Supplementary Material for "Improving Generalization of Meta-learning with Inverted Regularization at Inner-level"

Due to the space limitation of the main paper, we provide supplementary theoretical proof and supplementary experimental results in this Appendix, including: more detailed theoretical analyses, more experiment details, an additional experiment on Mini-ImageNet few-shot classification with limited tasks, an additional experiment on Meta-dataset with the first-order method and larger backbone, and an additional experiment on meta-reweighting with Minimax-Meta Regularization for robust learning.

A. Theoretical Analysis

In this section, we provide detailed proof derivations of the theoretical results in the main paper.

A.1. Lemmas

This section lists the Lemmas that help prove our main results.

Lemma 1. *(from [6]) Let ϕ be a λ-strongly convex and η-smooth function. Then, for any $\beta \leq \frac{2}{\lambda+\eta}$, we have*

$$\|(u - \beta\nabla\phi(u)) - (v - \beta\nabla\phi(v))\| \leq \left(1 - \frac{\beta\lambda\eta}{\lambda+\eta}\right)\|u - v\|$$

for any u and v.

Lemma 2. *(based on [5]) Let $f : \mathbb{R}^d \to \mathbb{R}$ be an function that is L-smooth, μ-strongly convex, and has gradient bounded by G. Consider a function $\boldsymbol{U}(\cdot)$ that describes the MAML inner-level update rule, with L2-Norm regularization parameterized by $\frac{\delta}{2}$: $\boldsymbol{U}(\boldsymbol{w}) = w - \alpha\nabla_w(f(w) + \frac{\delta}{2}\|w\|^2)$), with $\alpha \leq \frac{1}{2L}$, $\delta < \frac{1}{2\alpha}$. Then,*

$$\|\boldsymbol{U}(\boldsymbol{w}) - \boldsymbol{U}(\boldsymbol{v})\| \leq (1 - \alpha\delta - \alpha\mu)\|\boldsymbol{w} - \boldsymbol{v}\| \quad \forall \boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^d$$

Proof. Firstly, note that

$$\boldsymbol{U}(\boldsymbol{w}) = w - \alpha(\nabla f(w) + \delta w) = (1 - \alpha\delta)w - \alpha\nabla f(w)$$

The Jacobian of $\boldsymbol{U}(\cdot)$ is given by $\nabla\boldsymbol{U}(\boldsymbol{w}) = (1 - \alpha\delta)\boldsymbol{I} - \alpha\nabla^2 f(\boldsymbol{w})$.
Like in [5], we use $\boldsymbol{A} \succeq \boldsymbol{0}$ to denote the positive semi-definite nature of the matrix. Similarly, $\boldsymbol{A} \succeq \boldsymbol{B}$ means that $\boldsymbol{\theta}^T(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{\theta} \geq 0 \forall \boldsymbol{\theta}$. Since f is μ-strongly convex, and L-smooth, we could have $\mu\boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{w}) \preceq L\boldsymbol{I} \ \ \forall \boldsymbol{w} \in \mathbb{R}^d$. Then the Jacobian can be bounded by

$$(1 - \alpha\delta - \alpha L)\boldsymbol{I} \preceq \nabla\boldsymbol{U}(\boldsymbol{w}) \preceq (1 - \alpha\delta - \alpha\mu)\boldsymbol{I} \quad \forall \boldsymbol{w} \in \mathbb{R}^d$$

The upper bound implies $\|\nabla\boldsymbol{U}(\boldsymbol{w})\| \leq (1 - \alpha\mu - \alpha\delta) \ \ \forall \boldsymbol{w} \in \mathbb{R}^d$.
Let $\boldsymbol{\psi}(t) = \boldsymbol{v} + t(\boldsymbol{w} - \boldsymbol{v}), t \in [0, 1]$ be the line function connecting $\boldsymbol{w}$ and $\boldsymbol{v}$. Taking the line integral, we have

$$\begin{aligned}
\boldsymbol{U}(\boldsymbol{w}) - \boldsymbol{U}(\boldsymbol{v}) &= \int_{t=0}^{t=1} \nabla\boldsymbol{U}(\boldsymbol{\psi}(t))d\boldsymbol{\psi}(t) \\
&= \int_{t=0}^{t=1} \nabla\boldsymbol{U}(\boldsymbol{\psi}(t))\frac{d\boldsymbol{\psi}(t)}{dt}dt \\
&= \int_{t=0}^{t=1} \nabla\boldsymbol{U}(\boldsymbol{\psi}(t))(\boldsymbol{w} - \boldsymbol{v})dt \\
&= \left(\int_{t=0}^{t=1} \nabla\boldsymbol{U}(\boldsymbol{\psi}(t))dt\right)(\boldsymbol{w} - \boldsymbol{v})
\end{aligned}$$

Using the Cauchy-Schwartz inequality and $\|\nabla \boldsymbol{U}(\boldsymbol{w})\| \leq (1-\alpha\mu-\alpha\delta) \ \forall \boldsymbol{w}$, we have

$$\begin{aligned}
\|\boldsymbol{U}(\boldsymbol{w})-\boldsymbol{U}(\boldsymbol{v})\| &= \left\| \int_{t=0}^{t=1} \nabla \boldsymbol{U}(\boldsymbol{\psi}(t))(\boldsymbol{w}-\boldsymbol{v}) dt \right\| \\
&\leq \int_{t=0}^{t=1} \|\nabla \boldsymbol{U}(\boldsymbol{\psi}(t))(\boldsymbol{w}-\boldsymbol{v})\| dt \\
&\leq \int_{t=0}^{t=1} \|\nabla \boldsymbol{U}(\boldsymbol{\psi}(t))\| \|(\boldsymbol{w}-\boldsymbol{v})\| dt \\
&\leq \int_{t=0}^{t=1} (1-\alpha\mu-\alpha\delta)\|\boldsymbol{w}-\boldsymbol{v}\| dt \\
&= (1-\alpha\mu-\alpha\delta)\|\boldsymbol{w}-\boldsymbol{v}\| \int_{t=0}^{t=1} dt \\
&= (1-\alpha\mu-\alpha\delta)\|\boldsymbol{w}-\boldsymbol{v}\|
\end{aligned}$$

□

A.2. Main Results

We start by restating the assumptions we use to derive the results, and then we move on to prove our main results.

Assumption 1. *We assume the function $\ell(\cdot, z)$ satisfies the following properties for any $z \in \mathcal{Z}$:*

1. *(Strong convexity) $\ell(\cdot, z)$ is μ-strongly convex, i.e., $(\nabla \ell(w,z) - \nabla \ell(u,z))^T (w-u) \geq \mu \|w-u\|^2$;*
2. *(Lipschitz in function value) $\ell(\cdot, z)$ has gradients with norm bounded by G, i.e., $\|\nabla \ell(w,z)\| \leq G$;*
3. *(Lipschitz gradient) $\ell(\cdot, z)$ is L-smooth, i.e., $\|\nabla \ell(w,z) - \nabla \ell(u,z)\| \leq L\|w-u\|$;*
4. *(Lipschitz Hessian) $\ell(\cdot, z)$ has ρ-Lipschitz Hessian, i.e., $\left\|\nabla^2 \ell(w,z) - \nabla^2 \ell(u,z)\right\| \leq \rho \|w-u\|$*

Assumption 2. *We assume $\mathcal{F}_{\mathcal{Z}}$ is the Borel σ-algebra over $\mathcal{Z}$ and $\mathcal{Z}$ is a Polish space. And each p_i is a non-atomic distribution over $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$.*

A.2.1 Strongly Convexity and Smoothness

Lemma 3. *(based on [5]) Suppose f and $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ satisfy assumptions 1. We formulate the MAML's outer-level evaluation function with inner-level L2-Norm regularization parameterized by $\frac{\delta}{2}$ with f and $\hat{f}$, and let $\tilde{f}$ be the function evaluated after a one-step gradient update procedure, i.e.,*

$$\tilde{f}(w) := f(w - \alpha \nabla_w (\hat{f}(w) + \frac{\delta}{2}\|w\|^2))$$

then, with $\alpha < \frac{1}{2L}$, $\delta < \frac{1}{2\alpha}$ and $\frac{\alpha\rho G}{\mu} < (\frac{1}{2} - \alpha L)^2$, $\tilde{f}(\cdot)$ is $(-\alpha\rho G + (1-\alpha\delta-\alpha L)^2\mu)$ strongly convex and $(\alpha\rho G + (1-\alpha\delta-\alpha\mu)^2 L)$ smooth.

Proof. Let $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^d$ be two arbitrary points. Let $\boldsymbol{U}(\boldsymbol{w}) = w - \alpha \nabla_w (\hat{f}(w) + \frac{\delta}{2}\|w\|^2))$. Note that

$$\begin{aligned}
\boldsymbol{U}(\boldsymbol{w}) &= w - \alpha(\nabla \hat{f}(w) + \delta w) \\
&= (1-\alpha\delta)w - \alpha \nabla \hat{f}(w)
\end{aligned}$$

We use shorthand of $\tilde{\boldsymbol{w}} \equiv \boldsymbol{U}(\boldsymbol{w}), \tilde{\boldsymbol{v}} \equiv \boldsymbol{U}(\boldsymbol{v})$. Using the chain rule we could have

$$\begin{aligned}
\nabla \tilde{f}(\boldsymbol{w}) - \nabla \tilde{f}(\boldsymbol{v}) &= \nabla \boldsymbol{U}(\boldsymbol{w}) \nabla f(\tilde{\boldsymbol{w}}) - \nabla \boldsymbol{U}(\boldsymbol{v}) \nabla f(\tilde{\boldsymbol{v}}) \\
&= (\nabla \boldsymbol{U}(\boldsymbol{w}) - \nabla \boldsymbol{U}(\boldsymbol{v})) \nabla f(\tilde{\boldsymbol{w}}) + \nabla \boldsymbol{U}(\boldsymbol{v})(\nabla f(\tilde{\boldsymbol{w}}) - \nabla f(\tilde{\boldsymbol{v}}))
\end{aligned}$$

We first move towards the smoothness property. Taking the norm on both sides, based on triangle inequality, we have:

$$\begin{aligned}
\|\nabla \tilde{f}(\boldsymbol{w}) - \nabla \tilde{f}(\boldsymbol{v})\| &= \|(\nabla \boldsymbol{U}(\boldsymbol{w}) - \nabla \boldsymbol{U}(\boldsymbol{v})) \nabla f(\tilde{\boldsymbol{w}}) + \nabla \boldsymbol{U}(\boldsymbol{v})(\nabla f(\tilde{\boldsymbol{w}}) - \nabla f(\tilde{\boldsymbol{v}}))\| \\
&\leq \|(\nabla \boldsymbol{U}(\boldsymbol{w}) - \nabla \boldsymbol{U}(\boldsymbol{v})) \nabla f(\tilde{\boldsymbol{w}})\| + \|\nabla \boldsymbol{U}(\boldsymbol{v})(\nabla f(\tilde{\boldsymbol{w}}) - \nabla f(\tilde{\boldsymbol{v}}))\|
\end{aligned} \tag{7}$$

We could bound the first term on the RHS by

$$\begin{aligned}\|(\nabla \boldsymbol{U}(\boldsymbol{w}) - \nabla \boldsymbol{U}(\boldsymbol{v}))\nabla f(\tilde{\boldsymbol{w}})\| &\overset{(a)}{\leq} \|\nabla \boldsymbol{U}(\boldsymbol{w}) - \nabla \boldsymbol{U}(\boldsymbol{v})\|\|\nabla f(\tilde{\boldsymbol{w}})\| \\ &= \left\|\left((1-\alpha\delta)\boldsymbol{I} - \alpha\nabla^2 \hat{f}(\boldsymbol{w})\right) - \left((1-\alpha\delta)\boldsymbol{I} - \alpha\nabla^2 \hat{f}(\boldsymbol{v})\right)\right\| \|\nabla f(\tilde{\boldsymbol{w}})\| \\ &= \alpha \left\|\nabla^2 \hat{f}(\boldsymbol{w}) - \nabla^2 \hat{f}(\boldsymbol{v})\right\| \|\nabla f(\tilde{\boldsymbol{w}})\| \\ &\overset{(b)}{\leq} \alpha\rho\|\boldsymbol{w} - \boldsymbol{v}\|\|\nabla f(\tilde{\boldsymbol{w}})\| \\ &\overset{(c)}{\leq} \alpha\rho G\|\boldsymbol{w} - \boldsymbol{v}\|\end{aligned} \tag{8}$$

where (a) is due to Cauchy-Schwarz inequality, (b) is due to the Hessian Lipschitz property, and (c) is due to bounded gradient assumption. Similarly, we could bound the second term on (7)'s RHS by

$$\begin{aligned}\|\nabla \boldsymbol{U}(\boldsymbol{v})(\nabla f(\tilde{\boldsymbol{w}}) - \nabla f(\tilde{\boldsymbol{v}}))\| &= \left\|\left((1-\alpha\delta)\boldsymbol{I} - \alpha\nabla^2 \hat{f}(\boldsymbol{v})\right)(\nabla f(\tilde{\boldsymbol{w}}) - \nabla f(\tilde{\boldsymbol{v}}))\right\| \\ &\overset{(a)}{\leq} (1-\alpha\delta-\alpha\mu)\|\nabla f(\tilde{\boldsymbol{w}}) - \nabla f(\tilde{\boldsymbol{v}})\| \\ &\overset{(b)}{\leq} (1-\alpha\delta-\alpha\mu)L\|\tilde{\boldsymbol{w}} - \tilde{\boldsymbol{v}}\| \\ &\overset{(c)}{=} (1-\alpha\delta-\alpha\mu)L\|\boldsymbol{U}(\boldsymbol{w}) - \boldsymbol{U}(\boldsymbol{v})\| \\ &\overset{(d)}{\leq} (1-\alpha\delta-\alpha\mu)L(1-\alpha\delta-\alpha\mu)\|\boldsymbol{w} - \boldsymbol{v}\| \\ &= (1-\alpha\delta-\alpha\mu)^2 L\|\boldsymbol{w} - \boldsymbol{v}\|\end{aligned} \tag{9}$$

