Tern, Cedar Waxwing, Great Crested Flycatcher, Blue Grosbeak, White breasted Kingfisher, White eyed Vireo, Purple Finch, Cliff Swallow, Scissor tailed Flycatcher, Harris Sparrow, Western Grebe, Gadwall, American Goldfinch, Pine Warbler.

-meta-validation classes: Mockingbird, Vermilion Flycatcher, Cape May Warbler, Prothonotary Warbler, White crowned Sparrow, Ovenbird, Pomarine Jaeger, Indigo Bunting, Blue winged Warbler, Chipping Sparrow, Horned Grebe, Fox Sparrow, Green Violetear, Nashville Warbler, Least Tern, Marsh Wren.

-meta-testing classes: Rose breasted Grosbeak, Nighthawk, Long tailed Jaeger, Bronzed Cowbird, California Gull, Ivory Gull, Northern Fulmar, Brown Pelican, Ring billed Gull, Great Grey Shrike, White breasted Nuthatch, Mourning Warbler, Sage Thrasher, Horned Puffin, Pied Kingfisher, Shiny Cowbird, Scott Oriole, Red eyed Vireo, Song Sparrow, Winter Wren.

FGVCx-Fungi dataset. It contains about 1,500 wild mushroom species and has more than 100,000 fungi images. Here following [Yao et al., 2019] we select 100 species with 150 images per class. Then we split the meta-training, meta-validation and meta-testing classes as 64, 16 and 20 species respectively. In our experiments, all image size is resized to $84 \times 84 \times 3$.

-meta-training classes: Suillus granulatus, Phaeolus schweinitzii, Cystoderma amianthinum, Pycnoporellus fulgens, Psathyrella candolleana, Meripilus giganteus, Phellinus pomaceus, Laccaria laccata, Laccaria proxima, Amanita excelsa, Ganoderma pfeifferi, Clitopilus prunulus, Agaricus arvensis, Hericium coralloides, Plicatura crispa, Agrocybe praecox, Steccherinum ochraceum, Hypholoma fasciculare, Xerocomellus pruinatus, Xerocomellus chrysenteron, Crepidotus cesatii, Auricularia auricula-judae, Heterobasidion annosum, Entoloma clypeatum, Cortinarius torvus, Mycena tintinnabulum, Laetiporus sulphureus, Datronia mollis, Pholiota squarrosa, Cerioporus squamosus, Tricholoma terreum, Coprinellus micaceus, Cylindrobasidium laeve, Dacrymyces stillatus, Gloeophyllum sepiarium, Lycoperdon perlatum, Hygrophorus pustulatus, Clavulina coralloides, Xerocomus ferrugineus, Cortinarius alboviolaceus, Byssomerulius corium, Boletus edulis, Hymenopellis radicata, Basidioradulum radula, Cortinarius elatior, Schizophyllum commune, Cortinarius malicorius, Suillellus luridus, Ganoderma applanatum, Oligoporus guttulatus, Tubaria furfuracea, Cortinarius largus, Pleurotus ostreatus, Stereum hirsutum, Xylodon raduloides, Peniophora incarnata, Sutorius luridiformis, Flammulina velutipes var. velutipes, Phlebia radiata, Hygrocybe conica, Chlorophyllum olivieri, Armillaria ostoyae, Peniophora quercina, Mycena galericulata

-meta-validation classes: Agaricus impudicus, Daedaleopsis confragosa, Fomitopsis pinicola, Cortinarius anserinus, Mucidula mucida, Trametes versicolor, Stropharia cyanea, Ramaria stricta, Radulomyces confluens, Gliophorus psittacinus, Psathyrella spadiceogrisea, Coprinopsis lagopus, Daedalea quercina, Amanita muscaria, Armillaria lutea, Vuilleminia comedens

-meta-testing classes: Hygrocybe ceracea, Trametes hirsuta, Polyporus tuberaster, Lacrymaria lacrymabunda, Fistulina hepatica, Gymnopus dryophilus, Amanita rubescens, Fuscoporia ferrea, Craterellus undulatus, Tricholoma scalpturatum, Mycena pura, Russula depallens, Bjerkandera adusta, Trametes gibbosa, Tremella mesenterica, Cerioporus varius, Amanita fulva, Xylodon paradoxus, Cuphophyllus virgineus, Cortinarius flexipes

For the **second group-structured dataset** which contains Stanford Car [Krause et al., 2013], CUB Birds [Wah et al., 2011] and FGVCx-Fungi [Maji et al., 2018], we split each sub-dataset as follows.

