Improving Generalization of Meta-learning with Inverted Regularization at Inner-level

Lianzhe Wang[1]* Shiji Zhou[1]* Shanghang Zhang[2]† Xu Chu[1] Heng Chang[1] Wenwu Zhu[1]†
[1]Tsinghua University [2]National Key Laboratory for Multimedia Information Processing, Peking University.
{wanglz20,zsj17,changh17}@mails.tsinghua.edu.cn,
shanghang@pku.edu.cn, {chu_xu,wwzhu}@tsinghua.edu.cn

Abstract

Despite the broad interest in meta-learning, the generalization problem remains one of the significant challenges in this field. Existing works focus on meta-generalization to unseen tasks at the meta-level by regularizing the meta-loss, while ignoring that adapted models may not generalize to the task domains at the adaptation level. In this paper, we propose a new regularization mechanism for meta-learning – Minimax-Meta Regularization, which employs inverted regularization at the inner loop and ordinary regularization at the outer loop during training. In particular, the inner inverted regularization makes the adapted model more difficult to generalize to task domains; thus, optimizing the outer-loop loss forces the meta-model to learn meta-knowledge with better generalization. Theoretically, we prove that inverted regularization improves the meta-testing performance by reducing generalization errors. We conduct extensive experiments on the representative scenarios, and the results show that our method consistently improves the performance of meta-learning algorithms.

1. Introduction

Meta-learning has been proven to be a powerful paradigm for extracting well-generalized knowledge from previous tasks and quickly learning new tasks [47]. It has received increasing attention in many machine learning settings such as few-shot learning [10, 45, 46, 50] and robust learning [27, 39, 42], and can be deployed in many practical applications [7, 21, 29, 54]. The key idea of meta-learning is to improve the learning ability of agents through a learning-to-learn process. In recent years, optimization-based algorithms have emerged as a popular approach for realizing the learning-to-learn process in meta-learning [10, 28]. These methods formulate the problem as a bi-level optimization problem and have demonstrated impressive performance across various domains, leading to significant attention from the research community. The primary focus of our paper is to further advance this line of research.

The training process of meta-learning takes place at two levels [10, 19]. At the inner-level, a base model, which is initialized using the meta-model's parameters, adapts to each task by taking gradient descent steps over the support set. At the outer-level, a meta-training objective is optimized to evaluate the generalization capability of the initialization on all meta-training tasks over the query set, helping to ensure that the model is effectively optimized for the desired goal. With this learning-to-learn process, the final trained meta-model could be regarded as the model with good initialization to adapt to new tasks.

Despite the success of meta-learning, the additional level of learning also introduces a new source of potential overfitting [36], which poses a significant challenge to the generalization of the learned initialization. This generalization challenge is twofold: first, the meta-model must generalize to unseen tasks (*meta-generalization*); and second, the adapted model must generalize to the domain of a specific task, which we refer to as *adaptation-generalization*. As the primary objective of meta-learning is to achieve strong performance when adapting to new tasks, the ability of the meta-model to generalize well is critical. Recent works aim to address the meta-generalization problem by meta-regularizations, such as constraining the meta-initialization space [52], enforcing the performance similarity of the meta-model on different tasks [20], and augmenting meta-training data [33, 36, 51]. These approaches are verified to enhance generalization to unseen tasks. However, they do not address the problem of adaptation-generalization to the data distribution of meta-testing tasks.

To address this issue, we propose Minimax-Meta Regularization, a novel regularization mechanism that improves both adaptation-generalization and meta-generalization. Specifically, our approach particularly employs inverted regularization at the inner-level to hinder the adapted model's generalizability to the task domain. This forces the

*Equal contributions †Corresponding authors meta-model to learn hypotheses that better generalize to the task domains, which improves adaptation-generalization. Meanwhile, we use ordinary regularization at the outer-level to optimize the meta-model's generalization to new tasks, which helps meta-generalization. By improving both adaptation-generalization and meta-generalization simultaneously, our method results in a more robust and effective meta-learning regularization mechanism.

Theoretically, we prove that under certain assumptions, if we add L2-Norm as the regularization term to the inner-level loss function, the *inverted regularization* will reduce the generalization bound of MAML, while the *ordinary regularization* will increase the generalization bound. In terms of total test error, which includes both generalization error and training bias caused by regularization, the inverted L2-Norm also reduces the total test error when the reg parameter is selected within a negative interval. These results suggest that the regularization at the inner-level should be inverted. As it has been verified that ordinary regularization at the outer-level helps the meta-generalization, our theory implies that the proposed Minimax-Meta Regularization helps both meta-generalization and adaptation-generalization.

We conduct experiments on the few-shot classification problem for MAML [10] with different regularization types (ordinary/inverted) at the inner- and outer-level. The results demonstrate the efficacy of Minimax-Meta Regularization, and support the theoretical results that regularization at the inner-level improves test performance only when it's inverted. Additionally, we empirically verify that Minimax-Meta regularization can be applied with different types of regularization terms (norm/entropy), implying the flexibility for applying the proposed method in practice.

2. Related Work

Meta-learning. A line of meta-learning methods has sought to train recurrent neural networks that ingest entire datasets [8, 41]. However, they need to place constraints on the model architecture. Another line aims to learn a transferable metric space between samples from previous tasks [31, 34, 44, 49]. However, it is mainly limited to classification problems. In this paper, we focus on optimization-based meta-learning methods that learn a meta-initialization [10–13, 18, 26, 28, 35], which are well-generalized for meta-training tasks, being agnostic to both model architecture and problems. However, these approaches are shown to be overfitting the meta-training tasks [6, 40, 51, 53].