Here, (a) is due to $(1-\alpha\delta)\boldsymbol{I} - \alpha\nabla^2\hat{f}(v)$ being symmetric, semi-positive definite, and $\lambda_{\max}\left((1-\alpha\delta)\boldsymbol{I} - \alpha\nabla^2\hat{f}(v)\right) \leq (1-\alpha\delta) - \alpha\mu$ (see Lemma 2). Step (b) is due to $f(\cdot)$ is L-smooth. Step (c) is the use of short hand $\tilde{\boldsymbol{w}} \equiv \boldsymbol{U}(\boldsymbol{w}), \tilde{\boldsymbol{v}} \equiv \boldsymbol{U}(\boldsymbol{v})$. Finally, step (d) is achieved by using Lemma 2 on $\boldsymbol{U}(\cdot)$. Put the result of (8) and (9) into (7), we have

$$\begin{aligned}\|\nabla \tilde{f}(\boldsymbol{w}) - \nabla \tilde{f}(\boldsymbol{v})\| &\leq \|(\nabla \boldsymbol{U}(\boldsymbol{w}) - \nabla \boldsymbol{U}(\boldsymbol{v}))\nabla f(\tilde{\boldsymbol{w}})\| + \|\nabla \boldsymbol{U}(\boldsymbol{v})(\nabla f(\tilde{\boldsymbol{w}}) - \nabla f(\tilde{\boldsymbol{v}}))\| \\ &\leq \alpha\rho G\|\boldsymbol{w} - \boldsymbol{v}\| + (1-\alpha\delta-\alpha\mu)^2 L\|\boldsymbol{w} - \boldsymbol{v}\| \\ &= \left(\alpha\rho G + (1-\alpha\delta-\alpha\mu)^2 L\right)\|\boldsymbol{w} - \boldsymbol{v}\|\end{aligned}$$

and thus $\tilde{f}(\cdot)$ is $\alpha\rho G + (1-\alpha\delta-\alpha\mu)^2 L$ smooth.

Similarly, we first use triangle inequality to find the lower bound.

$$\begin{aligned}\|\nabla \tilde{f}(\boldsymbol{w}) - \nabla \tilde{f}(\boldsymbol{v})\| &= \|(\nabla \boldsymbol{U}(\boldsymbol{w}) - \nabla \boldsymbol{U}(\boldsymbol{v}))\nabla f(\tilde{\boldsymbol{w}}) + \nabla \boldsymbol{U}(\boldsymbol{v})(\nabla f(\tilde{\boldsymbol{w}}) - \nabla f(\tilde{\boldsymbol{v}}))\| \\ &\geq \|\nabla \boldsymbol{U}(\boldsymbol{v})(\nabla f(\tilde{\boldsymbol{w}}) - \nabla f(\tilde{\boldsymbol{v}}))\| - \|(\nabla \boldsymbol{U}(\boldsymbol{w}) - \nabla \boldsymbol{U}(\boldsymbol{v}))\nabla f(\tilde{\boldsymbol{w}})\|\end{aligned}$$

The second term on RHS has already been derived in (8). For the first term, we could bound it by

$$\begin{aligned}\|\nabla \boldsymbol{U}(\boldsymbol{v})(\nabla f(\tilde{\boldsymbol{w}}) - \nabla f(\tilde{\boldsymbol{v}}))\| &= \left\|\left((1-\alpha\delta)\boldsymbol{I} - \alpha\nabla^2 \hat{f}(\boldsymbol{v})\right)(\nabla f(\tilde{\boldsymbol{w}}) - \nabla f(\tilde{\boldsymbol{v}}))\right\| \\ &\overset{(a)}{\geq} (1-\alpha\delta-\alpha L)\|\nabla f(\tilde{\boldsymbol{w}}) - \nabla f(\tilde{\boldsymbol{v}})\| \\ &\overset{(b)}{\geq} (1-\alpha\delta-\alpha L)\mu\|\tilde{\boldsymbol{w}} - \tilde{\boldsymbol{v}}\| \\ &= (1-\alpha\delta-\alpha L)\mu\|(1-\alpha\delta)\boldsymbol{w} - \alpha\nabla\hat{f}(\boldsymbol{w}) - (1-\alpha\delta)\boldsymbol{v} + \alpha\nabla\hat{f}(\boldsymbol{v})\| \\ &\geq \mu(1-\alpha\delta-\alpha L)((1-\alpha\delta)\|\boldsymbol{w} - \boldsymbol{v}\| - \alpha\|\nabla\hat{f}(\boldsymbol{w}) - \nabla\hat{f}(\boldsymbol{v})\|) \\ &\overset{(c)}{\geq} \mu(1-\alpha\delta-\alpha L)((1-\alpha\delta)\|\boldsymbol{w} - \boldsymbol{v}\| - \alpha L\|\boldsymbol{w} - \boldsymbol{v}\|) \\ &\geq \mu(1-\alpha\delta-\alpha L)^2\|\boldsymbol{w} - \boldsymbol{v}\|\end{aligned}$$

Here (a) is due to $\lambda_{\min}\left(I-\alpha\delta-\alpha\nabla^2\hat{f}(v)\right)\geq 1-\alpha\delta-\alpha L$, (b) is due to $f(\cdot)$ being μ-strongly convex, and (c) is due to $\hat{f}(\cdot)$ being L-smooth. Put the results together, we have that

$$\begin{aligned}\|\nabla\tilde{f}(\boldsymbol{w})-\nabla\tilde{f}(\boldsymbol{v})\| &\geq \|\nabla\boldsymbol{U}(\boldsymbol{v})(\nabla f(\tilde{\boldsymbol{w}})-\nabla f(\tilde{\boldsymbol{v}}))\|-\|(\nabla\boldsymbol{U}(\boldsymbol{w})-\nabla\boldsymbol{U}(\boldsymbol{v}))\nabla f(\tilde{\boldsymbol{w}})\| \\ &\geq \left(\mu(1-\alpha\delta-\alpha L)^2-\alpha\rho G\right)\|\boldsymbol{w}-\boldsymbol{v}\|\end{aligned}$$

Thus the function $\tilde{f}(\cdot)$ is $\mu(1-\alpha\delta-\alpha L)^2-\alpha\rho G$ strongly convex. $\mu(1-\alpha\delta-\alpha L)^2-\alpha\rho G$ is positive since $\frac{\alpha\rho G}{\mu}<(\frac{1}{2}-\alpha L)^2$. □

A.2.2 Generalization Bound

Algorithm 2 MAML [4] (the Original Algorithm without Regularization)

Require: Datasets $\mathcal{S}=\left\{\mathcal{S}_i^{\text{in}},\mathcal{S}_i^{\text{out}}\right\}_{i=1}^m$; few-shot meta-query batch size K ; the number of training tasks sampled at each round r; the total number of iterations T.

1: Initialize the model parameters w^0.
2: **for** $t=0$ to $T-1$ **do**
3: Randomly sample r tasks from the set of m available tasks with indices stored in $\mathcal{B}_t$.
4: **for** each sampled task $\mathcal{T}_i$ **do**
5: Sample a size K support data batch $\mathcal{D}_i^{t,\text{ in}}$ from $\mathcal{S}_i^{\text{in}}$;
6: Sample a size b query data batch $\mathcal{D}_i^{t,\text{ out}}$ from $\mathcal{S}_i^{\text{out}}$;
7: Calculate $w_i^{t+1}:=w^t-\beta_t\nabla_{w^t}\hat{\mathcal{L}}\left(w^t-\alpha\nabla\hat{\mathcal{L}}\left(w^t,\mathcal{D}_i^{t,\text{ in}}\right),\mathcal{D}_i^{t,\text{ out}}\right)$;
8: **end for**
9: Meta-update $w^{t+1}:=\frac{1}{r}\sum_{i\in\mathcal{B}_t}w_i^{t+1}$
10: **end for**
11: **Return:** w^T

This section provides the derivation of the generalization bound of MAML with inner-level L2-Norm regularization. The proofs are based on the derivation framework based on algorithm stability proposed by [3]. The framework's preliminary is consistent with this work and shares the same assumptions as this work. To include the notations used in this section, we provide a restatement of the unregularized MAML steps in Algorithm 2.

We restate some notations for a clearer explanation. In the following of this appendix, we use the *hat* superscript to distinguish *empirical losses* from *population losses*. And we use the *tilde* superscript to denote the functions, algorithms, or processes *involving the inner-level regularization*, e.g.,

$$\begin{aligned}\hat{\tilde{F}}_i(w,\mathcal{S}_i) &:= \frac{1}{\binom{n}{K}}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\ |\mathcal{D}_i^{\text{in}}|=K}}\hat{\mathcal{L}}\left(w-\alpha\nabla_w\hat{\mathcal{L}}((w,\mathcal{D}_i^{\text{in}})+\frac{\delta}{2}\|w\|^2),\mathcal{S}_i^{\text{out}}\right) \\ &= \frac{1}{\binom{n}{K}}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\ |\mathcal{D}_i^{\text{in}}|=K}}\frac{1}{n}\sum_{z\in\mathcal{S}_i^{\text{out}}}\ell\left(w-\frac{\alpha}{K}\sum_{z'\in\mathcal{D}_i^{\text{in}}}\nabla_w\left(\ell(w,z')+\frac{\delta}{2}\|w\|^2\right),z\right)\end{aligned}$$

Similarly, $\tilde{F}(\cdot)$ and $\tilde{F}_i(\cdot,\mathcal{S}_i)$ are corresponding to functions with inner-level regularization, distinguished from $F(\cdot)$ and $F_i(\cdot,\mathcal{S}_i)$ corresponding to unregularized MAML. $\tilde{\mathcal{A}}$ also refers to the algorithm (MAML) with inner-level regularization, with output $\tilde{\mathcal{A}}(\mathcal{S})$.

Our goal is to bound

$$\mathbb{E}_{\tilde{\mathcal{A}},\mathcal{S}}[F(\tilde{\mathcal{A}}(\mathcal{S}))-\hat{F}(\tilde{\mathcal{A}}(\mathcal{S}),\mathcal{S})] \tag{10}$$

Which is the expected discrepancy between population loss and empirical loss. Note that the loss is evaluated using the original MAML's unregularized inner-updating rule at the test time, while the model $\tilde{\mathcal{A}}(\mathcal{S})$ is generated by MAML with inner-level regularization.

Then, we are going to bound (10) using the algorithm-stability-based framework proposed by [3]. We first include definitions and the key lemma of the framework.

Definition 1. *(**symmetric algorithm**)We define an algorithm $\mathcal{A}: \mathcal{Z}^n \to \mathbb{R}^d$ to be symmetric if its output distribution, denoted by $\mathcal{A}(\mathcal{S})$, remains unchanged under any permutation of the input set $\mathcal{S} \subset \mathcal{Z}^n$. In other words, if we take another set $\mathcal{S}'$ that is a permutation of $\mathcal{S}$, the distributions of $\mathcal{A}(\mathcal{S})$ and $\mathcal{A}(\mathcal{S}')$ would be similar.*

Definition 2. *((γ, K)-**uniformly stability**, from [3]) Consider the problem in (2) of the main paper, and let $\mathcal{A}$ be a randomized algorithm that produces output $w_{\mathcal{S}}$ given dataset $\mathcal{S}$. We say that $\mathcal{A}$ is (γ, K)-uniformly stable if the following condition holds: for any $i \in \{1, \dots m\}$, let $\bar{\mathcal{S}}$ be a dataset that is identical to $\mathcal{S}$ except that $\bar{\mathcal{S}}_i^{in}$ and $\bar{\mathcal{S}}_i^{out}$ differ from $\mathcal{S}_i^{in}$ and $\mathcal{S}_i^{out}$, respectively, in at most K and one data points. For any $\bar{z} \in \mathcal{Z}$ and any set of K distinct points $\{z_1, \dots, z_K\}$ in $\mathcal{Z}$,*

$$\mathbb{E}_{\mathcal{A}}\left[\left|\ell\left(w_{\mathcal{S}} - \alpha\nabla\hat{\mathcal{L}}\left(w_{\mathcal{S}}, \{z_j\}_{j=1}^K\right), \bar{z}\right) - \ell\left(w_{\bar{\mathcal{S}}} - \alpha\nabla\hat{\mathcal{L}}\left(w_{\bar{\mathcal{S}}}, \{z_j\}_{j=1}^K\right), \bar{z}\right)\right|\right] \leq \gamma$$

the expectation is with respect to the randomness of $\mathcal{A}$.