Stanford Car dataset. The Cars dataset contains 16,185 images of 196 classes of cars. Classes are typically at the level of Make, Model, Year, e.g. 2012 Tesla Model S or 2012 BMW M3 coupe. Here we select 100 species with

80 images in each species. We split the meta-training, meta-validation and meta-testing classes as 64, 16 and 20 species respectively. In our experiments, all image size is resized to $84 \times 84 \times 3$.

meta-training classes: Acura Integra Type R 2001, Acura TL Sedan 2012, Acura TL Type-S 2008, Acura TSX Sedan 2012, AM General Hummer SUV 2000, Aston Martin V8 Vantage Convertible 2012, Aston Martin V8 Vantage Coupe 2012, Audi 100 Wagon 1994, Audi A5 Coupe 2012, Audi R8 Coupe 2012, Audi S4 Sedan 2007, Audi S5 Convertible 2012, Audi S5 Coupe 2012, Audi S6 Sedan 2011, Audi TT Hatchback 2011, Audi TTS Coupe 2012, Audi V8 Sedan 1994, Bentley Continental Flying Spur Sedan 2007, Bentley Continental GT Coupe 2007, BMW 1 Series Coupe 2012, BMW 3 Series Sedan 2012, BMW 3 Series Wagon 2012, BMW 6 Series Convertible 2007, BMW M3 Coupe 2012, BMW M5 Sedan 2010, BMW M6 Convertible 2010, BMW X5 SUV 2007, BMW X6 SUV 2012, BMW Z4 Convertible 2012, Bugatti Veyron 16.4 Coupe 2009, Buick Enclave SUV 2012, Buick Rainier SUV 2007, Cadillac CTS-V Sedan 2012, Cadillac Escalade EXT Crew Cab 2007, Cadillac SRX SUV 2012, Chevrolet Avalanche Crew Cab 2012, Chevrolet Camaro Convertible 2012, Chevrolet Cobalt SS 2010, Chevrolet Corvette ZR1 2012, Chevrolet Impala Sedan 2007, Chevrolet Malibu Sedan 2007, Chevrolet Monte Carlo Coupe 2007, Chevrolet Silverado 1500 Classic Extended Cab 2007, Chevrolet Silverado 1500 Extended Cab 2012, Chevrolet Silverado 1500 Hybrid Crew Cab 2012, Chevrolet Silverado 1500 Regular Cab 2012, Chevrolet Sonic Sedan 2012, Chevrolet TrailBlazer SS 2009, Chevrolet Traverse SUV 2012, Chrysler 300 SRT-8 2010, Chrysler Aspen SUV 2009, Chrysler Crossfire Convertible 2008, Chrysler PT Cruiser Convertible 2008, Chrysler Sebring Convertible 2010, Daewoo Nubira Wagon 2002, Dodge Caliber Wagon 2007, Dodge Caliber Wagon 2012, Dodge Caravan Minivan 1997, Dodge Dakota Crew Cab 2010, Dodge Durango SUV 2012, Dodge Journey SUV 2012, Dodge Magnum Wagon 2008, Dodge Ram Pickup 3500 Crew Cab 2010, Dodge Ram Pickup 3500 Quad Cab 2009

