Consistent MetaReg: Alleviating Intra-task Discrepancy for Better Meta-knowledge

Pinzhuo Tian , **Lei Qi** , **Shaokang Dong** , **Yinghuan Shi** and **Yang Gao** *
National Key Laboratory for Novel Software Technology, Nanjing University, China
{tianpinzhuo, qilei.cs, shaokangdong}@gmail.com, {syh, gaoy}@nju.edu.cn

Abstract

In the few-shot learning scenario, the data-distribution discrepancy between training data and test data in a task usually exists due to the limited data. However, most existing meta-learning approaches seldom consider this intra-task discrepancy in the meta-training phase which might deteriorate the performance. To overcome this limitation, we develop a new consistent meta-regularization method to reduce the intra-task data-distribution discrepancy. Moreover, the proposed meta-regularization method could be readily inserted into existing optimization-based meta-learning models to learn better meta-knowledge. Particularly, we provide the theoretical analysis to prove that using the proposed meta-regularization, the conventional gradient-based meta-learning method can reach the lower regret bound. The extensive experiments also demonstrate the effectiveness of our method, which indeed improves the performances of the state-of-the-art gradient-based meta-learning models in the few-shot classification task.

1 Introduction

Learning quickly is a kind of ability of human intelligence, *e.g.*, children can recognize objects only from a few examples. However, this poses a great challenge to the existing deep learning models, which require large-scale training data to achieve promising performance. To tackle this problem, in recent years, meta-learning (*i.e.*, learning to learn) has drawn increasing interest in the machine learning community [Finn *et al.*, 2017; Rajeswaran *et al.*, 2019]. The goal of these methods is to learn the meta-knowledge across tasks, which can help model learn fast or adapt quickly in new tasks.

For meta-learning approaches, the core issue is – *How to learn effective meta-knowledge*. Regarding that the training data is usually limited in few-shot learning, these few data cannot describe the real data distribution, which results in the intra-task data-distribution discrepancy between training data and test data in each task. However, most existing meta-learning approaches usually ignore this discrepancy during the meta-knowledge learning phase. In this case, the learned inferior meta-knowledge might have a disadvantage to the final performance.

To this end, we propose a new Consistent Meta-regularization (CM) method, which could alleviate the intra-task discrepancy in the meta-training course. The motivation of our method is illustrated in Fig. 1. Although the across-task meta-knowledge can help the base-learner trained by the training data fit well on the test data, the base-learner indeed cannot work well on the test data because of the discrepancy. Therefore, we put forward the Consistent MetaReg to alleviate the intra-task discrepancy to help base-learner work well on the test data in each task for learning better meta-knowledge. Moreover, the proposed meta-regularization method can be easily integrated into existing gradient-based meta-learning models. Concretely, we obtain the base-learner for training data as the traditional meta-learning approaches. In particular, these traditional methods only utilize test data to update the meta-learner. Differently, our method considers that using test data to update both the meta-learner and base-learner, *i.e.*, we reversely use the test data as training data and acquire a new base-learner model for test data. To migrate the data distribution discrepancy between training data and test data in each task, we expect that two base-learner models trained on training data and test data to be consistent. Therefore, we develop Consistent Meta-regularization to achieve this goal.

In this paper, we demonstrate the efficacy of the Consistent Meta-regularization from the theoretical and experimental perspectives, respectively. On the theoretical aspect, we prove that using the proposed method, the conventional gradient-based meta-learning models can indeed achieve lower regret bound than before. In addition, we also validate CM on three benchmark datasets, *i.e.*, miniImageNet, tiered-ImageNet and office31. The experimental results show that our regularization method could learn more useful meta-knowledge by migrating the intra-task discrepancy. Integrated with our Consistent Meta-regularization, three existing state-of-the-art optimization-based meta-learning approaches perform better on all datasets.

In sum, our contributions in this paper are listed as follows:

- To the best of our knowledge, our method is the first one to consider the intra-task discrepancy problem in the typical meta-learning models, which is seldom touched

*Corresponding author

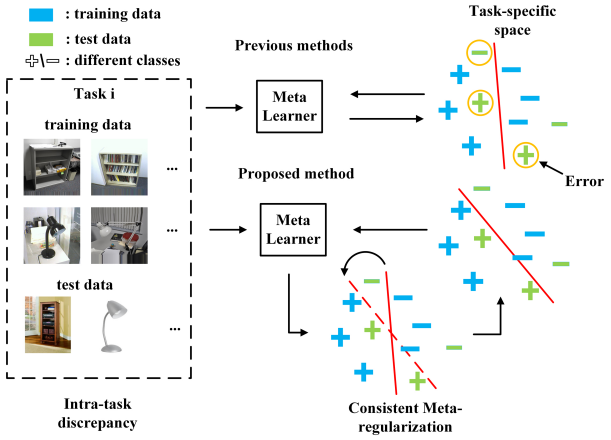

Figure 1: Concept. We utilize the 2-way few-shot classification task as an example.

in previous works.