Lemma 4. *(**stability and generalization error**, from [3]) Let F and $\hat{F}$ be the population and empirical losses defined in Equations (1) and (2) of the main paper, respectively. Suppose Assumption 2 holds and let $\mathcal{A}$ be a (possibly randomized) symmetric and (γ, K)-uniformly stable algorithm that produces output $w_{\mathcal{S}} \in \mathcal{W}$ given input dataset $\mathcal{S}$. Then, the expected difference between the population loss and the empirical loss of $w_{\mathcal{S}}$ is bounded by γ, i.e., $\mathbb{E}_{\mathcal{A},\mathcal{S}}\left[F\left(w_{\mathcal{S}}\right) - \hat{F}\left(w_{\mathcal{S}}, \mathcal{S}\right)\right] \leq \gamma$.*

Lemma 4 shows that if a symmetric algorithm could be proven to be (γ, K)-uniformly stable, we could bound its generalization error by capturing its stability parameter γ. Then we show the proof of generalization bound for (10) following this idea.

Proof of Theorem 1.

Theorem 1. *(generalization bound) If Assumption 1 and 2 hold. With $\alpha \leq \frac{1}{2L}, \beta_t \leq \frac{1}{\alpha\rho G + (1-\alpha\delta-\alpha\mu)^2 L}$, $\delta < \frac{1}{2\alpha}$ and $\frac{\alpha\rho G}{\mu} < (\frac{1}{2} - \alpha L)^2$. The model $\tilde{\mathcal{A}}(\mathcal{S})$ generated by the last iterate of MAML with regularized updating rule introduced in (6) of the main paper satisfies*

$$\mathbb{E}_{\tilde{\mathcal{A}},\mathcal{S}}[F(\tilde{\mathcal{A}}(\mathcal{S})) - \hat{F}(\tilde{\mathcal{A}}(\mathcal{S}), \mathcal{S})] \leq$$
$$\frac{2G^2(1+\alpha L)(1-\alpha\mu-\alpha\delta+(2+\alpha L-\alpha\delta)\alpha LK)}{mn}\left(\frac{1}{\alpha\rho G + (1-\alpha\delta-\alpha\mu)^2 L} + \frac{1}{-\alpha\rho G + (1-\alpha\delta-\alpha L)^2\mu}\right)$$

where the expectation is taken over the randomness of $\tilde{\mathcal{A}}$ and sampling of $\mathcal{S}$.

Proof. The result in Lemma 4 means that we could bound the generalization error of MAML with inner-level regularization $\tilde{\mathcal{A}}$ by proving its (γ, K)-uniformly stable as defined in Definition 2 and capture the γ parameter.

The Definition 2 of (γ, K)-uniformly stability means that there is a dataset $\bar{\mathcal{S}}$ which is the same as $\mathcal{S}$ except one i such that:

- $\bar{\mathcal{S}}_i^{\text{in}}$ has at most K data points different from $\mathcal{S}_i^{\text{in}}$. We denote the K samples in each dataset by $\{z_j\}_{j=1}^K$ and $\{\bar{z}_j\}_{j=1}^K$.
- $\bar{\mathcal{S}}_i^{\text{out}}$ has at most 1 data points different from $\mathcal{S}_i^{\text{out}}$. They are denoted by ζ and $\bar{\zeta}$.

We consider the two parallel processes of training $\{w^t\}$ and $\{\bar{w}^t\}$ using datasets $\mathcal{S}$ and $\bar{\mathcal{S}}$. The bar superscript is used to denote the process using $\bar{\mathcal{S}}$. $D_i^{t,\text{ out}}$ and $D_i^{t,\text{in}}$ are referring to indices of samples in $\mathcal{D}_i^{t,\text{ out}}$ and $\mathcal{D}_i^{t,\text{in}}$, respectively. We could assume the parallel using a same random machine to sample batches for generating $\{w^t\}$ and $\{\bar{w}^t\}$, i.e., $\mathcal{B}_t = \bar{\mathcal{B}}_t$, $D_i^{t,\text{ out}} = \bar{D}_i^{t,\text{ out}}$, and $D_i^{t,\text{in}} = \bar{D}_i^{t,\text{in}}$

We use v_t to denote the number of indices corresponding to $\{z_j\}_{j=1}^K$ (or $\{\bar{z}_j\}_{j=1}^K$) is chosen in $D_i^{t,\text{in}}$. And u_t is denoting the number of times that the index of sample ζ (or $\bar{\zeta}$) is chosen in $D_i^{t,\text{ out}}$, respectively. As shown in [3], recalling the definition of b and r from Algorithm 2, for each t, the expectations of v_t and u_t are given by

$$\mathbb{E}[v_t] = \frac{K^2 r}{nm}, \quad \mathbb{E}[u_t] = \frac{br}{nm} \tag{11}$$

Then we are coming to the main proof. We first claim that

$$\mathbb{E}_{\tilde{\mathcal{A}}}\left[\left\|w^T-\bar{w}^T\right\|\right] \leq \frac{2G(1-\alpha\mu-\alpha\delta+(2+\alpha L-\alpha\delta)\alpha LK)}{mn}\left(\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}\right) \tag{12}$$

when conditions in Assumption 1 are satisfied and $\{w^T,\bar{w}^T\}$ is generated using $\tilde{\mathcal{A}}$. The next is to prove this claim. To simplify the notation, let us define $\psi(w;\mathcal{D},z):=\ell(w-\alpha\nabla_w(\hat{\mathcal{L}}(w,\mathcal{D})+\frac{\delta}{2}\|w\|^2),z)$. Note that

$$\begin{aligned}\psi(w;\mathcal{D},z) &:= \ell(w-\alpha\nabla_w(\hat{\mathcal{L}}(w,\mathcal{D})+\frac{\delta}{2}\|w\|^2),z)\\ &= \ell((1-\alpha\delta)w-\alpha\nabla\hat{\mathcal{L}}(w,\mathcal{D}),z)\end{aligned}$$

and

$$\nabla\psi(w;\mathcal{D},z)=\left((1-\alpha\delta)I-\alpha\nabla^2\hat{\mathcal{L}}\left(w,\mathcal{D}\right)\right)\nabla\ell\left((1-\alpha\delta)w-\alpha\nabla\hat{\mathcal{L}}\left(w,\mathcal{D}\right),z\right).$$

Recalling from Lemma 1, we know for a λ-strongly convex and η-smooth function ϕ, we have

$$\|(u-\beta\nabla\phi(u))-(v-\beta\nabla\phi(v))\|\leq\left(1-\frac{\beta\lambda\eta}{\lambda+\eta}\right)\|u-v\|$$

for any u and v.

And Lemma 3 shows that for any batch $\mathcal{D}$ and any $z\in\mathcal{Z}$, $\psi(w;\mathcal{D},z)$ is $\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2L$ smooth and $-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu$ strongly convex. Hence, using Lemma 1, for any $j\in\mathcal{B}_t$ that $j\neq i$, we have

$$\left\|w_j^{t+1}-\bar{w}_j^{t+1}\right\|\leq\left(1-\beta_t\frac{1}{\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}}\right)\left\|w^t-\bar{w}^t\right\|. \tag{13}$$

Next, For the case that $i\in\mathcal{B}_t$, we could have

$$\begin{aligned}\left\|w_i^{t+1}-\bar{w}_i^{t+1}\right\| \leq & \frac{1}{b}\sum_{z\in\mathcal{D}_i^{t,\text{ out}}}\left\|\left(w^t-\beta_t\nabla\psi\left(w^t;\mathcal{D}_i^{t,\text{ in}},z\right)\right)-\left(\bar{w}^t-\beta_t\nabla\psi\left(\bar{w}^t;\bar{\mathcal{D}}_i^{t,\text{ in}},z\right)\right)\right\|\\ &+\frac{1}{b}\beta_t\sum_{z\in\bar{\mathcal{D}}_i^{t,\text{ out}}/\mathcal{D}_i^{t,\text{ out}}}\left\|\nabla\psi\left(\bar{w}^t;\bar{\mathcal{D}}_i^{t,\text{ in}},z\right)-\nabla\psi\left(w^t;\mathcal{D}_i^{t,\text{ in}},z\right)\right\|.\end{aligned} \tag{14}$$

To bound the second term on RHS of (14), we first consider that

$$\begin{aligned}\|\nabla\psi(w;\mathcal{D},z)\| &= \|\left((1-\alpha\delta)I-\alpha\nabla^2\hat{\mathcal{L}}\left(w,\mathcal{D}\right)\right)\nabla\ell\left((1-\alpha\delta)w-\alpha\nabla\hat{\mathcal{L}}\left(w,\mathcal{D}\right),z\right)\|\\ &\leq \|\left((1-\alpha\delta)I-\alpha\nabla^2\hat{\mathcal{L}}\left(w,\mathcal{D}\right)\right)\|\|\nabla\ell\left((1-\alpha\delta)w-\alpha\nabla\hat{\mathcal{L}}\left(w,\mathcal{D}\right),z\right)\|\\ &\leq (1-\alpha\mu-\alpha\delta)G.\end{aligned} \tag{15}$$

The last inequality is due to bounded gradient assumption for $\ell(\cdot,z)$ and $(1-\alpha\delta)\boldsymbol{I}-\alpha\nabla^2\hat{\mathcal{L}}\left(w,\mathcal{D}\right)$ being symmetric, semi-positive definite, and $\lambda_{\max}\left((1-\alpha\delta)\boldsymbol{I}-\alpha\nabla^2\hat{\mathcal{L}}\left(w,\mathcal{D}\right)\right)\leq(1-\alpha\delta)-\alpha\mu$ (see Lemma 2). Then, we have

$$\begin{aligned}\left\|\nabla\psi\left(\bar{w}^t;\bar{\mathcal{D}}_i^{t,\text{ in}},z\right)-\nabla\psi\left(w^t;\mathcal{D}_i^{t,\text{ in}},z\right)\right\| &\leq \left\|\nabla\psi\left(\bar{w}^t;\bar{\mathcal{D}}_i^{t,\text{ in}},z\right)\|+\|\nabla\psi\left(w^t;\mathcal{D}_i^{t,\text{ in}},z\right)\right\|\\ &\leq 2(1-\alpha\mu-\alpha\delta)G.\end{aligned} \tag{16}$$

Since $\left|\bar{\mathcal{D}}_i^{t,\text{out}}/\mathcal{D}_i^{t,\text{ out}}\right|=u_t$, using the above result, the second term of (14)'s RHS could be bounded by $2\beta_tu_tG(1-\alpha\mu-\alpha\delta)/b$, i.e.,

$$\begin{aligned}\left\|w_i^{t+1}-\bar{w}_i^{t+1}\right\| \leq & 2\beta_tG(1-\alpha\mu-\alpha\delta)\frac{u_t}{b}\\ &+\frac{1}{b}\sum_{z\in\mathcal{D}_i^{t,\text{ out}}}\left\|\left(w^t-\beta_t\nabla\psi\left(w^t;\mathcal{D}_i^{t,\text{ in}},z\right)\right)-\left(\bar{w}^t-\beta_t\nabla\psi\left(\bar{w}^t;\bar{\mathcal{D}}_i^{t,\text{ in}},z\right)\right)\right\|.\end{aligned} \tag{17}$$

For the second term, note that

$$\begin{aligned}
&\left\|\left(w^t-\beta_t\nabla\psi\left(w^t;\mathcal{D}_i^{t,\text{ in}},z\right)\right)-\left(\bar{w}^t-\beta_t\nabla\psi\left(\bar{w}^t;\bar{\mathcal{D}}_i^{t,\text{ in}},z\right)\right)\right\| \\
&\leq\left\|\left(w^t-\beta_t\nabla\psi\left(w^t;\mathcal{D}_i^{t,\text{ in}},z\right)\right)-\left(\bar{w}^t-\beta_t\nabla\psi\left(\bar{w}^t;\mathcal{D}_i^{t,\text{ in}},z\right)\right)\right\| \\
&+\beta_t\left\|\nabla\psi\left(\bar{w}^t;\mathcal{D}_i^{t,\text{ in}},z\right)-\nabla\psi\left(\bar{w}^t;\bar{\mathcal{D}}_i^{t,\text{ in}},z\right)\right\|.
\end{aligned}\tag{18}$$

For the first term on the RHS of (18), we could bound it similarly to the derivation of (13) by

$$\begin{aligned}
&\left\|\left(w^t-\beta_t\nabla\psi\left(w^t;\mathcal{D}_i^{t,\text{ in}},z\right)\right)-\left(\bar{w}^t-\beta_t\nabla\psi\left(\bar{w}^t;\mathcal{D}_i^{t,\text{ in}},z\right)\right)\right\| \\
&\leq\left(1-\beta_t\frac{1}{\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}}\right)\left\|w^t-\bar{w}^t\right\|.
\end{aligned}\tag{19}$$

And for the second term on the RHS of (18), we have

$$\begin{aligned}
&\left\|\nabla\psi\left(\bar{w}^t;\mathcal{D}_i^{t,\text{ in}},z\right)-\nabla\psi\left(\bar{w}^t;\bar{\mathcal{D}}_i^{t,\text{ in}},z\right)\right\| \\
&=\|\left((1-\alpha\delta)I-\alpha\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right)\right)\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right),z\right) \\
&\quad-\left((1-\alpha\delta)I-\alpha\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right)\right)\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right),z\right)\| \\
&\leq(1-\alpha\delta)\left\|\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right),z\right)-\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right),z\right)\right\|+ \\
&\alpha\|\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right)\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right),z\right) \\
&-\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right)\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right),z\right)\| \\
&\leq(1-\alpha\delta)\left\|\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right),z\right)-\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right),z\right)\right\|+ \\
&\alpha\|\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right)\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right),z\right)-\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right)\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right),z\right)+ \\
&\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right)\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right),z\right)-\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right)\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right),z\right)\| \\
&\leq(1-\alpha\delta)\left\|\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right),z\right)-\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right),z\right)\right\|+ \\
&\alpha\|\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right)\|\|\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right),z\right)-\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right),z\right)\|+ \\
&\alpha\|\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right)-\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right)\|\|\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right),z\right)\| \\
&\leq(1-\alpha\delta+\alpha L)\left\|\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right),z\right)-\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right),z\right)\right\|+ \\
&\alpha G\left\|\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right)-\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right)\right\|.
\end{aligned}\tag{20}$$