Meta-Regularization. Standard regularizations such as weight decay [22], dropout [15], and incorporating noise [1, 2, 48], which can significantly enhance the generality of single-level machine learning. However, it limits the flexibility of fast adaptation in the inner-level [51]. MR-MAML [52] constrains the search space of the meta-model and allows the adaptation to be sufficient at the inner-level. Jamal *et al.* [20] proposed TAML to enforce the meta-model to perform similarly across tasks. Rajenran *et al.* [37] explored an information-theoretic framework of meta-augmentation. Yao *et al.* [51] proposed two task augmentation methods – MetaMix and Channel Shuffle, which is theoretically proven to be generalized to unseen tasks. Ni *et al.* [33] investigated the distinct ways where data augmentation can be integrated at both the image and class levels. Rothfuss *et al.* [40] addressed the meta-generalization problem using the PAC-Bayesian framework. However, these works focus on meta-generalization, while adaptation-generalization is merely considered.

3. Preliminary

Model-Agnostic Meta-Learning (MAML) [10] with a single inner-step is adopted as the representative algorithm to derive the theoretical results in this paper. We follow the framework proposed by Fallah *et al.* [9] to make problem formulation for MAML with a single inner-step. We denote each data point by $z = (x, y) \in \mathcal{Z}$ and evaluate the performance of a model parameterized by $w \in \mathcal{W}$ using loss function $\ell(w, z)$. Tasks $\{\mathcal{T}_i\}_{i=1}^m$ are drawn from distributions $\{\mathcal{P}_i\}_{i=1}^m$, with corresponded *population loss* for model w defined as $\mathcal{L}_i(w) := \mathbb{E}_{z\sim p_i}[\ell(w, z)]$. Throughout the paper, we adopt the hat notation to denote empirical losses, i.e., $\hat{\mathcal{L}}(w; \mathcal{D}) := \frac{1}{|\mathcal{D}|}\sum_{z\in\mathcal{D}} \ell(w, z)$ means the *empirical loss* of model w with dataset $\mathcal{D}$.

$F_i(w)$ is defined to evaluate the performance of the model updated by one single stochastic gradient descent (SGD) from w, on task $\mathcal{T}_i$. $\mathcal{D}_i$ denotes a data batch consisting of K samples drawn from $\mathcal{P}_i$. The goal of MAML is to find a good model parameter w that generally performs well across different tasks after taking the SGD step:

$$\begin{aligned} &\min_{w\in\mathcal{W}} F(w) := \frac{1}{m}\sum_{i=1}^{m} F_i(w) \\ &= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathcal{D}_i}\mathbb{E}_{z\sim p_i}\left[\ell\left(w - \frac{\alpha}{K}\sum_{z'\in\mathcal{D}_i} \nabla\ell\left(w, z'\right), z\right)\right] \end{aligned} \tag{1}$$

However, directly solving (1) is usually impractical since the true task distributions $\{\mathcal{P}_i\}_{i=1}^m$ are usually unknown. Instead, the common practice is to approximate F_i by the empirical loss. For simplicity, suppose we have access to totally $2n$ training samples from each task $\mathcal{T}_i$, and we further group the samples into two distinct sets of size n: $\mathcal{S}_i^{\text{in}}$ for meta-training(support) at inner-level and $\mathcal{S}_i^{\text{out}}$ for meta-validation(query) at outer-level. Then, for each task $\mathcal{T}_i$, we have one corresponding training set $\mathcal{S}_i := \left\{\mathcal{S}_i^{\text{in}}, \mathcal{S}_i^{\text{out}}\right\}$. During training, each distinct K-shot data batch D_i is sampled from each $\mathcal{S}_i^{\text{in}}$ to serve as a meta-training(support) set.

The approximation of (1) is given by

$$\underset{w \in \mathcal{W}}{\arg\min} \hat{F}(w, \mathcal{S}) := \frac{1}{m} \sum_{i=1}^{m} \hat{F}_i\left(w, \mathcal{S}_i\right) \tag{2}$$

where $\mathcal{S} := \{\mathcal{S}_i\}_{i=1}^{m}$. And $\hat{F}_i$ stands for empirical loss that estimates F_i by

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

MAML solves the minimization problem in (2) by using each per-task gradient $\nabla \hat{F}_i\left(w, \mathcal{S}_i\right)$ to take SGD step at meta-level. Specifically, at each iteration t, for each sampled task data $\{\mathcal{D}_i^{t,\text{ in}}, \mathcal{D}_i^{t,\text{ out}}\}$, MAML calculates

$$w_i^{t+1} := w^t - \beta_t \nabla_{w^t} \hat{\mathcal{L}}\left(w^t - \alpha \nabla \hat{\mathcal{L}}\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right), \mathcal{D}_i^{t,\text{ out}}\right) \tag{3}$$

and update the model at the end of each iteration by

$$w^{t+1} := \frac{1}{r} \sum_{i \in \mathcal{B}_t} w_i^{t+1}$$

where $\mathcal{B}_t$ is the set of indices of r randomly chosen tasks at iteration t. When referring to the per-task adapted model in the paper, we denote it as ${w'}_i^t$ and its calculation is in fact embedded within (3), that is, ${w'}_i^t := w^t - \alpha \nabla \hat{\mathcal{L}}\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right)$.