meta-validation classes: Dodge Charger Sedan 2012, Dodge Charger SRT-8 2009, Dodge Durango SUV 2007, Eagle Talon Hatchback 1998, Ferrari 458 Italia Coupe 2012, Ferrari FF Coupe 2012, Fisker Karma Sedan 2012, Ford Edge SUV 2012, Ford Expedition EL SUV 2009, Ford F-150 Regular Cab 2007, Ford F-150 Regular Cab 2012, Ford F-450 Super Duty Crew Cab 2012, Ford Freestar Minivan 2007, Ford GT Coupe 2006, Ford Mustang Convertible 2007, Ford Ranger SuperCab 2011

meta-testing classes: Ford Fiesta Sedan 2012, Ford Focus Sedan 2007, Geo Metro Convertible 1993, GMC Acadia SUV 2012, GMC Canyon Extended Cab 2012, GMC Terrain SUV 2012, GMC Yukon Hybrid SUV 2012, Honda Odyssey Minivan 2007, Honda Odyssey Minivan 2012, HUMMER H2 SUT Crew Cab 2009, Hyundai Azera Sedan 2012, Hyundai Elantra Sedan 2007, Hyundai Elantra Touring Hatchback 2012, Hyundai Genesis Sedan 2012, Hyundai Santa Fe SUV 2012, Hyundai Tucson SUV 2012, Hyundai Veloster Hatchback 2012, Hyundai Veracruz SUV 2012, Isuzu Ascender SUV 2008, Jaguar XK XKR 2012

CUB Birds dataset. We use the same splitting meta-training, meta-validation and meta-testing classes as the first group-structured dataset. Please refer to the materials above.

FGVCx-Fungi dataset. We use the same splitting meta-training, meta-validation and meta-testing classes as the first group-structured dataset. Please refer to the materials above.

B MORE THEORETICAL RESULTS

Here we provide more theoretical results for MAML, namely the convergence analysis. Note the results in this section is not related to the results in the manuscript. Here we provide these results, since we want to make the analysis of MAML more compact and deeper. Now we provide the convergence guarantees for MAML whose formulation is as follows

$$\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) := \mathbb{E}_{T\sim\mathcal{T}}\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta} - \alpha\nabla\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta})), \tag{1}$$

where $\mathcal{L}_{D_T}(\boldsymbol{\theta}_T) = \frac{1}{K}\sum_{(\boldsymbol{x},\boldsymbol{y})\in D_T}\ell(f(\boldsymbol{\theta}_T,\boldsymbol{x}),\boldsymbol{y})$ with $D_T = D_T^{tr}$ or D_T^{ts} is the empirical risk on the dataset D_T, and α is a learning rate. From (1), one can observe that MAML actually approximately solves the inner-optimization problem $\min_{\boldsymbol{\theta}}\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta})$ via one gradient descent step, i.e. $\boldsymbol{\theta}_T = \boldsymbol{\theta} - \alpha\nabla\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta})$, increasing the difficulty of the subsequent convergence analysis. To solve this problem, MAML updates the parameter $\boldsymbol{\theta}$ via the SGD algorithm [Robbins and Monro, 1951]:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \beta\sum\nolimits_{T_i\sim\mathcal{T}}\nabla_{\boldsymbol{\theta}^t}\mathcal{L}_{D_{T_i}^{ts}}(\boldsymbol{\theta}^t - \alpha\nabla\mathcal{L}_{D_{T_i}^{tr}}(\boldsymbol{\theta}^t)), \tag{2}$$

Algorithm 1 Meta Framework for MAML

Input: initial point $\boldsymbol{\theta}^0$, learning rates α and β, task distribution $\mathcal{T}$.
for $t = 0, \cdots, S-1$ **do**
  sample a task mini-batch $\mathcal{S}^t = \{T_i\}_{i=1}^s$ as $T_i \sim \mathcal{T}$.
  for task T_i in $\mathcal{S}^t$ **do**
    compute gradient $\nabla \mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta}^t)$.
    update task-specific parameter $\boldsymbol{\theta}_{T_i}$ as $\boldsymbol{\theta}_{T_i} = \boldsymbol{\theta}^t - \alpha \nabla \mathcal{L}_{D_{T_i}^{tr}}(\boldsymbol{\theta}^t)$ for task T_i.
  end for
  update $\boldsymbol{\theta}^{t+1}$ via Eqn. (2).
end for
Output: $\boldsymbol{\theta}^S$

where $\boldsymbol{\theta}^t$ is a variable at the t-th iteration and β denotes a learning rate. See details in Algorithm 1 in the second page.