The last inequality is given by the smoothness and bounded gradient assumption for $\ell(\cdot,z)$. Next, we are going to bound the terms in (20). Note that

$$\begin{aligned}
&\left\|\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right),z\right)-\nabla\ell\left((1-\alpha\delta)\bar{w}^t-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right),z\right)\right\| \\
&\leq\alpha L\left\|\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right)-\nabla\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right)\right\|\leq 2\alpha LG\frac{v_t}{K}
\end{aligned}$$

and

$$\left\|\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\mathcal{D}_i^{t,\text{ in}}\right)-\nabla^2\hat{\mathcal{L}}\left(\bar{w}^t,\bar{\mathcal{D}}_i^{t,\text{ in}}\right)\right\|\leq 2L\frac{v_t}{K}.$$

Putting the above two results into (20), we have

$$\begin{aligned}\left\|\nabla\psi\left(\bar{w}^t;\mathcal{D}_i^{t,\text{ in}},z\right)-\nabla\psi\left(\bar{w}^t;\bar{\mathcal{D}}_i^{t,\text{ in}},z\right)\right\| &\leq 2(1-\alpha\delta+\alpha L)\alpha LG\frac{v_t}{K}+2\alpha LG\frac{v_t}{K}\\ &=2(2+\alpha L-\alpha\delta)\alpha LG\frac{v_t}{K}.\end{aligned} \tag{21}$$

Putting the result in (19) and (18) into (17), we have

$$\begin{aligned}\left\|w_i^{t+1}-\bar{w}_i^{t+1}\right\| \leq &\left(1-\beta_t\frac{1}{\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}}\right)\left\|w^t-\bar{w}^t\right\|\\ &+2\beta_tG((1-\alpha\mu-\alpha\delta)\frac{u_t}{b}+\alpha L(2+\alpha L-\alpha\delta)\frac{v_t}{K}).\end{aligned}$$

Along with (13), we have

$$\begin{aligned}\left\|\frac{1}{r}\sum_{j\in\mathcal{B}_t}w_j^{t+1}-\frac{1}{r}\sum_{j\in\mathcal{B}_t}\bar{w}_j^{t+1}\right\| \leq &\left(1-\beta_t\frac{1}{\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}}\right)\left\|w^t-\bar{w}^t\right\|\\ &+2\beta_tG((1-\alpha\mu-\alpha\delta)\frac{u_t}{rb}+\alpha L(2+\alpha L-\alpha\delta)\frac{v_t}{rK}),\end{aligned}$$

which indicates

$$\begin{aligned}\left\|w^{t+1}-\bar{w}^{t+1}\right\| \leq &\left(1-\beta_t\frac{1}{\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}}\right)\left\|w^t-\bar{w}^t\right\|\\ &+2\beta_tG((1-\alpha\mu-\alpha\delta)\frac{u_t}{rb}+\alpha L(2+\alpha L-\alpha\delta)\frac{v_t}{rK}).\end{aligned}$$

Using (11), we could take the expectation for both sides and get

$$\begin{aligned}\mathbb{E}_{\tilde{\mathcal{A}}}\left[\left\|w^{t+1}-\bar{w}^{t+1}\right\|\right] \leq &\left(1-\beta_t\frac{1}{\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}}\right)\mathbb{E}_{\tilde{\mathcal{A}}}\left[\left\|w^t-\bar{w}^t\right\|\right]\\ &+2\frac{\beta_tG}{mn}(1-\alpha\mu-\alpha\delta+(2+\alpha L-\alpha\delta)\alpha LK).\end{aligned}$$

The bound could be rewritten as

$$\mathbb{E}_{\tilde{\mathcal{A}}}\left[\left\|w^{t+1}-\bar{w}^{t+1}\right\|\right] \leq (1-\beta_t\lambda)\,\mathbb{E}_{\tilde{\mathcal{A}}}\left[\left\|w^t-\bar{w}^t\right\|\right]+\beta_t\eta,$$

where the λ and η are given by

$$\lambda:=\frac{1}{\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}},\quad \eta:=\frac{2G}{mn}(1-\alpha\mu-\alpha\delta+(2+\alpha L-\alpha\delta)\alpha LK).$$

In fact, the main claim (12) is equivalent to

$$\mathbb{E}_{\tilde{\mathcal{A}}}\left[\left\|w^t-\bar{w}^t\right\|\right] \leq \frac{\eta}{\lambda}.$$

For $t=1$, this is true because of $\beta_0 \leq \frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2L} \leq \frac{1}{\lambda}$. Then the result could be easily obtained by induction. The result could be written as.

$$\mathbb{E}_{\tilde{\mathcal{A}}}\left[\left\|w^T-\bar{w}^T\right\|\right] \leq \frac{2G(1-\alpha\mu-\alpha\delta+(2+\alpha L-\alpha\delta)\alpha LK)}{mn}(\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}).$$

Having proved the above result, we are ready to finish proving Theorem 1. We have

$$\begin{aligned}&\left|\ell\left(w^T-\alpha\nabla\hat{\mathcal{L}}\left(w^T,\{z_j\}_{j=1}^K\right),\bar{z}\right)-\ell\left(\bar{w}^T-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^T,\{z_j\}_{j=1}^K\right),\bar{z}\right)\right|\\ &\leq G\left\|\left(w^T-\alpha\nabla\hat{\mathcal{L}}\left(w^T,\{z_j\}_{j=1}^K\right),\bar{z}\right)-\left(\bar{w}^T-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^T,\{z_j\}_{j=1}^K\right),\bar{z}\right)\right\|\\ &\leq G\|w^T-\bar{w}^T\|+\alpha G\|\nabla\hat{\mathcal{L}}\left(w^T,\{z_j\}_{j=1}^K\right)-\nabla\hat{\mathcal{L}}\left(\bar{w}^T,\{z_j\}_{j=1}^K\right)\|\\ &\leq (1+\alpha L)G\|w^T-\bar{w}^T\|.\end{aligned}$$

Then,

$$\mathbb{E}_{\tilde{\mathcal{A}}}\left[\ell\left(w^T-\alpha\nabla\hat{\mathcal{L}}\left(w^T,\{z_j\}_{j=1}^K\right),\bar{z}\right)-\ell\left(\bar{w}^T-\alpha\nabla\hat{\mathcal{L}}\left(\bar{w}^T,\{z_j\}_{j=1}^K\right),\bar{z}\right)\right]\leq$$
$$\frac{2G^2(1+\alpha L)(1-\alpha\mu-\alpha\delta+(2+\alpha L-\alpha\delta)\alpha LK)}{mn}\left(\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}\right).$$

This means the algorithm is (γ, K)-uniformly stable with RHS as the γ parameter.
Finally, for the meta-testing phase learning objective of the original MAML (unregularized), by Lemma 4, the generalization bound in (10) is given by

$$\mathbb{E}_{\tilde{\mathcal{A}},\mathcal{S}}[F(\tilde{\mathcal{A}}(\mathcal{S}))-\hat{F}(\tilde{\mathcal{A}}(\mathcal{S}),\mathcal{S})]\leq$$
$$\frac{2G^2(1+\alpha L)(1-\alpha\mu-\alpha\delta+(2+\alpha L-\alpha\delta)\alpha LK)}{mn}\left(\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}\right),$$

where $F(\cdot)$ and $\hat{F}(\cdot,\mathcal{S})$ are population loss and empirical loss for unregularized MAML, respectively. The proof is complete. □

A.2.3 Training Bias

In this section, we give the proof of Theorem 2 on training bias bound in the paper.

Theorem 2. *(training bias bound) If Assumption 1 and 2 hold. With $\alpha\leq\frac{1}{2L}$, $\delta<\frac{1}{2\alpha}$ and $\frac{\alpha\rho G}{\mu}<(\frac{1}{2}-\alpha L)^2$. The training bias from MAML with inner-level L2 regularization to the original MAML is bounded by*

$$\mathbb{E}_{\mathcal{S}}\left[\hat{F}(\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}),\mathcal{S})-\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\right]\leq\frac{\alpha^2(\alpha\rho G+(1-\alpha\mu)^2L)((1-\alpha\mu-\alpha\delta)L\|w^*\|+G)^2\delta^2}{2(-\alpha\rho G+(1-\alpha L-\alpha\delta)^2\mu)^2}$$

where $\|w^*\|:=\max_{\mathcal{S}}\|\arg\min_w\hat{F}(w,\mathcal{S})\|$, the expectation is taken over sampling of $\mathcal{S}$.

Proof. The empirical loss of unregularized MAML is defined by

$$\hat{F}(w,\mathcal{S}):=\frac{1}{m}\sum_{i=1}^m\hat{F}_i\left(w,\mathcal{S}_i\right), \tag{22}$$

where $\hat{F}_i\left(\cdot,\mathcal{S}_i\right)$ is given by

$$\begin{aligned}\hat{F}_i\left(w,\mathcal{S}_i\right):&=\frac{1}{\binom{n}{k}}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}}\frac{1}{n}\sum_{z\in\mathcal{S}_i^{\text{out}}}\ell\left(w-\frac{\alpha}{K}\sum_{z'\in\mathcal{D}_i^{\text{in}}}\nabla\ell\left(w,z'\right),z\right)\\&=\frac{1}{\binom{n}{k}}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}}\hat{\mathcal{L}}\left(w-\alpha\nabla\hat{\mathcal{L}}\left(w,\mathcal{D}_i^{\text{in}}\right),\mathcal{S}_i^{\text{out}}\right).\end{aligned} \tag{23}$$

So (22) could also be written as

$$\hat{F}(w,\mathcal{S}):=\frac{1}{m}\frac{1}{\binom{n}{k}}\sum_{i=1}^m\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}}\hat{\mathcal{L}}\left(w-\alpha\nabla\hat{\mathcal{L}}\left(w,\mathcal{D}_i^{\text{in}}\right),\mathcal{S}_i^{\text{out}}\right). \tag{24}$$

For MAML with inner-level L2-Norm regularization, the corresponding empirical loss is given by

$$\hat{\tilde{F}}(w,\mathcal{S}):=\frac{1}{m}\frac{1}{\binom{n}{k}}\sum_{i=1}^m\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}}\hat{\mathcal{L}}\left(w-\alpha\nabla_w(\hat{\mathcal{L}}\left(w,\mathcal{D}_i^{\text{in}}\right)+\frac{\delta}{2}\|w\|^2),\mathcal{S}_i^{\text{out}}\right), \tag{25}$$

where δ is the parameter for regularization. Our goal is to bound the training bias

$$\hat{F}(\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}),\mathcal{S})-\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})=\hat{F}(\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}),\mathcal{S})-\hat{F}(\arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S}),\mathcal{S}). \tag{26}$$

To bound (26), we could first bound

$$\|\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S})-\arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S}).\| \tag{27}$$

We denote the two model parameters by

$$u:=\arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})$$

and

$$v:=\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}).$$

Lemma 3 shows the strongly convexity of both $\hat{\mathcal{L}}\left(w-\alpha\nabla\hat{\mathcal{L}}\left(w,\mathcal{D}_i^{\text{in}}\right),\mathcal{S}_i^{\text{out}}\right)$ and $\hat{\mathcal{L}}\left(w-\alpha\nabla_w(\hat{\mathcal{L}}\left(w,\mathcal{D}_i^{\text{in}}\right)+\frac{\delta}{2}\|w\|^2),\mathcal{S}_i^{\text{out}}\right)$, which also indicates the strongly convexity of $\hat{F}(\cdot,\mathcal{S})$ and $\hat{\tilde{F}}(\cdot,\mathcal{S})$. Suppose the optimal solution lies within $\mathcal{W}$ for $\hat{F}(\cdot,\mathcal{S})$ and $\hat{\tilde{F}}(\cdot,\mathcal{S})$, we have

$$\nabla\hat{F}(u,\mathcal{S})=\frac{1}{m}\frac{1}{\binom{n}{k}}\sum_{i=1}^{m}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}}\nabla_w\hat{\mathcal{L}}\left(w-\alpha\nabla\hat{\mathcal{L}}\left(w,\mathcal{D}_i^{\text{in}}\right),\mathcal{S}_i^{\text{out}}\right)\Bigg|_{w=u}=0 \tag{28}$$

and

$$\nabla\hat{\tilde{F}}(v,\mathcal{S})=\frac{1}{m}\frac{1}{\binom{n}{k}}\sum_{i=1}^{m}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}}\nabla_w\hat{\mathcal{L}}\left(w-\alpha\nabla_w(\hat{\mathcal{L}}\left(w,\mathcal{D}_i^{\text{in}}\right)+\frac{\delta}{2}\|w\|^2),\mathcal{S}_i^{\text{out}}\right)\Bigg|_{w=v}=0. \tag{29}$$