In the context of evaluating the performance of meta-learning algorithms, the test error is generally considered the most critical metric. This error represents the population loss of a meta-model, denoted as $\mathcal{A}(\mathcal{S})$, obtained by algorithm $\mathcal{A}$ with a given dataset $\mathcal{S}$. The test error can be decomposed into three distinct terms:

$$\mathbb{E}_{\mathcal{A},\mathcal{S}}\left[F(\mathcal{A}(\mathcal{S})) - \min_{\mathcal{W}} F\right] \quad \text{(test error)} \ =$$
$$\underbrace{\mathbb{E}_{\mathcal{A},\mathcal{S}}\left[\hat{F}(\mathcal{A}(\mathcal{S}), \mathcal{S}) - \min_{\mathcal{W}} \hat{F}(\cdot, \mathcal{S})\right]}_{\text{training error}}$$
$$+ \underbrace{\mathbb{E}_{\mathcal{A},\mathcal{S}}[F(\mathcal{A}(\mathcal{S})) - \hat{F}(\mathcal{A}(\mathcal{S}), \mathcal{S})]}_{\text{generalization error}} + \underbrace{\mathbb{E}_{\mathcal{S}}\left[\min_{\mathcal{W}} \hat{F}(\cdot, \mathcal{S})\right] - \min_{\mathcal{W}} F}_{\leq 0} \tag{4}$$

Fallah *et al.* [9] have shown that the first training error term will converge to zero as the number of training steps T increases, given that the loss function $\ell(w, z)$ satisfies certain assumptions, and that the third term is non-positive. Therefore, to improve the performance of the obtained model on the test error, we aim to apply regularization to reduce the generalization error term.

4. Method

In this section, we introduce the Minimax-Meta Regularization method for bi-level meta-learning and its application to the popular MAML algorithm. We also provide an intuitive explanation of the effectiveness of the inner-level inverted regularization.

4.1. Minimax-Meta Regularization

Our Minimax-Meta Regularization method is designed to improve the generalization performance of bi-level meta-learning by combining two types of regularizations: one at the outer-level and the other at the inner-level. In particular, we propose to use an ordinary regularization at the outer-level to encourage the meta-model to learn more generalized hypotheses, and an inverted regularization at the inner-level to increase the adaptation difficulty and help the meta-model improve generalization during training.

Specifically, when the regularizations involved can be achieved in the loss function, the Minimax-Meta Regularization shifts the learning objective of the inner level from $\hat{\mathcal{L}}\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right)$ to

$$\hat{\mathcal{L}}\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right) + \sigma^{in} Inverted\_Reg\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right),$$

and the learning objective of the outer level from $\hat{\mathcal{L}}\left({w'}_i^t, \mathcal{D}_i^{t,\text{ out}}\right)$ to

$$\hat{\mathcal{L}}\left({w'}_i^t, \mathcal{D}_i^{t,\text{ out}}\right) + \sigma^{out} Ordinary\_Reg\left({w'}_i^t, \mathcal{D}_i^{t,\text{ out}}\right),$$

where σ^{in} and σ^{out} are regularization coefficients.

The outer-level regularization term $Ordinary\_Reg\,(w, \mathcal{D})$ can be any classic ordinary regularization term, such as L1/L2-Norm or information entropy regularization, which encourages the meta-model to learn more generalized hypotheses. In contrast, the inner-level regularization term $Inverted\_Reg\,(w, \mathcal{D})$ should be an inverted regularization term, which could typically be achieved by changing the sign of an ordinary regularization term (e.g., negative L1/L2-Norm, inverted entropy regularization), and this increases the adaptation difficulty and forces the meta-model to learn better-generalized hypotheses.

It is worth noting that the inner-level inverted regularization is only added 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.

Intuition for Inverted Regularization at Inner-level. The intuition behind using inverted regularization at the inner-level is that it can help the meta-model learn better-generalized hypotheses (meta-knowledge) by increasing the

Algorithm 1 Minimax-MAML

```
Require: Datasets $\mathcal{S} = \left\{\mathcal{S}_i^{\text{in}}, \mathcal{S}_i^{\text{out}}\right\}_{i=1}^{m}$; total number of
    iterations $T$; regularization coefficients $\sigma^{in}$ and $\sigma^{out}$.
Initialize the meta-model $w^0$
for $t = 0$ to $T-1$ do
    Randomly sample $r$ tasks with indices stored in $\mathcal{B}_t$;
    for each sampled task $\mathcal{T}_i$ do
        Sample a support data batch $\mathcal{D}_i^{t,\text{ in}}$ from $\mathcal{S}_i^{\text{in}}$;
        Sample a query data batch $\mathcal{D}_i^{t,\text{ out}}$ from $\mathcal{S}_i^{\text{out}}$;
        (Inner-level) Compute per-task adapted parameters with gradient descent:
            $w'^t_i := w^t - \alpha\nabla_{w^t}\left(\hat{\mathcal{L}}\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right) + \sigma^{in} Inverted\_Reg\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right)\right);$
        (Outer-level) SGD step for meta-model, save per-task meta-weight for meta-update:
            $w_i^{t+1} := w^t - \beta_t\nabla_{w^t}\left(\hat{\mathcal{L}}\left(w'^t_i, \mathcal{D}_i^{t,\text{ out}}\right) + \sigma^{out} Ordinary\_Reg\left(w'^t_i, \mathcal{D}_i^{t,\text{ out}}\right)\right);$
    end for
    Meta-update $w^{t+1} := \frac{1}{r}\sum_{i\in\mathcal{B}_t} w_i^{t+1}$
end for
Return: $w^T$
```