Then we formally state our convergence results of Algorithm 1 in Theorem 4.

Theorem 4. ***(Convergence Analysis)*** *Suppose the loss function $\ell(f(\boldsymbol{\theta}, \boldsymbol{x}), \boldsymbol{y})$ is G-Lipschitz continuous, L_s-smooth, L_h-Hessian Lipschitz and third-order differential. Let $\Delta = F(\boldsymbol{\theta}^0) - \min_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$. Then by setting $\alpha \leq \frac{1}{L_s}$ and $\beta = \sqrt{\frac{2\Delta}{L_g L_a^2 G^2 S}}$, the sequence $\{\boldsymbol{\theta}^t\}$ produced by MAML, namely Algorithm 1, obeys*

$$\min_{0 \leq t < S} \mathbb{E}\Big[\big\|\nabla F(\boldsymbol{\theta}^t)\big\|_2^2\Big] \leq \frac{1}{S}\sum_{t=1}^{S} \mathbb{E}\Big[\big\|\nabla F(\boldsymbol{\theta}^t)\big\|_2^2\Big] \leq \sqrt{\frac{2\Delta L_g L_a^2 G^2}{S}},$$

where $L_a = 1 + \alpha L_s$ and $L_g = L_a^2 L_s + \alpha L_h G$ are two constants, and S is the total iteration number in Algorithm 1.

See its proof in Appendix E.2. Theorem 4, the L_h-Hessian Lipschitz function $g(\boldsymbol{\theta})$ means $\|\nabla^2 g(\boldsymbol{\theta}_1) - \nabla^2 g(\boldsymbol{\theta}_2)\|_2 \leq L_h \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$ with a constant L_h for any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

Theorem 4 shows that with along more iterations, the average gradient of the sequence $\{\boldsymbol{\theta}^t\}$ becomes smaller in expectation, and MAML converges at the rate of $\mathcal{O}(\frac{1}{\sqrt{S}})$, even for non-convex problems. Different from recent analysis works on MAML [Golmant, 2019, Finn et al., 2019, Fallah et al., 2019] with strong convexity or infinite mini-batch size assumptions, our theory gets rid of their restrictive assumptions and thus is more realistic.

C PROOF OF THE RESULTS IN SEC. 3.2

C.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 $\boldsymbol{\theta}_T^1 = \operatorname{argmin}_{\boldsymbol{\theta}_T} h_{D_T}(\boldsymbol{\theta}_T) = \boldsymbol{\theta}^* - \alpha \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^*)$. By using Taylor expansion, for $\boldsymbol{\theta}^*$ and any $\boldsymbol{\theta}_T$, there exists a constant $\lambda \in (0, L_s]$ such that

$$\mathcal{L}_{D_T}(\boldsymbol{\theta}_T) = \mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + h_{D_T}(\boldsymbol{\theta}_T) + \frac{1}{2}\left(\lambda - \frac{1}{\alpha}\right)\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^*\|^2. \tag{3}$$

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

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

where in ① we have used the optimality of $\boldsymbol{\theta}_{T,i}^1$ 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

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

Then combining the above two inequalities we arrive at

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

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:*

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

where $\boldsymbol{\theta}_T^t = \boldsymbol{\theta}^* - \alpha\left(\nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + \sum_{s=1}^{t-1} \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^s)\right)$ 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. Then for any q, the following bound holds that