Note that

$$\begin{aligned}&\nabla_w\hat{\mathcal{L}}\left(w-\alpha\nabla\hat{\mathcal{L}}\left(w,\mathcal{D}_i^{\text{in}}\right),\mathcal{S}_i^{\text{out}}\right)=\\&\left(I_d-\alpha\nabla^2\hat{\mathcal{L}}\left(w,\mathcal{D}_i^{in}\right)\right)\nabla\hat{\mathcal{L}}\left(w-\alpha\nabla\hat{\mathcal{L}}\left(w,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)\end{aligned} \tag{30}$$

and

$$\begin{aligned}&\nabla_w\hat{\mathcal{L}}\left(w-\alpha\nabla_w(\hat{\mathcal{L}}\left(w,\mathcal{D}_i^{\text{in}}\right)+\frac{\delta}{2}\|w\|^2),\mathcal{S}_i^{\text{out}}\right)=\\&\left((1-\alpha\delta)I_d-\alpha\nabla^2\hat{\mathcal{L}}\left(w,\mathcal{D}_i^{in}\right)\right)\nabla\hat{\mathcal{L}}\left((1-\alpha\delta)w-\alpha\nabla\hat{\mathcal{L}}\left(w,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right).\end{aligned} \tag{31}$$

By plugging (30) and (31) into (28) and (29) respectively, we have

$$\nabla\hat{F}(u,\mathcal{S})=\frac{1}{m}\frac{1}{\binom{n}{k}}\sum_{i=1}^{m}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}}\left(I_d-\alpha\nabla^2\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{in}\right)\right)\nabla\hat{\mathcal{L}}\left(u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)=0 \tag{32}$$

and

$$\nabla\hat{\tilde{F}}(v,\mathcal{S})=\frac{1}{m}\frac{1}{\binom{n}{k}}\sum_{i=1}^{m}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}}\left((1-\alpha\delta)I_d-\alpha\nabla^2\hat{\mathcal{L}}\left(v,\mathcal{D}_i^{in}\right)\right)\nabla\hat{\mathcal{L}}\left((1-\alpha\delta)v-\alpha\nabla\hat{\mathcal{L}}\left(v,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)=0. \tag{33}$$

Then, we could bound $\|\nabla\hat{\tilde{F}}(u,\mathcal{S})-\nabla\hat{\tilde{F}}(v,\mathcal{S})\|$ by

$$
\begin{aligned}
&\|\nabla\hat{\tilde{F}}(u,\mathcal{S})-\nabla\hat{\tilde{F}}(v,\mathcal{S})\| \\
&=\|\nabla\hat{\tilde{F}}(u,\mathcal{S})-0\|= \\
&=\|\nabla\hat{\tilde{F}}(u,\mathcal{S})-\nabla\hat{F}(u,\mathcal{S})\| \\
&=\|\frac{1}{m}\frac{1}{\binom{n}{k}}\sum_{i=1}^{m}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}}\left((1-\alpha\delta)I_d-\alpha\nabla^2\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{in}\right)\right)\nabla\hat{\mathcal{L}}\left((1-\alpha\delta)u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right) \\
&\quad-\frac{1}{m}\frac{1}{\binom{n}{k}}\sum_{i=1}^{m}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}}\left(I_d-\alpha\nabla^2\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{in}\right)\right)\nabla\hat{\mathcal{L}}\left(u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)\| \\
&\overset{(a)}{=}\|\frac{1}{m}\frac{1}{\binom{n}{k}}\sum_{i=1}^{m}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}}\Big[\left((1-\alpha\delta)I_d-\alpha\nabla^2\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{in}\right)\right)\nabla\hat{\mathcal{L}}\left((1-\alpha\delta)u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right) \\
&\quad-\left(I_d-\alpha\nabla^2\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{in}\right)\right)\nabla\hat{\mathcal{L}}\left(u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)\Big]\| \\
&=\|\frac{1}{m}\frac{1}{\binom{n}{k}}\sum_{i=1}^{m}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}} \\
&\quad\Big[\left((1-\alpha\delta)I_d-\alpha\nabla^2\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{in}\right)\right)\left(\nabla\hat{\mathcal{L}}\left((1-\alpha\delta)u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)-\nabla\hat{\mathcal{L}}\left(u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)\right) \\
&\quad-\alpha\delta\nabla\hat{\mathcal{L}}\left(u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)\Big]\| \\
&\overset{(b)}{\leq}\frac{1}{m}\frac{1}{\binom{n}{k}}\sum_{i=1}^{m}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}} \\
&\quad\|\left((1-\alpha\delta)I_d-\alpha\nabla^2\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{in}\right)\right)\left(\nabla\hat{\mathcal{L}}\left((1-\alpha\delta)u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)-\nabla\hat{\mathcal{L}}\left(u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)\right) \\
&\quad-\alpha\delta\nabla\hat{\mathcal{L}}\left(u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)\| \\
&\overset{(c)}{\leq}\frac{1}{m}\frac{1}{\binom{n}{k}}\sum_{i=1}^{m}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}} \\
&\quad\Big(\|\left((1-\alpha\delta)I_d-\alpha\nabla^2\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{in}\right)\right)\left(\nabla\hat{\mathcal{L}}\left((1-\alpha\delta)u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)-\nabla\hat{\mathcal{L}}\left(u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)\right)\| \\
&\quad+\alpha|\delta|\|\nabla\hat{\mathcal{L}}\left(u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right),\mathcal{D}_i^{\text{out}}\right)\|\Big) \\
&\overset{(d)}{\leq}\frac{1}{m}\frac{1}{\binom{n}{k}}\sum_{i=1}^{m}\sum_{\substack{\mathcal{D}_i^{\text{in}}\subset\mathcal{S}_i^{\text{in}}\\|\mathcal{D}_i^{\text{in}}|=K}} \\
&\quad\big((1-\alpha\mu-\alpha\delta)L\|((1-\alpha\delta)u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right))-(u-\alpha\nabla\hat{\mathcal{L}}\left(u,\mathcal{D}_i^{\text{in}}\right))\| \\
&\quad+\alpha|\delta|G\big) \\
&=\frac{1}{m}\frac{1}{\binom{n}{k}}m\binom{n}{k}\big((1-\alpha\mu-\alpha\delta)L\|\alpha\delta u\|+\alpha|\delta|G\big) \\
&=\big(\alpha|\delta|(1-\alpha\mu-\alpha\delta)L\|u\|+\alpha|\delta|G\big) \\
&=\alpha|\delta|\big((1-\alpha\mu-\alpha\delta)L\|u\|+G\big).
\end{aligned}
\tag{34}
$$

Here, (a) is by combining each term with the same index within the two summations, (b) and (c) are due to triangle inequality, and (d) is due to the strongly convex, smooth and bounded gradient property of $\hat{\mathcal{L}}\left(\cdot, \mathcal{D}_i^{\text{in}}\right)$.
By the definition of u and v, (34) actually shows

$$\|\nabla\hat{\tilde{F}}(\arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S}),\mathcal{S}) - \nabla\hat{\tilde{F}}(\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}),\mathcal{S})\| \leq \alpha|\delta|\big((1-\alpha\mu-\alpha\delta)L\|\arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\| + G\big). \tag{35}$$

Lemma 3 also indicates that $\hat{\tilde{F}}(\cdot,\mathcal{S})$ is $(-\alpha\rho G + (1-\alpha L-\alpha\delta)^2\mu)$ strongly-convex, so we could bound $\|\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}) - \arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\|$ by

$$\begin{aligned}\|\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}) - \arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\| &\leq \frac{1}{(-\alpha\rho G + (1-\alpha L-\alpha\delta)^2\mu)}\|\nabla\hat{\tilde{F}}(\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}),\mathcal{S}) - \nabla\hat{\tilde{F}}(\arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S}),\mathcal{S})\| \\ &\leq \frac{1}{(-\alpha\rho G + (1-\alpha L-\alpha\delta)^2\mu)}\alpha|\delta|\big((1-\alpha\mu-\alpha\delta)L\|\arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\| + G\big).\end{aligned} \tag{36}$$

Take square for both sides of (36), we have

$$\|\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}) - \arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\|^2 \leq \left(\frac{1}{(-\alpha\rho G + (1-\alpha L-\alpha\delta)^2\mu)}\alpha|\delta|\big((1-\alpha\mu-\alpha\delta)L\|\arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\| + G\big)\right)^2, \tag{37}$$

and then take the expectation over the sampling of $\mathcal{S}$, we further have

$$\mathbb{E}_{\mathcal{S}}\left[\|\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}) - \arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\|^2\right] \leq \left(\frac{1}{(-\alpha\rho G + (1-\alpha L-\alpha\delta)^2\mu)}\alpha|\delta|\big((1-\alpha\mu-\alpha\delta)L\|w^*\| + G\big)\right)^2, \tag{38}$$

where $\|w^*\| := \max_{\mathcal{S}}\|\arg\min_w \hat{F}(w,\mathcal{S})\|$, the maximum is taken over sampling of $\mathcal{S}$.
Recall from Lemma 3 that $\hat{F}(\cdot,\mathcal{S})$ is $(\alpha\rho G + (1-\alpha\mu)^2 L)$ smooth, and note that $\nabla\hat{F}(\arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S}),\mathcal{S}) = 0$, we could bound (26) by

$$\begin{aligned}\hat{F}(\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}),\mathcal{S}) - \min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S}) &= \hat{F}(\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}),\mathcal{S}) - \hat{F}(\arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S}),\mathcal{S}) \\ &\leq \frac{1}{2}(\alpha\rho G + (1-\alpha\mu)^2 L)\|\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}) - \arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\|^2.\end{aligned} \tag{39}$$

Finally, by taking the expectation, we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}}\left[\hat{F}(\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}),\mathcal{S}) - \min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\right] \\ &\leq \mathbb{E}_{\mathcal{S}}\left[\frac{1}{2}(\alpha\rho G + (1-\alpha\mu)^2 L)\|\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}) - \arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\|^2\right] \\ &= \frac{1}{2}(\alpha\rho G + (1-\alpha\mu)^2 L)\mathbb{E}_{\mathcal{S}}\left[\|\arg\min_{\mathcal{W}}\hat{\tilde{F}}(\cdot,\mathcal{S}) - \arg\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\|^2\right] \\ &\overset{(a)}{\leq} \frac{1}{2}(\alpha\rho G + (1-\alpha\mu)^2 L)\left(\frac{1}{(-\alpha\rho G + (1-\alpha L-\alpha\delta)^2\mu)}\alpha|\delta|\big((1-\alpha\mu-\alpha\delta)L\|w^*\| + G\big)\right)^2 \\ &= \frac{\alpha^2(\alpha\rho G + (1-\alpha\mu)^2 L)((1-\alpha\mu-\alpha\delta)L\|w^*\| + G)^2\delta^2}{2(-\alpha\rho G + (1-\alpha L-\alpha\delta)^2\mu)^2}.\end{aligned} \tag{40}$$

(a) is because of (38). The proof is complete.

□

A.3. Further Analysis

A.3.1 Property of Generalization Error Bound

In 4.1, we claimed that if we regard the generalization bound as a function $GB(\delta)$, its derivative $GB'(\delta)$ would be positive for $\delta \in (-\infty, \frac{1}{2\alpha})$, i.e.,

$$GB'(\delta) > 0 \quad \forall \delta \in (-\infty, \frac{1}{2\alpha}) \tag{41}$$

In this section, we provide proof of this claim.

Proof. Based on the result of Theorem 1, the function $GB(\delta)$ is given by

$$GB(\delta) = \frac{2G^2(1+\alpha L)(1-\alpha\mu-\alpha\delta+(2+\alpha L-\alpha\delta)\alpha LK)}{mn}(\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2 L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}).$$

Taking its derivative, we have

$$\begin{aligned}
&GB'(\delta) = \\
&\frac{2\alpha G^2(1+\alpha L)}{mn(-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu)}\cdot \\
&(-(\alpha LK+1)((1-\alpha\delta-\alpha L)^2\mu-\alpha\rho G)+2\mu(1-\alpha\delta-\alpha L)(\alpha LK(2+\alpha L-\alpha\delta)+(1-\alpha\delta-\alpha\mu)) \\
&+ \\
&\frac{2\alpha G^2(1+\alpha L)}{mn(\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2 L)}\cdot \\
&(-(\alpha LK+1)((1-\alpha\delta-\alpha\mu)^2 L+\alpha\rho G)+2L(1-\alpha\delta-\alpha\mu)(\alpha Lk(2+\alpha L-\alpha\delta)+(1-\alpha\delta-\alpha\mu))).
\end{aligned} \tag{42}$$

To prove (41), we are going to prove both terms on RHS of (42) are greater than 0 for $\delta \in (-\infty, \frac{1}{2\alpha})$.
For the first term on RHS of (42), having $\frac{2\alpha G^2(1+\alpha L)}{mn(-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu)} > 0$, we only need to prove

$$-(\alpha LK+1)((1-\alpha\delta-\alpha L)^2\mu-\alpha\rho G)+2\mu(1-\alpha\delta-\alpha L)(\alpha LK(2+\alpha L-\alpha\delta)+(1-\alpha\delta-\alpha\mu)>0. \tag{43}$$

(43) could be re-written by

$$-(\alpha LK+1)((1-\alpha\delta-\alpha L)^2\mu-\alpha\rho G)+2\mu(1-\alpha\delta-\alpha L)((\alpha LK+1)(2+\alpha L-\alpha\delta)-(1+\alpha L+\alpha\mu))>0. \tag{44}$$

Then, since $\mu > 0$, $(\alpha LK+1) > 0$ and $(1-\alpha\delta-\alpha L) > 0$, when $\delta < \frac{1}{2\alpha}$, by dividing both sides of (44) by $\mu(1-\alpha\delta-\alpha L)(\alpha LK+1)$, we find this inequality is equivalent to

$$-(1-\alpha\delta-\alpha L-\frac{\alpha\rho G}{\mu(1-\alpha\delta-\alpha L)})+2(2+\alpha L-\alpha\delta-\frac{1+\alpha L+\alpha\mu}{\alpha LK+1})>0. \tag{45}$$

(45) is equivalent to

$$3+3\alpha L-\alpha\delta+\frac{\alpha\rho G}{\mu(1-\alpha\delta-\alpha L)}-\frac{2(1+\alpha L+\alpha\mu)}{\alpha LK+1}>0. \tag{46}$$

Since $\alpha LK+1>1$, (46) is true if

$$3+3\alpha L-\alpha\delta+\frac{\alpha\rho G}{\mu(1-\alpha\delta-\alpha L)}-2(1+\alpha L+\alpha\mu)>0. \tag{47}$$

By recombining the LHS of (47), we obtain its equivalent form

$$(1-\alpha\mu-\alpha\delta)+\alpha(L-\mu)+\frac{\alpha\rho G}{\mu(1-\alpha\delta-\alpha L)}>0. \tag{48}$$

When $\delta < \frac{1}{2\alpha}$, (48) is true since $(1-\alpha\mu-\alpha\delta) > 0$, $\frac{\alpha\rho G}{\mu(1-\alpha\delta-\alpha L)} > 0$, and $\alpha(L-\mu)$ is non-negative. This proves (43) to be true and shows the first term of (42)'s RHS is greater than 0 for $\delta \in (-\infty, \frac{1}{2\alpha})$.