adaptation difficulty during training. Specifically, by making the adapted model more difficult to learn a generalized hypothesis by fitting the meta-support set, the meta-model is forced to learn better-generalized meta-knowledge to achieve good performance on the meta-query set. In this sense, we can think of the Minimax-Meta Regularization as a form of "adversarial training" for the meta-model, which can improve its generalization performance during training. Importantly, the "adversarial training" is only applied during the training phase and is not used during meta-testing. Thus, the meta-model does not carry the "adversarial training" burden in the actual deployment after learning better-generalized meta-knowledge, which can lead to better generalization in the new environment.

While the concept of using inverted regularization at the inner-level to improve generalization may seem too intuitional or counterintuitive to some, we provide a theoretical analysis in the next section to support its utility.

4.2. Application to MAML

To apply Minimax-Meta Regularization to MAML, we modify the MAML algorithm by adding the regularization to the inner- and outer-level training objective. The modified algorithm, which we refer to as Minimax-MAML, is shown in Algorithm 1. Note that this modification for Minimax-Meta Regularization is also generally applicable to other MAML variants.

5. Theoretical Analysis

In this section, we provide an analysis of the effectiveness of inverted regularization in meta-learning by taking L2-Norm regularization at the inner-level of the single-step MAML algorithm as a typical example, which is very possible to generalize to other regularization.

It is important to note that the process of adding regularization often involves changes to the loss function during training. This means if the model is obtained by a new regularized algorithm $\tilde{\mathcal{A}}$, it is usually optimized for a different function $\hat{\hat{F}}(\cdot)$ instead of the original $F(\cdot)$ (e.g., added weight-norm in the inner-level). However, in the meta-testing phase, the model's test error is still calculated using $F(\cdot)$. As a result, to evaluate the test error change with a new regularized method $\tilde{\mathcal{A}}$, instead of directly adopting (4)'s decomposition in Preliminary, we need to further decompose the test error by

$$
\begin{aligned}
&\mathbb{E}_{\tilde{\mathcal{A}},\mathcal{S}}\left[F(\tilde{\mathcal{A}}(\mathcal{S})) - \min_{\mathcal{W}} F\right] \quad \text{(test error)} \; = \\
&\underbrace{\mathbb{E}_{\tilde{\mathcal{A}},\mathcal{S}}\left[\hat{F}(\tilde{\mathcal{A}}(\mathcal{S}),\mathcal{S}) - \hat{F}(\arg\min_{\mathcal{W}}\hat{\hat{F}}(\cdot,\mathcal{S}),\mathcal{S})\right]}_{\text{training error}} \\
&+ \underbrace{\mathbb{E}_{\tilde{\mathcal{A}},\mathcal{S}}[F(\tilde{\mathcal{A}}(\mathcal{S})) - \hat{F}(\tilde{\mathcal{A}}(\mathcal{S}),\mathcal{S})]}_{\text{generalization error}} \\
&+ \underbrace{\mathbb{E}_{\mathcal{S}}\left[\min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\right] - \min_{\mathcal{W}} F}_{\leq 0} \\
&+ \underbrace{\mathbb{E}_{\mathcal{S}}\left[\hat{F}(\arg\min_{\mathcal{W}}\hat{\hat{F}}(\cdot,\mathcal{S}),\mathcal{S}) - \min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\right]}_{\text{training bias}}
\end{aligned}
\tag{5}
$$

where $\hat{\hat{F}}(\cdot)$ refers to the regularized empirical loss function. (5) has one more training bias term compared to (4), which is caused by the changing of the objective function. Usually, regularization would reduce the expected generalization error while increasing the training bias. The goal of regularization is to decrease test error by reducing generalization error while trading off training bias.

Adding L2-Norm regularization at the inner-level for MAML could be obtained by changing the inner-level training objective from $\hat{\mathcal{L}}\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right)$ to $(\hat{\mathcal{L}}\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right) + \frac{\delta}{2}\|w^t\|^2)$, where δ is the regularisation parameter. The meta updating rule would be accordingly changed from (3) to:

$$
\begin{aligned}
&w_i^{t+1} := \\
&w^t - \beta_t\nabla_{w^t}\hat{\mathcal{L}}\left(w^t - \alpha\nabla_{w^t}(\hat{\mathcal{L}}\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right) + \frac{\delta}{2}\|w^t\|^2)), \mathcal{D}_i^{t,\text{ out}}\right)
\end{aligned}
\tag{6}
$$

Here δ can be either positive or negative to represent the ordinary and inverted regularization, respectively. We treat

δ as a variable and analyze how its value would influence the generalization error and the training bias of the total error introduced in (5).

The analysis of generalization error closely follows the work of [9], and holds the same assumptions about function $\ell(\cdot, z)$ and task distribution as follows.

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 $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$*

5.1. Generalization Error

We derive our generalization bound for MAML with L2 regularization at the inner-level through the theoretical framework proposed by [9], which mainly adopts an algorithmic stability approach for the derivation. We denote the algorithm combines MAML with inner-level regularization as $\tilde{\mathcal{A}}$, and the below generalization bound could be obtained. We provide detailed proof in Appendix.