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

and

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

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

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

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

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

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

$$\begin{aligned}
&h_{D_T}(\boldsymbol{\theta}_{T,i}^q) - h_{D_T}(\boldsymbol{\theta}_T^q) \leq \left\langle \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^*) - \nabla \mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1} \left[\nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t) - \nabla \mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_{T,i}^t)\right], \boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q \right\rangle \\
=& \left\langle \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^*) - \nabla \mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1} \left[\nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t) - \nabla \mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_T^t) + \nabla \mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_T^t) - \nabla \mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_{T,i}^t)\right], \boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q \right\rangle \\
\leq& \frac{2qG}{K}\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\| + \sum_{t=1}^{q-1} \left\langle \nabla \mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_T^t) - \nabla \mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_{T,i}^t), \boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q \right\rangle \\
\leq& \frac{2qG}{K}\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\| + \sum_{t=1}^{q-1} \left\|\nabla \mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_T^t) - \nabla \mathcal{L}_{D_T^{(i)}}(\boldsymbol{\theta}_{T,i}^t)\right\| \cdot \left\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\right\| \\
\leq& \frac{2qG}{K}\left\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\right\| + L_s \left\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\right\| \sum_{t=1}^{q-1} \left\|\boldsymbol{\theta}_T^t - \boldsymbol{\theta}_{T,i}^t\right\|.
\end{aligned}$$

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 $(\boldsymbol{x}, \boldsymbol{y}) \sim T$

$$|\ell(f(\boldsymbol{\theta}_{T,i}^q, \boldsymbol{x}), \boldsymbol{y}) - \ell(f(\boldsymbol{\theta}_T^q, \boldsymbol{x}), \boldsymbol{y})| \leq G\|\boldsymbol{\theta}_{T,i}^q - \boldsymbol{\theta}_T^q\| \leq \frac{2G^2\left[(1+2\alpha L_s)^q - 1\right]}{L_s K}. \tag{4}$$

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 (4). 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

$$\|\nabla\ell(f(\boldsymbol{\theta}_T^q, \boldsymbol{x}), \boldsymbol{y}) - \nabla\ell(f(\boldsymbol{\theta}_{T,i}^q, \boldsymbol{x}), \boldsymbol{y})\| \leq L_s\|\boldsymbol{\theta}_T^q - \boldsymbol{\theta}_{T,i}^q\| \leq \frac{2G\left[(1+2\alpha L_s)^q - 1\right]}{K}. \tag{5}$$

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 (5). The proof is concluded. □

C.2 PROOF OF THEOREM 1

Here we provide the 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

$$\left|\mathbb{E}_{D_T\sim T}\left[\mathcal{L}(\boldsymbol{\theta}_T^q)-\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q)\right]\right|\leq\frac{2G^2\left[(1+2\alpha L_s)^q-1\right]}{L_sK}. \tag{6}$$

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

$$\mathbb{E}_{D_T}[\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q)-\mathcal{L}_{D_T}(\boldsymbol{\theta}_T^*)]\leq\frac{1}{2\alpha}\mathbb{E}_{D_T}\left[\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_T^*\|^2\right]. \tag{7}$$

Then we can show the following

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

where ① uses Eqn. (6) and ② employs inequality (7). 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

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

This proves the results in the theorem. □

C.3 PROOF OF THEOREM 2

Here we provide proof of Theorem 2.

Proof. Consider a fixed task $T \sim \mathcal{T}$ and its associated random sample $D_T \sim T$ of size K. Then we consider q gradient steps to obtain the adapted parameter $\boldsymbol{\theta}_T^q = \boldsymbol{\theta}^* - \alpha \left(\nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}^*) + \sum_{t=1}^{q-1} \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t) \right)$. We can show the following inequity

$$
\begin{aligned}
\left\| \mathbb{E}_{D_T} \left[\nabla \mathcal{L}(\boldsymbol{\theta}_T^q) \right] \right\|^2 =& \left\| \mathbb{E}_{D_T} \left[\nabla \mathcal{L}(\boldsymbol{\theta}_T^q) - \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) \right] + \mathbb{E}_{D_T} \left[\nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) \right] \right\|^2 \\
\leq & 2 \left\| \mathbb{E}_{D_T} \left[\nabla \mathcal{L}(\boldsymbol{\theta}_T^q) - \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) \right] \right\|^2 + 2 \left\| \mathbb{E}_{D_T} \left[\nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) \right] \right\|^2 \\
\leq & 2 \left\| \mathbb{E}_{D_T} \left[\nabla \mathcal{L}(\boldsymbol{\theta}_T^q) - \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) \right] \right\|^2 + 2 \mathbb{E}_{D_T} \left[\left\| \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) \right\|^2 \right] \\
\overset{①}{\leq} & \frac{8G^2 \left[(1 + 2\alpha L_s)^q - 1 \right]^2}{K^2} + 2 \mathbb{E}_{D_T} \left[\left\| \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^q) \right\|^2 \right],
\end{aligned}
$$