- We develop Consistent Meta-regularization to reduce the intra-task discrepancy, which can be easily introduced into the existing gradient-based meta-learning models.
- We demonstrate the meta-learning methods with the proposed CM can have lower regret bound in theory, and extensive experiments also highlight that adding CM into the conventional meta-learning frameworks can indeed achieve better performance.

2 Related Work

The topic of meta-learning (or learning to learn) was introduced and studied several decades ago [Schmidhuber, 1987; Bengio *et al.*, 1992; Thrun and Pratt, 1998]. Early work mostly focused on learning how to dynamically adjust the inductive bias [Utgoff, 1986] or hypothesis space for a learning algorithm [Vilalta and Drissi, 2002]. In recent years, two-level framework is popular in the present meta-learning approaches. The motivation of this framework is that meta-level (meta-learner) is used to learn meta-knowledge, which can help base-level (task-specific model or base-learner) learn fast or adapt quickly in a new task. The two-level framework can be broadly divided into three categories.

- **Metric-based method.** In these methods, a non-parametric similarity function is used as base-learner to evaluate the similarity between examples. The meta-learner is trained to learn useful meta-knowledge in the predefined metric space, *e.g.*, Euclidean distance based prototypical networks [Snell *et al.*, 2017], cosine similarity based recurrence with attention mechanisms [Vinyals *et al.*, 2016].
- **Model-based method.** The meta-learner is usually designed as a parameterized predictor to generate base-learner parameters. For example, Ravi *et al*. [Ravi and Larochelle, 2017] used recurrent neural network as meta-learner to direct the updating for base-learner. Munkhdalai *et al*. [Munkhdalai and Yu, 2017] designed a meta-learner, which uses loss gradients from base-learner to predict parameters for base-learner.
- **Gradient-based method.** Finn *et al*. [Finn *et al.*, 2017] proposed model-agnostic meta learning (MAML) for deep models. MAML is similar to the initialization of deep networks. The meta-learner aims to learn a good initialization for all tasks, and the base-learner merely requires a few gradient steps from this initialization to achieve great performance. However, there are still many limitations for MAML. Some works [Li *et al.*, 2017; Rajeswaran *et al.*, 2019; Na *et al.*, 2019] were developed to further improve it. Besides MAML, some gradient-based methods [Bertinetto *et al.*, 2019; Lee *et al.*, 2019] leveraged bilevel framework to optimize meta-learning algorithms.

Our work is the most related to gradient-based methods. Compared with metric-based methods, gradient-based methods can be broadly applied in many areas, such as reinforcement learning [Rakelly *et al.*, 2019] and NLP [Xie *et al.*, 2019]. Moreover, the gradient-based methods do not introduce additional parameters or require a particular learner architecture. These factors result in that gradient-based methods become a promising and hot research topic in meta-learning recently. However, the previous gradient-based methods ignore the intra-task discrepancy, which hinders meta-learning models to effectively learn meta-knowledge. The proposed Consistent Meta-regularization method aims to overcome this limitation, which can be integrated into MAML or bilevel gradient-based methods and enhance their performance.

3 Gradient-based Meta-learning Method

In this section, we firstly introduce the problem formulation for meta-learning in the context of supervised learning. Then, two important kinds of gradient-based meta-learning approaches are enumerated.

3.1 Meta-Learning Problem Formulation

The problem formulation of meta-learning follows [Amit and Meir, 2018]. We assume that all tasks share the sample space $\mathcal{Z}$, hypothesis space $\mathcal{H}$ and loss function $\ell : \mathcal{H} \times \mathcal{Z} \to [0, 1]$. N different tasks are sampled from an unknown task distribution τ as meta-training set $\{T_1, \cdots, T_n\}$. And each task T_i has an observed dataset S_i^{tr}. The goal of the meta-learner is to extract meta-knowledge from the meta-training set so that learning a new task only needs few training data. We denote that the meta-knowledge P comes in the form of a distribution over hypotheses. During learning a new task, the base-learner uses the task-specific observed dataset S^{tr} and the meta-knowledge P to learn a base-learner $Q(P, S^{\text{tr}})$ over $\mathcal{H}$. The quality of meta-knowledge P is measured by the expected loss for learning new tasks, as defined by

$$er(P, \tau) := \mathbb{E}_{T\sim\tau}\mathbb{E}_{S^{\text{tr}}\sim T}\mathbb{E}_{h\sim Q(P,S^{\text{tr}})}\mathbb{E}_{z\sim T}\ell(h, z). \quad (1)$$