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) satisfies*

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

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

The generalization bound could be regarded as a function $GB(\delta)$, and its derivative $GB'(\delta)$ is positive $\forall \delta \in (-\infty, \frac{1}{2\alpha})$[1]. It suggests that $GB(\delta)$ is monotonically increasing if $\delta \in (-\infty, \frac{1}{2\alpha})$, implying that L2 regularization at the inner-level decreases the generalization bound of MAML only when it's inverted (i.e. $\delta < 0$). And ordinary regularization (i.e. $\delta \in (0, \frac{1}{2\alpha})$) at the inner-level would increase the generalization bound.

5.2. Training Bias

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*

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

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

The training bias bound could also be regarded as a function $TB(\delta)$. We could observe that $TB(\delta) > TB(0) = 0$ for $\delta \neq 0$, which suggests that training bias is inevitable when regularization is adopted. Another important finding is that for any legal choice of $\delta_0 > 0$, we have $TB(-\delta_0) < TB(\delta_0)$ [2], which suggests that the inverted regularization has less corruption to training bias bound at the inner-level than the ordinary regularization with the same coefficient.

5.3. Test Error

Since the training error term in the test error (5) vanishes with iteration T as long as the outer-level loss is strongly-convex [9], the training error term could be negligible for $\delta < \frac{1}{2\alpha}$. So we could just consider the training bias and generalization error for bounding the test error, i.e.,

$$\mathbb{E}_{\tilde{\mathcal{A}},\mathcal{S}}\left[F(\tilde{\mathcal{A}}(\mathcal{S})) - \min_{\mathcal{W}} F\right] \leq \\ \underbrace{\frac{2G^2(1 + \alpha L)(1 - \alpha\mu - \alpha\delta + (2 + \alpha L - \alpha\delta)\alpha LK)}{mn} \cdot \left(\frac{1}{\alpha\rho G + (1 - \alpha\mu - \alpha\delta)^2 L} + \frac{1}{-\alpha\rho G + (1 - \alpha L - \alpha\delta)^2 \mu}\right)}_{\text{generalization error bound } GB(\delta)} \\ + \underbrace{\frac{\alpha^2(\alpha\rho G + (1 - \alpha\mu)^2 L)((1 - \alpha\mu - \alpha\delta)L\|w^*\| + G)^2 \delta^2}{2(-\alpha\rho G + (1 - \alpha L - \alpha\delta)^2 \mu)^2}}_{\text{training bias bound } TB(\delta)}$$

The test error bound could be described by $TE(\delta) := TB(\delta) + GB(\delta)$. When δ is positive, we have $TB(\delta) > TB(0)$ and $GB(\delta) > GB(0)$ (since $GB'(\delta) > 0 \ \forall \delta \in (-\infty, \frac{1}{2\alpha})$), which suggests ordinary regularization at the inner-level worsens the model's test error bound. Instead, for inverted regularization, since $TE'(0) = TB'(0) + GB'(0) = 0 + GB'(0) > 0$, there must be an interval $[\delta^*, 0)$ in which all values can be used as the inverted regularization parameter to decrease the test error bound.

[1] Derivation is included in Appendix A.3.1. $\delta \geq \frac{1}{2\alpha}$ are excluded from discussion because they may break the convexity of the meta loss function.

[2] Derivation is included in Appendix A.3.2

6. Experiments

We conduct extensive experiments on three types of classical meta-learning tasks: few-shot classification, few-shot regression, and robust reweighting. The experiments include: **i)** an empirical verification of the regularization at inner- and outer-level on the Mini-Imagenet few-shot classification task, which demonstrates the effectiveness of both the inverted regularization at inner-level and the ordinary regularization at outer-level; **ii)** further experiments on few-shot classification and regression benchmarks to compare our Minimax-Meta regularized algorithms with other representative methods; **iii)** a few-shot learning experiment on a limited number of tasks evaluating generalization of different regularization strategies; and **iv)** an experiment on meta-reweighting for robust learning, which demonstrates the broad applicability of our method to different meta-learning problems. *(Due to page-size limitations, the experiments on limited tasks, Meta-Dataset with larger backbones, and meta-reweighting are included in the Appendix)*

6.1. Few-shot Classification

We first conduct experiments on the few-shot classification task, one of the most popular tasks to evaluate meta-learning algorithms. To verify the effectiveness of our approach, we adapt Minimax-Meta Regularization into bi-level optimization meta-learning algorithms and make a benchmark to compare with other methods.

6.1.1 Experimental Setup

Datasets. For the few-shot classification task, we experiment on the Mini-Imagenet [38, 49] and Omniglot [23] datasets. The Mini-Imagenet [38] is sampled from ImageNet with 600 instances of 100 classes. 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.

Experimental details. We select MAML [10] as the representative bi-level optimization meta-learning algorithm to conduct the experiment. The few-shot benchmark settings for Omniglot and Mini-ImageNet experiments provided in [4] are adopted for our experiment build. Details about the experiment can be found in Appendix.

To verify the theoretical results and show the effectiveness of our regularization design, we first conduct an empirical verification experiment on Mini-ImageNet using L2-Norm as the regularizer.