Then, we move on to prove the second term of (42)'s RHS is greater than 0 for $\delta \in (-\infty, \frac{1}{2\alpha})$. Having $\frac{2\alpha G^2(1+\alpha L)}{mn(-\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2 L)} > 0$, we only need to prove

$$-(\alpha LK+1)((1-\alpha\delta-\alpha\mu)^2 L+\alpha\rho G)+2L(1-\alpha\delta-\alpha\mu)(\alpha LK(2+\alpha L-\alpha\delta)+(1-\alpha\delta-\alpha\mu)>0. \tag{49}$$

Similar to the proof of the first term, when $\delta < \frac{1}{2\alpha}$ we could obtain an equivalent inequality for (49) by dividing its both sides by $L(1-\alpha\delta-\alpha\mu)(\alpha LK+1)$ and reordering:

$$3+2\alpha L+\alpha\mu-\alpha\delta-\frac{\alpha\rho G}{L(1-\alpha\delta-\alpha\mu)}-2\frac{1+\alpha L+\alpha\mu}{\alpha LK+1}>0. \tag{50}$$

Since $\frac{\alpha\rho G}{\mu} < (\frac{1}{2}-\alpha L)^2$ and $\alpha Lk+1>1$, (50) is true if

$$3+2\alpha L+\alpha\mu-\alpha\delta-\frac{\mu(\frac{1}{2}-\alpha L)^2}{L(1-\alpha\delta-\alpha\mu)}-2(1+\alpha L+\alpha\mu)\geq 0. \tag{51}$$

Note that we have

$$\frac{\mu(\frac{1}{2}-\alpha L)^2}{L(1-\alpha\delta-\alpha\mu)} \overset{(a)}{<} \frac{\mu(1-\alpha\delta-\alpha L)^2}{L(1-\alpha\delta-\alpha\mu)} \overset{(b)}{\leq} \frac{L(1-\alpha\delta-\alpha L)^2}{L(1-\alpha\delta-\alpha L)} = 1-\alpha\delta-\alpha L, \tag{52}$$

where (a) is true when $\delta < \frac{1}{2\alpha}$ and (b) is because of $\mu \leq L$. By plugging (52) into (51), we have (51) being true when

$$3+2\alpha L+\alpha\mu-\alpha\delta-(1-\alpha\delta-\alpha L)-2(1+\alpha L+\alpha\mu)\geq 0 \tag{53}$$

(53) is equivalent to

$$\alpha L-\alpha\mu\geq 0. \tag{54}$$

This is obviously true since we have $\alpha>0$ and $L\geq\mu$. This proves (49) to be true and shows the second term of (42)'s RHS is greater than 0 for $\delta \in (-\infty, \frac{1}{2\alpha})$.
Since both terms of (42)'s RHS being greater than 0 for $\delta < \frac{1}{2\alpha}$ has been proved, the conclusion $GB'(\delta)>0 \forall \delta \in (-\infty, \frac{1}{2\alpha})$ in (41) is obtained. The proof is complete. □

A.3.2 Property of Training Bias Bound

In 4.2, we claimed that if we regard the training bais bound as a function $TB(\delta)$, for a legal positive choice of δ-value δ_0, we would always have $TB(\delta_0) > TB(-\delta_0)$, i.e.,

$$TB(\delta_0) > TB(-\delta_0) \quad \forall \delta_0 \in (0, \frac{1}{2\alpha}) \tag{55}$$

In this section, we provide proof of this claim.

Proof. From the result of Theorem 2, the training bias bound function of δ is given by

$$TB(\delta)=\frac{\alpha^2(\alpha\rho G+(1-\alpha\mu)^2 L)((1-\alpha\mu-\alpha\delta)L\|w^*\|+G)^2\delta^2}{2(-\alpha\rho G+(1-\alpha L-\alpha\delta)^2\mu)^2}. \tag{56}$$

It's easy to find that $TB(\delta)>0$ for any $\delta\neq 0$. So the conclusion (55) is equivalent to

$$\frac{TB(\delta_0)}{TB(-\delta_0)}>1 \quad \forall \delta_0 \in (0, \frac{1}{2\alpha}). \tag{57}$$

By plugging (56) into (57), (57) is equivalent to

$$\frac{((1-\alpha\mu-\alpha\delta_0)L\|w^*\|+G)^2}{((1-\alpha\mu+\alpha\delta_0)L\|w^*\|+G)^2}\frac{(-\alpha\rho G+(1-\alpha L+\alpha\delta_0)^2\mu)^2}{(-\alpha\rho G+(1-\alpha L-\alpha\delta_0)^2\mu)^2}>1 \quad \forall \delta_0 \in (0, \frac{1}{2\alpha}). \tag{58}$$

By taking the square root for both sides, (58) is equivalent to

$$\frac{((1-\alpha\mu-\alpha\delta_0)L\|w^*\|+G)}{((1-\alpha\mu+\alpha\delta_0)L\|w^*\|+G)}\frac{(-\alpha\rho G+(1-\alpha L+\alpha\delta_0)^2\mu)}{(-\alpha\rho G+(1-\alpha L-\alpha\delta_0)^2\mu)}>1 \quad \forall\delta_0\in(0,\frac{1}{2\alpha}). \tag{59}$$

A sequence of equivalent transformations on (59) can be performed as follows:

(59) $\iff$

$$\frac{((1-\alpha\mu-\alpha\delta_0)+\frac{G}{L\|w^*\|})}{((1-\alpha\mu+\alpha\delta_0)+\frac{G}{L\|w^*\|})}\frac{(-\frac{\alpha\rho G}{\mu}+(1-\alpha L+\alpha\delta_0)^2)}{(-\frac{\alpha\rho G}{\mu}+(1-\alpha L-\alpha\delta_0)^2)}>1 \quad \forall\delta_0\in(0,\frac{1}{2\alpha}).$$

$\iff$

$$\frac{((1-\alpha\mu+\frac{G}{L\|w^*\|})-\alpha\delta_0)}{((1-\alpha\mu+\frac{G}{L\|w^*\|})+\alpha\delta_0)}\frac{((-\frac{\alpha\rho G}{\mu}+(1-\alpha L)^2+\alpha^2{\delta_0}^2)+2\alpha\delta_0(1-\alpha L))}{((-\frac{\alpha\rho G}{\mu}+(1-\alpha L)^2+\alpha^2{\delta_0}^2)-2\alpha\delta_0(1-\alpha L))}>1 \quad \forall\delta_0\in(0,\frac{1}{2\alpha}).$$

$\iff$

$$\begin{aligned}&((1-\alpha\mu+\frac{G}{L\|w^*\|})-\alpha\delta_0)((-\frac{\alpha\rho G}{\mu}+(1-\alpha L)^2+\alpha^2{\delta_0}^2)+2\alpha\delta_0(1-\alpha L))>\\&((1-\alpha\mu+\frac{G}{L\|w^*\|})+\alpha\delta_0)((-\frac{\alpha\rho G}{\mu}+(1-\alpha L)^2+\alpha^2{\delta_0}^2)-2\alpha\delta_0(1-\alpha L)) \quad \forall\delta_0\in(0,\frac{1}{2\alpha}).\end{aligned}$$

$\iff$

$$\begin{aligned}&((1-\alpha\mu+\frac{G}{L\|w^*\|})(-\frac{\alpha\rho G}{\mu}+(1-\alpha L)^2+\alpha^2{\delta_0}^2)-2\alpha^2{\delta_0}^2(1-\alpha L))\\&+(2\alpha\delta_0(1-\alpha L)(1-\alpha\mu+\frac{G}{L\|w^*\|})-\alpha\delta_0(-\frac{\alpha\rho G}{\mu}+(1-\alpha L)^2+\alpha^2{\delta_0}^2))>\\&((1-\alpha\mu+\frac{G}{L\|w^*\|})(-\frac{\alpha\rho G}{\mu}+(1-\alpha L)^2+\alpha^2{\delta_0}^2)-2\alpha^2{\delta_0}^2(1-\alpha L))\\&-(2\alpha\delta_0(1-\alpha L)(1-\alpha\mu+\frac{G}{L\|w^*\|})-\alpha\delta_0(-\frac{\alpha\rho G}{\mu}+(1-\alpha L)^2+\alpha^2{\delta_0}^2)) \quad \forall\delta_0\in(0,\frac{1}{2\alpha}).\end{aligned}$$

$\iff$

$$(2\alpha\delta_0(1-\alpha L)(1-\alpha\mu+\frac{G}{L\|w^*\|})-\alpha\delta_0(-\frac{\alpha\rho G}{\mu}+(1-\alpha L)^2+\alpha^2{\delta_0}^2))>0 \quad \forall\delta_0\in(0,\frac{1}{2\alpha}).$$

$\iff$

$$2(1-\alpha L)(1-\alpha\mu+\frac{G}{L\|w^*\|})>-\frac{\alpha\rho G}{\mu}+(1-\alpha L)^2+\alpha^2{\delta_0}^2 \quad \forall\delta_0\in(0,\frac{1}{2\alpha}). \tag{60}$$

For (60)'s LHS, since $\mu\leq L$ and $\alpha\leq\frac{1}{2L}$, we have

$$2(1-\alpha L)(1-\alpha\mu+\frac{G}{L\|w^*\|})>2(1-\alpha L)(1-\alpha L+0)\geq 2(1-\frac{1}{2L}\cdot L)(1-\frac{1}{2L}\cdot L+0)=\frac{1}{2} \quad \forall\delta_0\in(0,\frac{1}{2\alpha}). \tag{61}$$

For (60)'s RHS, since $\alpha\leq\frac{1}{2L}$ and $\delta_0<\frac{1}{2\alpha}$, we have

$$-\frac{\alpha\rho G}{\mu}+(1-\alpha L)^2+\alpha^2{\delta_0}^2<0+(1-\frac{1}{2L}\cdot L)^2+\alpha^2(\frac{1}{2\alpha})^2=\frac{1}{2} \quad \forall\delta_0\in(0,\frac{1}{2\alpha}). \tag{62}$$

Taking the result of (61) and (62) altogether, (60) is true since

$$2(1-\alpha L)(1-\alpha\mu+\frac{G}{L\|w^*\|})>\frac{1}{2}>-\frac{\alpha\rho G}{\mu}+(1-\alpha L)^2+\alpha^2{\delta_0}^2 \quad \forall\delta_0\in(0,\frac{1}{2\alpha}).$$

The proven (55) is equivalent to the desired conclusion. The proof is complete. □

B. Supplementary Experiment Details

This section provides more details about the experimental settings and hyper-parameter choices.

B.1. Few-shot Classification

Algorithm 3 MAML with inner- and outer-level regularization

Require: Datasets $\mathcal{S} = \left\{\mathcal{S}_i^{\text{in}}, \mathcal{S}_i^{\text{out}}\right\}_{i=1}^{m}$; few-shot meta-query batch size K ; the number of training tasks sampled at each round r; the total number of iterations T.

Require: Regularization term $Reg(w, \mathcal{D})$; Inner-level regularization selector $\sigma^{in} \in \{-1, 0, 1\}$, Outer-level regularization selector $\sigma^{out} \in \{-1, 0, 1\}$.