If we assume the meta-knowledge P is sampled from the hyper-posterior $\mathcal{Q}(P)$, we can ideally obtain the best meta-knowledge P^* via minimizing the following formula as

$$er(\mathcal{Q}, \tau) := \mathbb{E}_{P\sim\mathcal{Q}}(P, \tau). \quad (2)$$

However, Eq. 2 is not computable. Recent meta-learning algorithms use the idea from [Vinyals *et al.*, 2016]: "*make the test and train conditions much match*". The observed data S_i^{tr} in each task T_i is viewed as training dataset. Besides S_i^{tr}, a test dataset S_i^{ts} is also sampled from T_i. Then the algorithms learn good meta-knowledge by minimizing the empirical risk

$$\hat{er}(\mathcal{Q}, T_1, \cdots, T_n) := \mathbb{E}_{P\sim\mathcal{Q}} \frac{1}{n} \sum_{i=1}^{n} \hat{er}(Q(P, S_i^{\text{tr}}), S_i^{\text{ts}}). \quad (3)$$

3.2 MAML

Model-agnostic meta-learning (MAML) is a pretty significant gradient-based meta-learning method, which has been widely applied to many fields [Al-Shedivat *et al.*, 2018; Javed and White, 2019]. The initialization of model is regarded as meta-knowledge in MAML. Thus, the goal of MAML is to meta-learn the initial model parameter $\boldsymbol{\theta}$ to generalize over the task distribution τ. The empirical loss function in MAML is

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; S_i^{\text{tr}}); S_i^{\text{ts}}), \quad (4)$$

where α is the stepsize. When encountering a new task T_j, the task-specific predictor $\boldsymbol{\theta}_j$ can be easily obtained in a single (or a few) gradient step from the inital $\boldsymbol{\theta}$.

3.3 Bilevel Method

Recently, the bilevel gradient-based method has also attracted many attentions, which achieves the state-of-the-art performance in many computer vision tasks [Lee *et al.*, 2019; Tian *et al.*, 2019]. Bilevel meta-learning framework is proposed in [Franceschi *et al.*, 2018], and its formulation is as

$$\min\{f(\lambda) : \lambda \in \Lambda\}, \quad (5)$$

where function $f : \Lambda \to \mathbb{R}$ is defined at $\lambda \in \Lambda$ as

$$f(\lambda) = \inf\{E(w_\lambda, \lambda) : w_\lambda \in \arg\min_{u\in\mathbb{R}^d} \mathcal{L}_\lambda(u)\}. \quad (6)$$

In Eq. , $E : \mathbb{R}^d \to \mathbb{R}$ is the outer objective. For every $\lambda \in \Lambda$, $L_\lambda : \mathbb{R}^d \to \mathbb{R}$ is the inner objective. The two-level meta-learning algorithm can be easily formulated as a bilevel problem. We can use the outer objective to learn meta-knowledge, and the inner objective is utilized to learn base-learner. Then the empirical loss Eq. 3 can be rewritten as

$$\begin{gathered}\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{\theta}, \text{Alg}(\boldsymbol{\theta}, S_i^{\text{tr}}); S_i^{\text{ts}}), \\ \text{where } \text{Alg}(\boldsymbol{\theta}, S_i^{\text{tr}}) = \min_{\mathbf{w}} \mathcal{L}_{\text{base}}(\mathbf{w}; \boldsymbol{\theta}, S_i^{\text{tr}}).\end{gathered} \quad (7)$$

In Eq. 7, $\boldsymbol{\theta}$ and $\mathbf{w}$ are the meta-parameter and the base-learner parameter, respectively. $\mathcal{L}$ and $\mathcal{L}_{\text{base}}$ denote the meta-loss and task-specific loss, respectively. And Alg is a base-learner algorithm, which computes task-specific parameter based on meta-parameter $\boldsymbol{\theta}$. In fact, MAML can be regarded a special case of bilevel gradient-based meta-learning method, which is analyzed in [Rajeswaran *et al.*, 2019].