In other few-shot classification experiments, we use a combination of L2-Norm and output-entropy as the regularizer to further improve the generalization. (Although we only use L2-Norm as the sample regularizer to derive the theoretical results in Section 5, the use of inverted regularization can cover many other regularizers in practice, including the entropy regularizer.) That is, in this part of the experiment, when we say that ”adding ordinary regularization” at a certain level, its corresponding learning objective will include minimizing the L2-Norm of model weights and maximizing the entropy of the model’s output prediction (improves generalization); when we say that ”adding inverted regularization” at a certain level, its corresponding learning objective will include maximizing of the L2-Norm of model weights and minimizing the entropy of the model’s output prediction (hinders generalization). And we keep the magnitude of the L2-Norm parameter = 5e-4 and the magnitude of the information entropy parameter = 2.0 across the experiments, i.e., the difference between ordinary and inverted regularization in this group of few-shot learning experiments is only the sign of the regularization term. Note that we only add regularization at the training phase, so the inner-levels are not regularized in meta-testing time.

6.1.2 Empirical Verification for regularization at inner- and outer-level.

To verify our view that the regularization at the inner- and outer-level should respectively be inverted and ordinary, we conduct two experiments for MAML [10] with different regularization methods on Mini-Imagenet 5-way few-shot problem. There are five regularization methods being compared: *no regularization*, *regularize the outer-level*, *regularize the inner-level*, *invertedly regularize the inner-level*, and *Minimax-Meta Regularization*. In the first experiment, We only use L2-Norm regularization to match the setting of theoretical analysis. In the second experiment, We use L2-Norm & entropy combined regularization to verify whether inverted inner-level regularization is suitable for different types of regularizers and whether a combination of multiple regularizers leads to better generalization. We follow [4]’s setting to build the experiment with 48-48-48-48 conv backbone and use the ensemble of per-epoch models to generate more stable results (MAML baseline achieves higher performance under this setting compared to classic 32-32-32-32 conv backbone implementations [10]), The results are respectively presented in Table 1 and 2. Based on the results, we make the following observations:

Inner-level inverted regularization enhances the generalization performance. Compare the results from “no regularization” and “invertedly regularize the inner-level”, we observe that adding inner inverted regularization achieves accuracy improvements in both 1-shot and 5-shot experiments, which verifies the efficacy of the inner inverted regularization. This is aligned with our intuition and theoretical

Table 1. Test accuracy of MAML with different types of regularization in the Mini-Imagenet 5-way MAML Few-shot Classification experiment (*L2-Norm as regularization objective only*). Backbone: 48-48-48-48 conv. We report the test accuracy with a 95% confidence interval for the mean.

Mini-Imagenet 5-way Few-shot Classification for MAML *(Reg Objective: L2-Norm)*				
Regularization Type	Outer Reg	Inner Reg	1-Shot	5-Shot
no regularization	-	-	49.58±0.45%	65.39±0.50%
regularize the outer-level	*Ordinary*	-	49.90±0.54%	66.47±1.21%
regularize the inner-level	-	*Ordinary*	49.28±0.37%	64.80±0.25%
invertedly regularize the inner-level	-	*Inverted*	49.92±0.42%	66.05±0.68%
Minimax-Meta Regularization	*Ordinary*	*Inverted*	**50.25±0.38%**	**68.17±0.92%**

Table 2. Test accuracy of MAML with different types of regularization in the Mini-Imagenet 5-way MAML Few-Shot Classification experiment (*Combining L2-Norm and output entropy as regularization objective*). Backbone: 48-48-48-48 conv. We report the test accuracy with a 95% confidence interval for the mean.

Mini-Imagenet 5-way Few-Shot Classification for MAML *(Reg Objective: L2-Norm & Entropy)*				
Regularization Type	Outer Reg	Inner Reg	1-Shot	5-Shot
no regularization	-	-	49.58±0.45%	65.39±0.50%
regularize the outer-level	*Ordinary*	-	50.23±0.67%	67.18±0.88%
regularize the inner-level	-	*Ordinary*	48.07±1.01%	64.32±0.35%
invertedly regularize the inner-level	-	*Inverted*	49.96±0.33%	65.91±0.41%
Minimax-Meta Regularization	*Ordinary*	*Inverted*	**50.85±0.37%**	**69.36±0.34%**

result.

Inner-level ordinary regularization impairs the generalization performance. Compare the results from "no regularization" and "regularize the inner-level", we observe that adding inner ordinary regularization suffers from accuracy impairments. This observation is also consistent with our intuition and theoretical findings.

Outer-level ordinary regularization enhances the generalization performance. Compare the results from "no regularization" and "regularize the outer-level", we observe that adding outer regularization can get accuracy improvements, which verifies the efficacy of adding ordinary regularization at the outer-level.

The outer-level ordinary regularization and inner-level inverted regularization are compatible. We observe that Minimax-Meta Regularization outperforms solely outer-level or inverted inner-level regularization, indicating compatibility between the regularizations at the two distinct levels. This aligns with the intuition that meta and adaptation generalization are not conflicting.

Inner-level inverted regularization and the outer-level ordinary regularization are suitable for combined regularizer We observe consistent effects across L2-Norm regularizer and L2-Norm & entropy combined regularizer when using different regularization strategies. Furthermore, combining the L2-Norm and entropy regularizer led to improved performance compared to using L2-Norm regularizer alone.

6.1.3 Minimax-Meta Regularization for Few-shot Classification

So far, we have proved that Minimax-Meta Regularization is a promising regularization strategy for bi-level meta-learning. Here, we do experiments to further test the effectiveness of Minimax-Meta Regularization.