where in ① we use Lemma 4:

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

This completes the proof. □

D PROOFS OF THE RESULTS IN SEC. 4

Here we presents the proofs of Corollary 1 in Sec. 4.

D.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 $\boldsymbol{\theta}_T^q = \mathcal{A}(\{\boldsymbol{\theta}_i^*\}_{i=1}^m, T) - \alpha \big(\nabla \mathcal{L}_{D_T}(\mathcal{A}(\{\boldsymbol{\theta}_i^*\}_{i=1}^m, T)) + \sum_{t=1}^{q-1} \nabla \mathcal{L}_{D_T}(\boldsymbol{\theta}_T^t) \big)$ 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. □

E PROOFS OF THE RESULTS IN APPENDIX 2

In this section, we provide proofs of the results in Appendix B.

E.1 AUXILIARY LEMMA

Lemma 5. *Under assumptions in both Theorem 4, $F(\boldsymbol{\theta})$ is L_g-smooth, where $L_g = L_a^2 L_s + \alpha L_h G$ and $L_a = (1 + \alpha L_s)$. That is, for any $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$, we have*

$$
F(\boldsymbol{\theta}^1) \leq F(\boldsymbol{\theta}^2) + \left\langle \nabla F(\boldsymbol{\theta}^2), \boldsymbol{\theta}^1 - \boldsymbol{\theta}^2 \right\rangle + \frac{L_g}{2} \|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2\|^2.
$$

Proof. Now we prove $F(\boldsymbol{\theta})$ is L_g-smooth. Specifically, by using Taylor expansion we have

$$
F(\boldsymbol{\theta}^1) = F(\boldsymbol{\theta}^2) + \left\langle \nabla F(\boldsymbol{\theta}^2), \boldsymbol{\theta}^1 - \boldsymbol{\theta}^2 \right\rangle + \int_0^1 \left(\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2 \right)^T \nabla^2 F(\boldsymbol{\theta}) \left(\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2 \right) dt,
$$

where $\boldsymbol{\theta} = t\boldsymbol{\theta}^1 + (1-t)\boldsymbol{\theta}^2$. Let $\boldsymbol{\theta}^T = \boldsymbol{\theta} - \alpha \nabla \mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta}))$. Since the loss function $\ell(f(\boldsymbol{\theta}, \boldsymbol{x}), \boldsymbol{y})$ is L_s-smooth and also L_h-Hessian Lipschitz, then $\mathcal{L}_{D_T}(\boldsymbol{\theta}) = \frac{1}{K} \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in D_T} \ell(f(\boldsymbol{\theta}, \boldsymbol{x}), \boldsymbol{y})$ where $\mathcal{L}_{D_T}(\boldsymbol{\theta})$ could be $\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta})$ and $\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta})$, is

L_s-smooth and also L_h-Hessian Lipschitz. Besides, for $\ell(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})$, it is G-Lipschitz continuous. Therefore, its gradient with respect to $\boldsymbol{\theta}$ can be bounded as $\|\nabla\ell(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})\|\leq G$. Then we also have $\|\nabla\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta})\|\leq G$. Finally we can bound