Moreover, since both MAML and bilevel gradient-based method can be optimized by gradient descent, it means these approaches can be easily implemented in existing deep learning frameworks (*e.g.*, Pytorch and Tensorflow).

4 Our Method

4.1 Consistent Meta-regularization

As mentioned above, how to learn good meta-knowledge, which can generalize well over the task distribution τ, is very important for meta-learning algorithm. Although MAML or bilevel based methods can indeed learn meta-knowledge by minimizing Eq. 4 and Eq. 7, respectively, all of them ignore the intra-task discrepancy in the meta-training set (*i.e.*, the data-distribution discrepancy between training data and test data in each task because few samples in the few-shot setting cannot describe the real data-distribution of a dataset), which is not conducive to learn good meta-knowledge.

Our proposed Consistent Meta-regularization (CM) method aims to mitigate the impact of the intra-task discrepancy on gradient-based meta-learning models. Specially, we firstly train the base-learner for the training data S_i^{tr} in the task T_i.Then we reversely exploit the test data S_i^{ts} from the task T_i as training data, and train a new base-learner for the test data. For simplicity, two base-learners for training data S_i^{tr} and S_i^{ts} in task T_i are denoted as M_i^{tr} and M_i^{ts}, respectively. Finally, we minimize the difference between M_i^{tr} and M_i^{ts} as a regularization which can be inserted into the traditional meta-loss to alleviate the intra-task discrepancy. Differently, the test data S_i^{ts} is just used to minimize the meta-loss in traditional gradient-based meta-learning.

In this paper, we directly use F-norm of the difference between parameters of M_i^{tr} and M_i^{ts} to measure the gap between the two models. And we can easily integrate the proposed regularization into existing gradient-based meta-learning models (*i.e.*, MAML and bilevel method). For example, we can define the Consistent Meta-regularization loss for bilevel method as

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{\theta}, M_i^{\text{tr}}(\boldsymbol{\theta}, S_i^{\text{tr}}); S_i^{\text{ts}}) + \delta \|M_i^{\text{tr}} - M_i^{\text{ts}}\|_F, \quad (8)$$

where δ is the regularization parameter. If we consider that the base-learner is a deep model [Finn *et al.*, 2017], which contains K layers, the Consistent Meta-regularization loss can be rewritten as

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; S_i^{\text{tr}}); S_i^{\text{ts}}) + \delta \sum_{j=1}^{K} \|(L_j)_i^{\text{tr}} - (L_j)_i^{\text{ts}}\|_F, \quad (9)$$

where L_j is the j-th layer in base-learner. The Consistent Meta-regularization gradient-based meta-learning algorithm is summarized in Alg. 1.

Remark. The traditional gradient-based meta-learning approaches (GBML) can be considered as a bilevel problem. The meta-learner object is used to learn meta-knowledge, and the base learner utilizes the meta knowledge to learn a base model in the task-specific space. In the meta-training phase, these GBML approaches firstly optimize the base-learner, then utilize test dataset S_i^{ts} in each task T_i to minimize meta-loss to learn cross-task meta-knowledge. However, if the base-learner learned by training data in task-specific space cannot generalize well on test data, the meta-loss computed by test data will contain two errors (*i.e.*, true meta-knowledge

Algorithm 1 The proposed CM for meta supervised learning

Require: τ: distribution over tasks, regularization parameter δ, meta-learner step size η, gradient-based meta-learning algorithm: GBML.
Output: Meta-parameter $\boldsymbol{\theta}$

```
Randomly initialize θ.
while not converged do
    Sample mini-batch of tasks {T_i}_{i=1}^B ~ τ.
    for each task T_i do
        Sample N datapoints S_i^tr = {x^m, y^m} from T_i.
        Train a base-learner M_i^tr for S_i^tr by GBML.
        Sample K datapoints S_i^ts = {x^k, y^k} from T_i.
        Train a base-learner M_i^ts for S_i^ts by GBML.
    end for
    Update meta-parameter with gradient descent:
    θ ← θ − η (1/B) Σ_{i=1}^B ∇_θ (L(S_i^ts) + δ‖M_i^tr − M_i^ts‖_F).
end while
```

error and model discrepancy error). Thus, it is difficult to obtain good meta-knowledge. Particularly, this problem will be worse when the training data and test data in each task have a large domain discrepancy. Our proposed regularization eliminates the intra-task discrepancy by pulling M_i^{tr} and M_i^{ts} (*i.e.*, the models for training data and test data) closer in the task-specific space. Therefore, based on our method, the meta-loss is merely from the true meta-knowledge error, which can guide the meta-learner to extract better meta-knowledge.