The experiments are conducted on Omniglot and Mini-ImageNet datasets. We implement Minimax-Meta Regularization for bi-level meta-learning algorithms: *MAML* [10], which is the most representative bi-level meta-learning algorithm; *MAML++* [4], which is an adapted version of MAML with additional techniques for performance improvements. L2-Norm & entropy combined regularizer is adopted in this experiment.

Representative algorithms with comparable backbone structures are selected for making the comparison. We use the 64-64-64-64 conv backbone for the Mini-ImageNet experiment to make a fairer comparison with other methods. The results are shown in Table 3 and 4.

The results suggest that Minimax-Meta Regularization generally improves test performances. Minimax-MAML++ achieves the best performance on both datasets.

6.2. Minimax-Meta Regularization for Few-shot Regression

We then conduct experiments on the few-shot regression task to test the efficacy of Minimax-Meta Regularization.

Table 3. Omniglot 20-way 1-shot experiment. We report the test accuracy with a 95% confidence interval for the mean.

*the * indicates result generated in our experiment.*

Omniglot 20-way 1-Shot Classification	
	Accuracy
Meta-SGD [28]	95.93±0.38%
Prototypical Net [44]	96.00%
Meta-Networks [32]	97.00%
GNN [16]	97.40%
Relation Network [46]	97.60±0.20%
R2-D2 [5]	96.24±0.05%
SNAIL [31]	97.64±0.30%
TAML(Entropy) [20]	95.62±0.50%
MAML [10]*	94.20±0.41%
Minimax-MAML(ours)*	95.76±0.39%
MAML++ [4]*	97.21±0.51%
Minimax-MAML++(ours)*	**97.77±0.06%**

Table 4. Mini-Imagenet 5-way few-shot experiment. We report the test accuracy with a 95% confidence interval for the mean.

*the * indicates result generated in our experiment.*

Mini-Imagenet 5-way Few-Shot Classification			
Approach	Backbone	1-Shot Accuracy	5-Shot Accuracy
Meta-SGD [28]	64-64-64-64	50.47±1.87%	64.03±0.94%
Prototypical Nets [44]	64-64-64-64	49.42±0.78%	68.20±0.66%
GNN [16]	64-96-128-256	50.33±0.36%	66.41±0.63%
R2-D2 [5]	64-64-64-64	49.50±0.20%	65.40±0.20%
LR-D2 [5]	96-192-384-512	51.90±0.20%	68.70±0.20%
MetaOptNet [25]	64-64-64-64	53.23±0.59%	69.51±0.48%
TAML(Entropy) [20]	64-64-64-64	51.73±1.88%	66.05±0.85%
MAML-Meta Dropout [24]	32-32-32-32	51.93±0.67%	67.42±0.52%
MAML-MMCF [51]	32-32-32-32	50.35±1.82%	64.91±0.96%
MAML [10]*	64-64-64-64	50.20±1.65%	65.86±0.61%
Minimax-MAM(ours)*	64-64-64-64	51.70±0.42%	68.41±1.28%
MAML++ [4]*	64-64-64-64	52.96±0.78%	70.02±0.55%
Minimax-MAML++(ours)*	64-64-64-64	**53.28±0.35%**	**71.70±0.23%**

6.2.1 Experimental Setup

Datasets. We follow the few-shot regression experiment setting proposed in [40] to build the experiment. One synthetic and three real-world few-shot regression datasets are included. The synthetic dataset is created by a 2-dimensional mixture of Cauchy distributions plus random GP functions. One real-world dataset is SwissFEL [30] which corresponds to Swiss Free Electron Laser's calibration sessions. Another two datasets are from the PhysioNet 2012 challenge [43], which contains time-series data related to patients' health metrics, in particular, the Glasgow Coma Scale (GCS) and the hematocrit value (HCT).

Experimental details. We implement Minimax-MAML for the regression task by adding inverted and ordinary L2-Norm at the inner-level and outer-level of MAML, respectively. To obtain optimal results, unlike the single-inner-step MAML implemented in [40], we perform three inner update steps for the meta-training of Minimax-MAML. In order to verify the effect of minimax, we also compared the results of unregularized MAML with three inner steps.

6.2.2 Experimental Results

As shown in Table 5, the Minimax-Meta Regularization improved the performance in all four datasets. And Minimax-MAML achieves near-best performance on the synthetic Cauchy datasets and outperforms other algorithms on the two Physionet datasets. The results suggest that the Minimax-Meta Regularization could improve the performance of the few-shot regression task for meta-learning.

Table 5. Test RMSE comparison of algorithms in four meta-learning environments for few-shot regression.

*the * indicates the result generated in our experiment, other results are reported from [40]*

	Cauchy	SwissFel	Physionet-GCS	Physionet-HCT
MLL-GP [14]	0.216±0.003	0.974±0.093	1.654±0.094	2.634±0.144
MLAP [3]	0.219±0.004	0.486±0.026	2.009±0.248	2.470±0.039
NP [17]	0.224±0.008	0.471±0.053	2.056±0.209	2.594±0.107
PACOH-GP [40]	0.209±0.008	0.376±0.024	1.498±0.081	2.361±0.047
PACOH-NN [40]	**0.195±0.001**	**0.372±0.002**	1.561±0.061	2.405±0.017
MAML [10](1 inner step)	0.219±0.004	0.730±0.057	1.895±0.141	2.413±0.113
MAML [10](3 inner steps)*	0.212±0.003	0.535±0.042	1.532±0.074	2.396±0.047
Minimax-MAML*	0.201±0.002	0.477±0.026	**1.483±0.052**	**2.343±0.019**