Initialize the model parameters w^0 randomly.
for $t = 0$ to $T - 1$ **do**
  Randomly select r tasks from the set of m available tasks with indices stored in $\mathcal{B}_t$.
  for each sampled task $\mathcal{T}_i$ **do**
    Sample a size K support data batch $\mathcal{D}_i^{t,\text{ in}}$ from $\mathcal{S}_i^{\text{in}}$;
    Sample a size b query data batch $\mathcal{D}_i^{t,\text{ out}}$ from $\mathcal{S}_i^{\text{out}}$;
    (Inner-level) Compute adapted parameters with gradient descent:
      $w'^t_i := w^t - \alpha \nabla_{w^t} \left(\hat{\mathcal{L}} \left(w^t, \mathcal{D}_i^{t,\text{ in}} \right) + \sigma^{in} Reg \left(w^t, \mathcal{D}_i^{t,\text{ in}} \right) \right)$;
    (Outer-level) SGD step for meta-model, save the per-task weight for meta-update:
      $w_i^{t+1} := w^t - \beta_t \nabla_{w^t} \left(\hat{\mathcal{L}} \left(w'^t_i, \mathcal{D}_i^{t,\text{ out}} \right) + \sigma^{out} Reg \left(w'^t_i, \mathcal{D}_i^{t,\text{ out}} \right) \right)$;
  end for
  Meta-update $w^{t+1} := \frac{1}{r} \sum_{i \in \mathcal{B}_t} w_i^{t+1}$
end for
Return: w^T

The experiment setup for Omniglot and Mini-ImageNet follows [1]. **Datasets.** For the few-shot classification task, we experiment on the Mini-Imagenet [10, 15] and Omniglot [8] datasets. The Mini-Imagenet [10] is sampled from ImageNet with 600 instances of 100 classes. Each image is resized into 84×84. In the experiment, the Mini-Imagenet dataset is split into 64 classes for training, 12 classes for validation, and 24 classes for testing. The Omniglot dataset is a collection of 1623 character classes with different alphabets. Each class in the dataset contains 20 instances. The classes are shuffled and divided into the training, validation, and test sets, with 1150, 50, and 423 instances in the experiment. **Models.** We use the classic 4-layer convolution backbone models [1, 4] in the experiments. Each convolution layer has conv-filters of 3×3 size and is followed by batchnorm and max-pooling. For the Omniglot dataset, we use the backbone model with 64 filters in each convotion layer (i.e., the backbone is 64-64-64-64 conv model). For the empirical verification experiment on Mini-ImageNet, the 48-48-48-48 conv backbone model is adopted. And for the experiment that comparing Minimax-MAML and Minimax-MAML++ with other baseline methods on Mini-ImageNet, we use the 64-64-64-64 conv backbone model to make a fairer comparison with other methods. **Training.** All the MAML experiments take 5 inner-steps. In one experiment, the training takes 150 epochs for 64-64-64-64 conv model and 120 epochs for 48-48-48-48 conv model, and each epoch consists of 500 iterations. The task batch size for all Omniglot experiments is 16. Mini-Imagenet experiments use task batch sizes of 4 and 2 for 1-shot and 5-shot experiments, respectively. After each epoch, the model's performance is evaluated on the validation set. When the training is complete, a prediction of the test set is made by the ensemble of the best 5 per-epoch-models on the validation set (following [1], all the MAML-type methods' results are generated under this paradigm). The Adam optimizer is adopted for the model training, with a scheduled learning rate starting from 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.99$. Cross-entropy loss is adopted as the loss function for all the models in the experiments.

Regularization:.

The pseudo-code for implementing MAML with inner/outer-level regularization in the experiment is shown in Algorithm 3. For the Few-shot image classification experiments, the MAML-type methods are sharing the same form of regularization objective (except the first verification experiment isolating the L2-Norm regularizer). The regularization is achieved by combining the L2-Norm regularization and output entropy regularization, i.e., the regularization term $Reg(w, \mathcal{D})$ in Algorithm 3

is given by

$$Reg(w, \mathcal{D}) = -\gamma^{entropy} H(w, \mathcal{D}) + \gamma^{norm} \frac{1}{2} \|w\|^2, \tag{63}$$

where $H(w, \mathcal{D})$ denotes information entropy of the output generated by model w for data batch D:

$$H(w, \mathcal{D}) = -\mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}} \sum_{i=1}^{K} p_w(y = i \mid \mathbf{x}) \log p_w(y = i \mid \mathbf{x}),$$

where K is the number of classes and $p_w(y = i|x)$ represents the probability of prediction to class i generated by model w. Entropy represents the diversity of model output, and using negative entropy as a regularization objective can encourage the model to make more conservative outputs and suppress overconfident outputs, thereby avoiding overfit. Entropy regularization is also referred to as label-smooth sometimes. (*Since entropy needs to be maximized in order to serve as a regularizer, it is necessary to include a negative sign when incorporating it into the loss function.*)

In order to explain entropy regularizer more specifically, we provide a PyTorch-based sample implementation here:

```
class Self_Entropy(torch.nn.Module):
    def __init__(self, reduction = True):
        super(Self_Entropy, self).__init__()
        self.reduction = reduction

    def forward(self, x):
        b = F.softmax(x, dim=1) * F.log_softmax(x, dim=1)
        if self.reduction:
            b = -1.0 * b.mean()
        else:
            b = -1.0 * b.sum()
        return b

self_entropy = Self_Entropy()
```

$\gamma^{entropy}$ and γ^{norm} in (63) are positive hyper-parameters controlling the regularization rate. We use $\gamma^{entropy} = 2.0$ and $\gamma^{norm} = 5e - 4$ for all of the few-shot classification experiments.

In Algorithm 3, the selectors σ^{in} and σ^{out} respectively determine the type of regularization for the inner-level and outer-level. The values of σ^{in} and σ^{out} can be 1, 0 or -1, corresponding to ordinary regularization, non-regularization, and inverted regularization respectively. For instance, in the empirical verification experiment that uses a combined regularizer, to evaluate the effect of inverted inner-level regularization, we set $\{\sigma^{in}, \sigma^{out}\}$ as $\{-1, 0\}$. Similarly, we set $\{\sigma^{in}, \sigma^{out}\}$ as $\{1, 0\}$ to evaluate the effect of ordinary regularization at inner-level.

In terms of other regularization types, we have $\{\sigma^{in}, \sigma^{out}\} = \{0, 1\}$ for *regularize the outer-level*, $\{\sigma^{in}, \sigma^{out}\} = \{1, 1\}$ for *regularize the loss function* (in the limited-tasks experiment), and $\{\sigma^{in}, \sigma^{out}\} = \{-1, 1\}$ for *minimax-meta regularization*. Since the original MAML doesn't have any regularization, it is equivalent to having δ^{in} = 0 and δ^{out} = 0. (see Table 1 and 2 in the main paper.)

It is worth noting that we only add the inner-level inverted regularization during the training phase, and we do not use it for the meta-testing phase. Specifically, during the meta-testing phase, which evaluates the performance of the learned meta-model on new tasks, we only adapt the model without any additional regularization to avoid influencing its task-specific performance.

Implementation.

The implementation of inner- and outer-level regularizations is simple and straightforward, and often involves only modifications to loss functions. Assuming that cross-entropy is used as the classification loss, and the combined regularization term $Reg(w, \mathcal{D}) = -\gamma^{entropy} H(w, \mathcal{D}) + \gamma^{norm} \frac{1}{2} \|w\|^2$ is adopted, we provide a PyTorch implementation example of Minimax-Meta Regularization to further explain the regularizations and demonstrate the simplicity of implementation.

During training, at the inner-level, the invertedly regularized loss $\hat{\mathcal{L}}\left(w^t, \mathcal{D}_i^{t,\text{in}}\right) - Reg\left(w^t, \mathcal{D}_i^{t,\text{in}}\right)$ now can be expressed by $\hat{\mathcal{L}}\left(w^t, \mathcal{D}_i^{t,\text{in}}\right) - (-\gamma^{entropy} H(w^t, \mathcal{D}_i^{t,\text{in}}) + \gamma^{norm} \frac{1}{2} \|w^t\|^2)$, which could be implemented as:

```
    # inner-loop training loss of MAML, with inverted regularization.
    loss = F.cross_entropy(preds, y) - (- gamma_e * self_entropy(preds) + gamma_n * l2_norm(weights))
```

where *preds* is the model's prediction for the input data batch, *y* is the true label batch and *weights* stores the weight values of the model.

Similarly, at the outer-level, the ordinarily regularized loss $\hat{\mathcal{L}}\left(w'^t_i, \mathcal{D}_i^{t,\text{ out}}\right) + Reg\left(w'^t_i, \mathcal{D}_i^{t,\text{ out}}\right)$ now can be expressed by $\hat{\mathcal{L}}\left(w'^t_i, \mathcal{D}_i^{t,\text{ out}}\right) + (-\gamma^{entropy} H(w'^t_i, \mathcal{D}_i^{t,\text{ out}}) + \gamma^{norm} \frac{1}{2}\|w'^t_i\|^2)$, which could be implemented as:

```
    # outer-loop training loss of MAML, with ordinary regularization.
    loss = F.cross_entropy(preds, y) + (- gamma_e * self_entropy(preds) + gamma_n * l2_norm(weights))
```

Since the built-in norm/weight-decay methods in popular libraries usually do not support negative parameters, the $l2\_norm$ function may require manual implementation, but it is also easy to accomplish.

When training is complete, during the testing phase, which evaluates the performance of the learned meta-model on new tasks, we adapt the model to each new task without any additional regularization:

```
    # meta-testing phase loss of MAML, without additional regularization
    loss = F.cross_entropy(preds, y)
```

The aforementioned implementation example can be readily incorporated into the widely-used open-source MAML directory *"How to train your MAML in Pytorch"* proposed in [1], which serves as the basis for our experimental setup.

B.2. Few-shot Regression

The experiment setting follows the few-shot regression experiment in [12]. **Datasets** One synthetic and three real-world few-shot regression datasets are considered. The synthetic dataset is created by a 2-dimensional mixture of Cauchy distributions plus random GP functions. One real-world dataset is SwissFEL [9] which corresponds to Swiss Free Electron Laser's calibration sessions. Another two datasets are from the PhysioNet 2012 challenge [13], which contains time-series data related to patients' health metrics, in particular, the Glasgow Coma Scale (GCS) and the hematocrit value (HCT). Cauchy contains 20 tasks, and each task consists of 20 samples. SwissFel contains 5 tasks, and each task consists of 200 samples. Each Physionet dataset contains 100 tasks, and each task consists of $4 \sim 24$ samples. **Models** We use a fully-connected neural network with 4 layers with each 32 neurons as the base-learner model, aligning with the base-learner structure adopted by other baseline methods. $ReLU$ is used as activation. MAML takes 3 inner steps in our experiment.

Regularization For regression problems, the output entropy used in the classification experiments cannot again be used as the regularization objective. So we adopt L2-Norm as the only regularization objective. Let γ be the parameter controlling the magnitude and direction of the L2-Norm regularization, the inner-level regularization rate parameter γ^{in} is set to negative (inverted regularization) and the outer-level regularization rate parameter γ^{out} is set to positive (ordinary regularization). We use hyper-parameter search to select the value of parameters. Specifically, we use $\{\gamma^{in}, \gamma^{out}\} = \{-1e\text{-}3, 1e\text{-}3\}$, $\{\gamma^{in}, \gamma^{out}\} = \{-1e\text{-}2, 1e\text{-}2\}$, $\{\gamma^{in}, \gamma^{out}\} = \{-5e\text{-}3, 5e\text{-}3\}$, and $\{\gamma^{in}, \gamma^{out}\} = \{-5e\text{-}2, 5e\text{-}2\}$ for Cauchy, SwissFel, Physionet-GCS, and Physionet-HCT experiment respectively. The number of iterations in each experiment is determined using the validation set.

C. Supplementary Experimental Analysis

Due to the space limitation of the main paper, we provide supplementary experimental results in this section, including an additional experiment on Mini-ImageNet few-shot classification with limited tasks, an additional experiment on Meta-dataset with the first-order method and larger backbone, and an additional experiment on meta-reweighting with Minimax-Meta Regularization for robust learning.

C.1. Mini-ImageNet Few-shot Classification with Limited Tasks

To further illustrate the generalization ability of Meta-Minimax regularization, we conduct an experiment to compare it with other common regularization strategies on meta-learning with the limited number of training tasks. The fewer the task number is, the easier the meta-model would overfit.

In the implementation of N-way few-shot classification experiments, in the training phase, each task is sampled by combining N training classes as one N-way classification task. That is, for a dataset with M training classes available, there would be accordingly $\binom{M}{N}$ training tasks available. So we could restrict the number of training tasks by restricting the number of training classes.

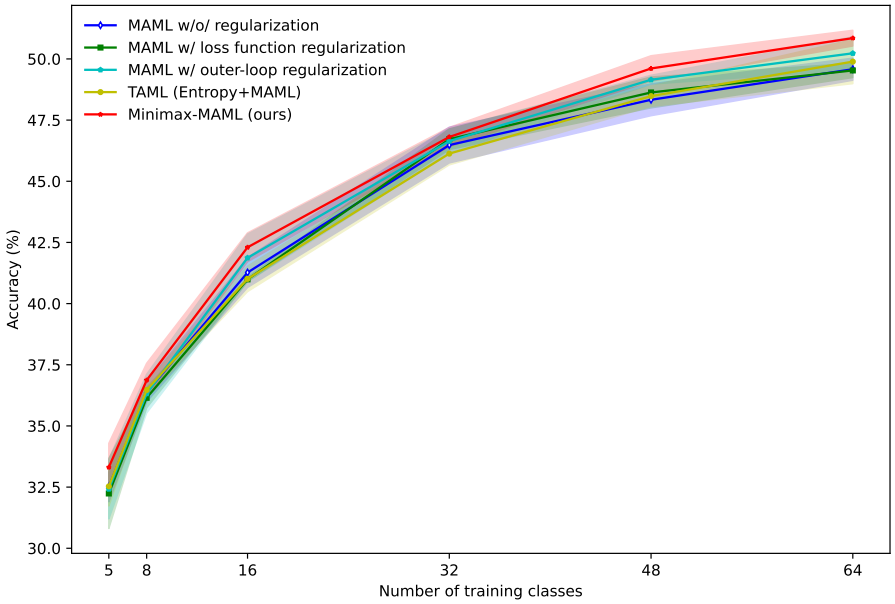

Figure 1. Test accuracies (%) with varying training classes number. The shaded region denotes the 95% confidence interval.