4.2 Regret Bound for Consistent MetaReg

To further explain the efficacy of our method, a theoretical analysis is given in this part. Specifically, we prove our method can achieve a lower regret bound when compared with the method without using Consistent Meta-regularization in the online convex optimization framework.

We use the same notation with Sec. 3.1, and the S_i^{ts} in each task T_i contains m_i samples, *i.e.*, $S_i^{\text{ts}} = \{(x_{i,j}, y_{i,j}), \cdots, (x_{i,m_i}, y_{i,m_i})\}$. In addition, we assume that the learner incurs the loss $\hat{\ell}_{i,j} := \ell(\hat{y}_{i,j}, y_{i,j})$, when a sample $(x_{i,j}, y_{i,j})$ is revealed.

Definition 1. *The prediction error of a task T_i is*

$$\hat{L}_i = \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{\ell}_{i,j}. \tag{10}$$

The average error (empirical risk) of a meta-learning algorithm after N tasks is

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{\ell}_{i,j}. \tag{11}$$

Given meta-knowledge P, we already know the best predictor h_i^* for task T_i. The regret of a task T_i is

$$\begin{aligned}\mathcal{R}_i(P) &= \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{\ell}_{i,j} - \inf_{h_i\in Q(P,S_i^{\text{ts}})}\sum_{j=1}^{m_i}\ell(h_i, y_{i,j}) \\ &= \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{\ell}_{i,j} - \sum_{j=1}^{m_i}\ell(h_i^*, y_{i,j}).\end{aligned} \tag{12}$$

Then, we introduce the definition of online convex programming problem, following [Zinkevich, 2003].

Definition 2. *An online convex programming problem consists of a feasible set $F \subseteq R^n$ and a infinite sequence $\{f_1, f_2, \cdots\}$, where each $f_t : F \to \mathbb{R}$ is a convex function.*

At each time step t, an online player selects a vector $\mathbf{x}_t \in F$. After the vector is selected, it receives the cost function f_t.

Lemma 1. *[Hazan* et al.*, 2007] If we use gradient descent to select vector at each step,* i.e.*, $\mathbf{x}_t = \Pi_F(\mathbf{x}_{t-1} - \eta_t\nabla f_{t-1}(\mathbf{x_{t-1}}))$. Here, Π_F denotes the projection onto nearest point in F, $\eta_t > 0$ is the step size.*

$$R_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \inf_{\mathbf{x}\in F}\sum_{t=1}^{T} f_t(\mathbf{x}) \le \frac{1}{2}\frac{G^2}{\sigma}\log T,$$

where $\|\nabla f\| \le G$ and $\nabla^2 f \succeq \sigma I$. G and σ are positive constants.

Next, we use Lemma 1 to prove Theorem 1.

Theorem 1. *If we use parameter $\boldsymbol{\theta}$ in a convex set Θ to learn meta-knowledge, and $\ell : \Theta \to \mathbb{R}$ is a convex function. $\boldsymbol{\theta}$ is optimized by gradient descent. And the within-task algorithm has a regret bound $\mathcal{R}_i(\boldsymbol{\theta}) \le \beta(\boldsymbol{\theta}, m_i)$ for any $\boldsymbol{\theta}$, then*

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{\ell}_{i,j} \le &\min_{\boldsymbol{\theta}\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\sum_{j=1}^{m_i}\ell(h_i^*, y_{i,j}) \\ &+ \frac{1}{N}\sum_{i=1}^{N}\beta(\boldsymbol{\theta}, m_i) + \frac{1}{2}\frac{G^2}{\sigma}\frac{\log N}{N}.\end{aligned}$$

Proof.

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{\ell}_{i,j} = \frac{1}{N}\sum_{i=1}^{N}\hat{L}_i \\ &\le \min_{\boldsymbol{\theta}\in\Theta}(\frac{1}{N}\sum_{i=1}^{N}\hat{L}_i) + \frac{1}{2}\frac{G^2}{\sigma}\frac{\log N}{N} \quad \text{(Lemma. 1)} \\ &= \min_{\boldsymbol{\theta}\in\Theta}(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\sum_{j=1}^{m_i}\hat{\ell}_{i,j}) + \frac{1}{2}\frac{G^2}{\sigma}\frac{\log N}{N} \\ &\le \min_{\boldsymbol{\theta}\in\Theta}\frac{1}{N}\sum_{i=1}^{N}(\beta(\boldsymbol{\theta}, m_i) + \inf_{h_i\in Q(P,S_i^{\text{ts}})}\frac{1}{m_i}\sum_{j=1}^{m_i}\ell_{i,j}) \\ &\quad + \frac{1}{2}\frac{G^2}{\sigma}\frac{\log N}{N} \\ &= \min_{\boldsymbol{\theta}\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\sum_{j=1}^{m_i}\ell(h_i^*, y_{i,j}) + \frac{1}{N}\sum_{i=1}^{N}\beta(\boldsymbol{\theta}, m_i) \\ &\quad + \frac{1}{2}\frac{G^2}{\sigma}\frac{\log N}{N}.\end{aligned}$$