7. Conclusion

This paper studies the generalization problem of bi-level optimization-based meta-learning. While most of the existing works focus on meta-generalization to unseen tasks at the meta-level, they leave out that adapted models may not be generalized to the task domain at the adaptation-level. We give an intuitive explanation of why the inverted regularization at the inner-level could improve the adaptation generalization of meta-learning. We provide theoretical support for this intuition by deriving generalization error and training bias bound. We empirically verify that both *inverted regularization at inner-level* and *ordinary regularization at outer-level* improve the test performance of meta-learning. Based on the aligned theoretical and empirical results, we propose meta-learning with Minimax-Meta Regularization, combining regularization at inner- and outer-level. Finally, we conduct experiments on multiple meta-learning tasks to show the efficacy of the proposed method.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China No. 2020AAA0106300, National Natural Science Foundation of China (No. 62250008, 62222209, 62102222, 61936011). Shanghang would like to thank the support from CCF-DiDi GAIA Collaborative Research Funds for Young Scholars. We gratefully thank Jiaming Liu, Yuzhou Cao and Wenpeng Zhang for their valuable discussions.

References

[1] Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018. 2

[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 2

[3] Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended pac-bayes theory. In *International Conference on Machine Learning*, pages 205–214. PMLR, 2018. 8

[4] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. *arXiv preprint arXiv:1810.09502*, 2018. 6, 7, 8

[5] Luca Bertinetto, Joao F Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. 8

[6] Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Task-robust model-agnostic meta-learning. *arXiv preprint arXiv:2002.04766*, 2020. 2

[7] Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. Investigating meta-learning algorithms for low-resource natural language understanding tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1192–1197, 2019. 1

[8] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016. 2

[9] 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. 2, 3, 5

[10] 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. 1, 2, 6, 7, 8

[11] Chelsea Finn and Sergey Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. In *International Conference on Learning Representations*, 2018. 2

[12] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9537–9548, 2018. 2

[13] Sebastian Flennerhag, Andrei Rusu, Razvan Pascanu, Francisco Visin, Hujun Yin, and Raia Hadsell. Meta-learning with warped gradient descent. In *International Conference on Learning Representations 2020*, 2020. 2

[14] Vincent Fortuin and Gunnar Rätsch. Deep mean functions for meta-learning in gaussian processes. *arXiv preprint arXiv:1901.08098*, 2019. 8

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2

[16] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018. 8

[17] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018. 8

[18] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *International Conference on Learning Representations*, 2018. 2

[19] Mike Huisman, Jan N Van Rijn, and Aske Plaat. A survey of deep meta-learning. *Artificial Intelligence Review*, 54(6):4483–4541, 2021. 1

[20] 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. 1, 2, 8

[21] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019. 1

[22] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957, 1992. 2

[23] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 6

[24] Hae Beom Lee, Taewook Nam, Eunho Yang, and Sung Ju Hwang. Meta dropout: Learning to perturb latent features for generalization. In *International Conference on Learning Representations*, 2020. 8

[25] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10649–10657. IEEE Computer Society, 2019. 8

[26] Yoonho Lee and Seungjin Choi. Gradient-based meta-learning with learned layerwise metric and subspace. In *International Conference on Machine Learning*, pages 2927–2936. PMLR, 2018. 2

[27] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5051–5059, 2019. 1

[28] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017. 1, 2, 8

[29] Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, 2019. 1

[30] 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. 8

[31] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, 2018. 2, 8

[32] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *International Conference on Machine Learning*, pages 2554–2563. PMLR, 2017. 8

[33] Renkun Ni, Micah Goldblum, Amr Sharaf, Kezhi Kong, and Tom Goldstein. Data augmentation for meta-learning. In *International Conference on Machine Learning*, pages 8152–8161. PMLR, 2021. 1, 2

[34] Boris N Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: task dependent adaptive metric for improved few-shot learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 719–729, 2018. 2

[35] Eunbyung Park and Junier B Oliva. Meta-curvature. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3314–3324, 2019. 2

[36] Janarthanan Rajendran, Alexander Irpan, and Eric Jang. Meta-learning requires meta-augmentation. In *Advances in Neural Information Processing Systems*, pages 5705–5715, 2020. 1

[37] Janarthanan Rajendran, Alex Irpan, and Eric Jang. Meta-learning requires meta-augmentation. *arXiv preprint arXiv:2007.05549*, 2020. 2

[38] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 6

[39] 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. 1

[40] 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. 2, 8

[41] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016. 2

[42] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in Neural Information Processing Systems*, 32:1919–1930, 2019. 1

[43] 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. 8

[44] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4080–4090, 2017. 2, 8

[45] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019. 1

[46] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 1, 8

[47] Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012. 1

[48] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015. 2

[49] 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. 2, 6

[50] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020. 1

[51] Huaxiu Yao, Long-Kai Huang, Linjun Zhang, Ying Wei, Li Tian, James Zou, Junzhou Huang, et al. Improving generalization in meta-learning via task augmentation. In *International Conference on Machine Learning*, pages 11887–11897. PMLR, 2021. 1, 2, 8

[52] Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. Meta-learning without memorization. *arXiv preprint arXiv:1912.03820*, 2019. 1, 2

[53] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7343–7353, 2018. 2

[54] Tianhe Yu, Chelsea Finn, Annie Xie, Sudeep Dasari, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *arXiv preprint arXiv:1802.01557*, 2018. 1