$$
\begin{aligned}
&\left\|\nabla^2 F(\boldsymbol{\theta})\right\| \\
=&\left\|\mathbb{E}_{T\sim\mathfrak{T}}\left[(\boldsymbol{I}-\alpha\nabla^2\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta}))\nabla^2_{\boldsymbol{\theta}^T}\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta}^T)(\boldsymbol{I}-\alpha\nabla^2\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta}))-\alpha\nabla^3\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta})\otimes\nabla_{\boldsymbol{\theta}^T}\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta}^T)\right]\right\| \\
\leq&\max_{T\sim\mathfrak{T}}\left\|(\boldsymbol{I}-\alpha\nabla^2\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta}))\nabla^2_{\boldsymbol{\theta}^T}\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta}^T)(\boldsymbol{I}-\alpha\nabla^2\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta}))-\alpha\nabla^3\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta})\otimes\nabla_{\boldsymbol{\theta}^T}\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta}^T)\right\| \\
\leq&\max_{T\sim\mathfrak{T}}\left\|\boldsymbol{I}-\alpha\nabla^2\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta})\right\|^2\cdot\left\|\nabla^2_{\boldsymbol{\theta}^T}\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta}^T)\right\|+\alpha\left\|\nabla^3\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta})\right\|\cdot\left\|\nabla_{\boldsymbol{\theta}^T}\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta}^T)\right\| \\
\leq&(1+\alpha L_s)^2L_s+\alpha L_hG:=L_g.
\end{aligned}
$$

So it yields

$$
F(\boldsymbol{\theta}^1)\leq F(\boldsymbol{\theta}^2)+\left\langle\nabla F(\boldsymbol{\theta}^2),\boldsymbol{\theta}^1-\boldsymbol{\theta}^2\right\rangle+\int_0^1 L_g\|\boldsymbol{\theta}^1-\boldsymbol{\theta}^2\|^2dt=F(\boldsymbol{\theta}^2)+\left\langle\nabla F(\boldsymbol{\theta}^2),\boldsymbol{\theta}^1-\boldsymbol{\theta}^2\right\rangle+\frac{L_g}{2}\|\boldsymbol{\theta}^1-\boldsymbol{\theta}^2\|^2.
$$

The proof is completed. □

E.2 PROOF OF THEOREM 4

Here we provide proof of Theorem 4.

Proof. Here we prove the convergence of the MAML algorithm. For brevity, we let

$$
\boldsymbol{v}_t=\sum_{T_i\sim\mathfrak{T}}(\boldsymbol{I}-\alpha\nabla^2_{\boldsymbol{\theta}^t}\mathcal{L}_{D_{T_i}^{tr}}(\boldsymbol{\theta}^t))\nabla_{\boldsymbol{\theta}_{T_i}}\mathcal{L}_{D_{T_i}^{ts}}(\boldsymbol{\theta}_{T_i})
$$

denote the gradient of the sampled mini-batch. Then by using Lemma 5, we can obtain

$$
\begin{aligned}
\mathbb{E}\left[F(\boldsymbol{\theta}^{t+1})\right]=&\mathbb{E}\left[F\Big(\boldsymbol{\theta}^t-\beta_t\boldsymbol{v}_t\Big)\right] \\
\leq&\mathbb{E}\Big[F(\boldsymbol{\theta}^t)-\beta_t\left\langle\nabla F(\boldsymbol{\theta}^t),\boldsymbol{v}_t\right\rangle+\frac{\beta_t^2L_g}{2}\left\|\boldsymbol{v}_t\right\|^2\Big] \\
=&\mathbb{E}\Big[F(\boldsymbol{\theta}^t)-\beta_t\|\nabla F(\boldsymbol{\theta}^t)\|^2+\frac{\beta_t^2L_g}{2}\left\|\boldsymbol{v}_t\right\|^2\Big].
\end{aligned}
$$

Then we rearrange the above inequality and obtain

$$
\mathbb{E}\left[\left\|\nabla F(\boldsymbol{\theta}^t)\right\|^2\right]\leq\frac{1}{\beta_t}\mathbb{E}\left[F(\boldsymbol{\theta}^t)-F(\boldsymbol{\theta}^{t+1})\right]+\frac{\beta_tL_g}{2}\mathbb{E}[\|\boldsymbol{v}_t\|^2]. \tag{8}
$$