We take Mini-ImageNet 5-way 1-shot as the experiment scenario. The experiment setting follows the same setting of Mini-ImageNet empirical verification experiments. In the original experiment, there are 64 classes available for training. We restrict the number of training classes to 48/32/16/8/5 in this limited classes experiment. And Meta-Minimax regularization is compared with original MAML and MAML with common regularization: *MAML w/ outer-loop regularization, MAML w/ loss function regularization* (*MAML w/ loss function regularization* means simply adding ordinary reg term in loss function at both inner- and outer-level during the training). We also implemented TAML(Entropy+MAML) proposed by [7] for comparison.

Figure 1 shows the accuracy curve of experiment results with the varying number of training classes. The result suggests Meta-Minimax regularization continuously outperforms other methods under the limited task number scenario and improves the accuracy to a certain margin even under a very small task number.

C.2. Meta-Dataset Few-shot Classification Experiment

To test the effectiveness of Minimax-Regularization on larger backbones and to validate if it is fitful for first-order meta-learning methods, We further conduct an experiment using first-order MAML (fo-MAML) and ResNet-12 backbone on Meta-Dataset [14]. Meta-Dataset creates a dataset of datasets benchmark for meta-learning. In our experiment, we only train the model on the ILSVRC training set and test on the ILSVRC testing test and other 8 datasets (Omniglot, Aircraft, Birds, Textures, QuickDraw, VGG Flower, Traffic, MSCOCO). The experiment settings follow the benchmark proposed in [14], and we use the open PyTorch repository provided by [2] to build the experiment. We implement Minimax-Meta Regularization for fo-MAML to compare it with the original version.

Each model takes 6 inner-loop steps in the training phase, and additional 5 inner-loop steps are adopted for the testing phase. The inner stepsize is set to 0.3 for each model. Adam optimizer is adopted for the meta-model updating with a learning rate of 0.00025. Each model completed 10000 training updates. We repeat the experiment for 3 independent runs and report the mean accuracies and 95% confidence intervals.

Like in the Mini-ImageNet and Omniglot experiments, we adopt the entropy & L2-Norm combined regularization term to achieve the Minimax-Meta Regularization. We use $\gamma^{entropy} = 2.0$ and $\gamma^{norm} = 3e\text{-}5$ as the reg magnitude coefficients for both inner- and outer-level regularizations.

Table 6 shows the experiment results. The results suggest that the Minimax-fo-MAML generalizes better on all 9 testing

datasets.

C.3. Meta-reweighting with Minimax Regularization for Robust Learning

To verify the general effectiveness of our proposed methods, we further conduct experiments on the meta-learning problem of meta-reweighting for robust learning.

C.3.1 Experimental Setup

For this experiment, we evaluate the performance of our proposed method and baselines on a robust-learning task: the noisy MNIST dataset. The dataset is created by randomly flipping the labels of 40% of the training images, resulting in 10000 training images with 40% incorrectly labeled data. Each image has a dimension of 28x28, and the task is to classify them into ten handwritten digits $(0 \sim 9)$. There is also a clean validation set consisting of 100 correctly labeled images with balanced categories available for helping the training process on the noisy set.

Algorithm 4 Minimax Meta-Reweighting.

Require: model θ_0, noisy training set D_f, clean validation set D_g, training batch size n, validation batch size m, inner-level regularization parameter γ^{in},outer-level regularization parameter γ^{out}

Ensure: θ_T

for $t = 0$ to $T-1$ **do**
  Sample a n-size mini-Batch data $\{X_f, y_f\}$ from $\mathcal{D}_f$;
  Sample a m-size mini-Batch data $\{X_g, y_g\}$ from $\mathcal{D}_g$;
  Forward X_f using model θ_t, get predicted labels $\hat{y}_f$;
  Set temporary example weights to zero: $\epsilon = 0$;
  Calculate weighted loss on noisy data batch: $l_f = \sum_{i=1}^{n} \epsilon_i C\left(y_{f,i}, \hat{y}_{f,i}\right)$;
  Calculate $\hat{\theta}_t = \theta_t - \alpha \nabla_{\theta_t} l_f$;
  Forward X_g using model $\hat{\theta}_t$, get predicted labels $\hat{y}_g$;
  Evaluate loss on clean data batch, with inverted entropy reg:
    $l_g = \frac{1}{m} \sum_{i=1}^{m} \left(C\left(y_{g,i}, \hat{y}_{g,i}\right) + \gamma^{in} Entropy(\hat{y}_{g,i})\right)$;
  Calculate new example weights $\tilde{w} = \max(-\nabla_{\epsilon} l_g, 0)$, and normalize $w = \frac{\tilde{w}}{\sum_j \tilde{w} + \delta\left(\sum_j \tilde{w}\right)}$;
  Calculate new weighted loss on noisy data batch, with ordinary entropy reg:
    $\hat{l}_f = \sum_{i=1}^{n} w_i \left(C\left(y_{f,i}, \hat{y}_{f,i}\right) - \gamma^{out} Entropy(\hat{y}_{f,i})\right)$;
  $\theta_{t+1} \leftarrow \text{OptimizerStep}\left(\theta_t, \nabla_{\theta_t} \hat{l}_f\right)$;
end for

The basic robust-learning baseline we evaluate here is Meta-Reweighting introduced in [11]. The Meta-Reweighitng algorithm learns to assign weights to training examples for robust learning. To determine the example weights, Meta-Reweighting performs a meta gradient descent step on the mini-batch example weights (which are initialized from zero) to minimize the loss on a clean, unbiased validation set.

Our method adds the Minimax-Meta Regularization on top of Meta-Reweighting. We add ordinary regularization at the outer-level, where the optimal weights are adopted for meta-update. And inverted regularization is added at the inner-level, where the weighted inner-model fits the clean unbiased validation set for optimal weight calculation. Intuitively, through the meta-weighted learning process, such a regularization method makes the model become more conservative when updating based on the noisy training data in the outer loop and values the diversity of predictions more, thereby resisting overfit.

Table 6. Few-shot classification results on **Meta-Dataset** using models trained on ILSVRC. Backbone: **ResNet-12**
Datasets except ILSVRC are only used for testing, and we report the test accuracy with a 95% confidence interval. Each model completed 10000 training updates.

Method	ILSVRC (test)	Omniglot	Aircraft	Birds	Textures	QuickDraw	VGG Flower	Traffic	MSCOCO
fo-MAML	38.24±2.30	44.75±6.26	28.06±2.43	37.64±3.56	39.41±4.50	42.57±3.79	58.55±5.20	36.62±2.85	42.38±5.09
Minimax-fo-MAML(ours)	**40.53±1.54**	**68.43±3.53**	**30.95±2.97**	**41.09±0.40**	**45.12±1.41**	**51.57±2.68**	**66.23±0.89**	**38.83±2.71**	**45.15±0.85**

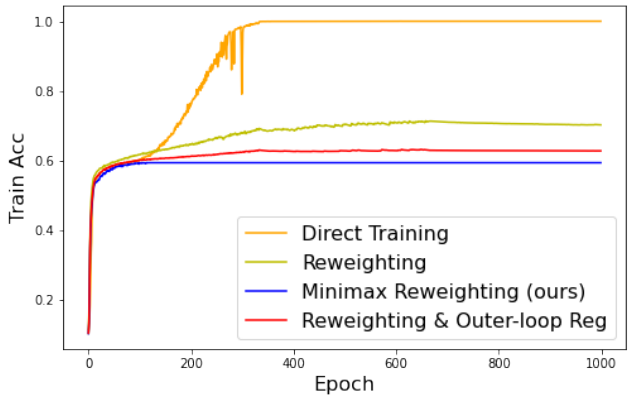

Figure 2. Training accuracy curve. Since 40% of training samples are incorrectly labeled, the model keeps a training accuracy of around 60% would be considered resistant to overfitting during the training.

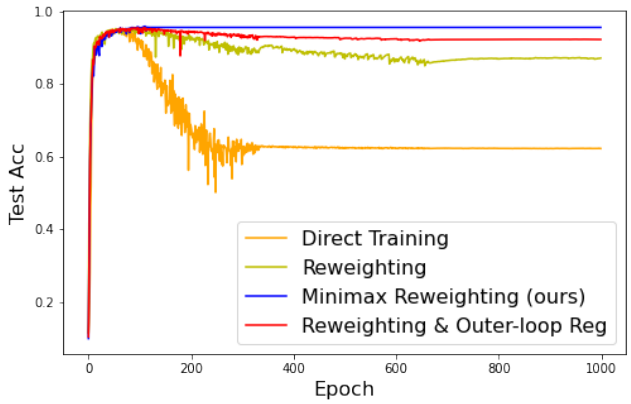

Figure 3. Test accuracy curve. Since the test dataset is clean, the model that can maintain a higher test accuracy is considered with better learning robustness (less affected by the noise in the training set).

At the same time, the inner model was encouraged to make sharper predictions on the clean validation set by the inverted regularization.

The regularization objective used in our method is maximizing output entropy (minimizing output entropy at the inner-level). We call our method Minimax Meta-Reweighting. The pseudo-code for implementation is shown in Algorithm 4. In our experiment, we use a $\gamma^{in} = 0.25$ and $\gamma^{out} = 2.0$.

For each method in the experiment, we use the LeNet-5 as the backbone model and train the model for 1000 epochs. The learning rates for the first 1/3, the middle 1/3, and the last 1/3 training epochs are set to 1e-2, 1e-3, and 1e-4, respectively.

C.3.2 Results and Analysis

Under this setting, models are extremely prone to overfit the noisy dataset during the training phase. To understand the models' performance, we could look at the training and testing curves in Figure 2 and 3. Since the training set is noisy, models overfitted to the train set would show significant performance deduction on the clean test set.

From the perspective of robust learning, the *direct training* method sets the lower performance bound to some extent. Since it does not have any denoising ability, it quickly overfits the training set during the training. It reaches peak accuracy on the clean test set around the 80th epoch, and starts to overfit after that. We could identify the overfitting characteristic from the training and testing accuracy curve. Since 40% of the labels in the training set are incorrect, once the model starts

to predict the training data with an accuracy larger than 60%, it fits the distribution of the noisy training data instead of the ground truth distribution. At the same time, the performance deduction on the clean test set would also start. Finally, we could observe the training accuracy and testing accuracy of the directly trained model to converge to nearly 100% and 60%, respectively, which indicates a complete overfit. On the contrary, the model with optimal learning robustness should not overfit the train set, keep a train accuracy value close to 60% and maintain the optimal performance on the clean test set.

Compared to direct training, the training curve of Meta-Reweighting baseline [11] shows a significant improvement in the learning robustness. However, it still suffers from overfitting. It neither completely overfits the training dataset nor ignores all the noises; its training accuracy converges to around 70%. After around the 100th epoch, the Meta-Reweighting model experienced continual test accuracy deduction and finally maintained test accuracy at around 87.5%.

Minimax-Reweighting nearly reached the optimal learning robustness under this setting. The training accuracy of Minimax-Reweighting stuck at around 60% with rarely any change throughout the training phase. And the testing accuracy maintained a peak value of around 95.5% without observable deduction.

To further evaluate the effectiveness of Minimax-Reweighting, we implemented the outer-loop-only regularization on top of the Meta-Reweighting algorithm to make comparisons. While this approach did show improvement from the baseline method, it was unable to achieve the same level of performance as Minimax-Meta Regularization, as shown in Figure 2 and 3.

References

[1] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. *arXiv preprint arXiv:1810.09502*, 2018. 16, 18

[2] Malik Boudiaf, Ziko Imtiaz Masud, Jérôme Rony, Jose Dolz, Ismail Ben Ayed, and Pablo Piantanida. Mutual-information based few-shot classification, 2021. 19

[3] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks. *Advances in Neural Information Processing Systems*, 34, 2021. 4, 5

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017. 4, 16

[5] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930. PMLR, 2019. 1, 2

[6] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016. 1

[7] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11719–11727, 2019. 19

[8] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 16

[9] Christopher J Milne, Thomas Schietinger, Masamitsu Aiba, Arturo Alarcon, Jürgen Alex, Alexander Anghel, Vladimir Arsov, Carl Beard, Paul Beaud, Simona Bettoni, et al. Swissfel: the swiss x-ray free electron laser. *Applied Sciences*, 7(7):720, 2017. 18

[10] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 16

[11] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343. PMLR, 2018. 20, 22

[12] Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, and Andreas Krause. Pacoh: Bayes-optimal meta-learning with pac-guarantees. In *International Conference on Machine Learning*, pages 9116–9126. PMLR, 2021. 18

[13] Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 Computing in Cardiology*, pages 245–248. IEEE, 2012. 18

[14] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019. 19

[15] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638, 2016. 16