Since the loss function $\ell(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})$ is L_s-smooth and also L_h-Hessian Lipschitz, then $\mathcal{L}_{D_T}(\boldsymbol{\theta})=\frac{1}{K}\sum_{(\boldsymbol{x},\boldsymbol{y})\in D_T}\ell(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})$ where $\mathcal{L}_{D_T}(\boldsymbol{\theta})$ could be $\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta})$ and $\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta})$, is L_s-smooth and also L_h-Hessian Lipschitz. Besides, for $\ell(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})$, it is G-Lipschitz continuous. Therefore, its gradient with respect to $\boldsymbol{\theta}$ can be bounded as $\|\nabla\ell(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})\|\leq G$. Then we also have $\|\nabla\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta})\|\leq G$. So we can further establish

$$
\begin{aligned}
\mathbb{E}[\|\boldsymbol{v}_t\|^2]=&\mathbb{E}\left[\left\|\sum_{T_i\sim\mathfrak{T}}(\boldsymbol{I}-\alpha\nabla^2_{\boldsymbol{\theta}^t}\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta}^t))\nabla_{\boldsymbol{\theta}_{T_i}}\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta}_{T_i})\right\|^2\right] \\
\leq&\max_{T_i}\left\|(\boldsymbol{I}-\alpha\nabla^2_{\boldsymbol{\theta}^t}\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta}^t))\nabla_{\boldsymbol{\theta}_{T_i}}\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta}_{T_i})\right\|^2 \\
\leq&\max_{T_i}\left\|(\boldsymbol{I}-\alpha\nabla^2_{\boldsymbol{\theta}^t}\mathcal{L}_{D_T^{tr}}(\boldsymbol{\theta}^t))\right\|^2\cdot\left\|\nabla_{\boldsymbol{\theta}_{T_i}}\mathcal{L}_{D_T^{ts}}(\boldsymbol{\theta}_{T_i})\right\|^2 \\
\overset{①}{\leq}&L_a^2G^2,
\end{aligned}
$$

where $L_a=1+\alpha L_s$.

By summing up Eqn. (8) from $t=0$ to $t=S-1$ and setting $\beta_t=\beta$, we can establish

$$\min_t \mathbb{E}\left[\left\|\nabla F(\boldsymbol{\theta}^t)\right\|^2\right] \leq \frac{1}{S}\sum_{t=0}^{S-1}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{\theta}^t)\right\|^2\right] \leq \frac{1}{S\beta}\mathbb{E}\left[F(\boldsymbol{\theta}^0)-F(\boldsymbol{\theta}^S)\right]+\frac{\beta L_g}{2}L_a^2G^2 \overset{①}{=} \sqrt{\frac{2\Delta L_g L_a^2 G^2}{S}},$$

where ① holds by setting $\beta=\sqrt{\frac{2\Delta}{L_g L_a^2 G^2 S}}$. In this way, one can observe that the convergence rate is $\mathcal{O}\left(1/\sqrt{S}\right)$. The proof is completed. □

References

A. Fallah, A. Mokhtari, and A. Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. *arXiv preprint arXiv:1908.10400*, 2019.

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.

C. Finn, A. Rajeswaran, S. Kakade, and S. Levine. Online meta-learning. *arXiv preprint arXiv:1902.08438*, 2019.

N. Golmant. On the convergence of model-agnostic meta-learning. *http://noahgolmant.com/writings/maml.pdf*, 2019.

J. Krause, M. Stark, J. Deng, and F. Li. 3D object representations for fine-grained categorization. In *Int'l IEEE Workshop on 3D Representation and Recognition*, 2013.

A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. 2018 FGCVx fungi classification challenge. *fungi-challenge-fgvc-2018*, 2018.

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

A. Rajeswaran, C. Finn, S. Kakade, and S. Levine. Meta-learning with implicit gradients. In *Proc. Conf. Neural Information Processing Systems*, pages 113–124, 2019.

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

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

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

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.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

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