□

Theorem 1 shows that the regret bound of average error is related with h_i^*. As mentioned in Sec. 3.1, the traditional meta-learning algorithm uses S_i^{tr} to train the task-specific model M_i^{tr} for task T_i. However, the intra-task discrepancy

causes that the model trained by S_i^{tr} is far from h_i^*. Since h_i^* denotes the best predictor for S_i^{ts} in the test set, our proposed Consistent Meta-regularization supposes that M_i^{ts} (*i.e.*, the model trained by the test set S_i^{ts}) is close to h_i^*. Thus, our method aims to pull the M_i^{tr} closer to M_i^{ts}, which can reduce the gap between M_i^{tr} and h_i^* to achieve lower regret bound.

5 Experiments

In this part, we first describe the implementation details. Then, we present results on three benchmark datasets for the few-shot classification task, including two derivatives of ImageNet [Russakovsky *et al.*, 2015] (*i.e.*, miniImageNet [Vinyals *et al.*, 2016], tieredImageNet [Ren *et al.*, 2018]) and office31 [Saenko *et al.*, 2010]. Particularly, office-31 involves the larger intra-task data-distribution discrepancy when compared with miniImageNet and tieredImageNet. Finally, we analyze the impact of our method on learning speed, and the parameter sensitivity of regularization parameter δ.

5.1 Implement Details

We use the same architecture as the embedding model used in [Vinyals *et al.*, 2016], which has 4 modules with a 3×3 convolutions and 64 filters, followed by batch normalization, a ReLU nonlinearity, and 2×2 max-pooling in all experiments. And all the images are resized to 84×84. Adam with learning rate 0.001 is used as optimizer. During the meta-training process, we sample 5 training (support) samples and 5 test (query) samples as S_i^{tr} and S_i^{ts} in each task T_i, respectively. The number of test example for each task is set as 10 during the meta-testing phase. All methods are trained on a single NVIDIA 1080 Ti.

We insert the proposed CM method into three state-of-the-art gradient-based meta-learning approaches, *i.e.*, MAML, R2-D2 [Bertinetto *et al.*, 2019] and MetaOptNet-SVM [Lee *et al.*, 2019] to validate its effectiveness. Particularly, all tricks in R2-D2 and MetaOptNet-SVM aiming to improve the performance are not used in our experiments, such as learnable scale parameter to adjust the prediction score[1].

5.2 Experiments on ImageNet Derivatives

MiniImageNet is a standard benchmark for few-shot image classification, consisting of 100 randomly chosen classes from ILSVRC-2012 [Russakovsky *et al.*, 2015]. These classes are randomly split into 64, 16 and 20 classes for meta-training, meta-validation, and meta-testing respectively. Each class contains 600 images. The tieredImageNet benchmark is a larger subset of ILSVRC-2012, composed of 608 classes grouped into 34 high-level categories. These are divided into 20 categories for meta-training, 6 categories for meta-validation, and 8 categories for meta-testing. This corresponds to 351, 97 and 160 classes for meta-training, meta-validation, and meta-testing respectively.

We perform the pre-processing for the two datasets following [Lee *et al.*, 2019]. Because of the slow convergence rate of MAML, we trained MAML 60,000 iterations on miniImageNet and tieredImagNet. The other methods are trained by 30,000 and 50,000 iterations on miniImageNet and tieredImageNet, respectively. The meta batch-size is set as 4 and 8 for MAML and the other methods. The parameter δ is set as 1 for MAML and 5 for MetaOptNet-SVM and R2-D2.

Results. Tab. 1 reports the results on the 5-way miniImageNet and tieredImageNet. All reported results are average performances with 95% confidence interval over 2000 tasks randomly sampled from meta-testing split. OptNet denotes metaOptNet-SVM, ProNet is the Prototypical Networks. And MAML-CM, *etc.* represent the methods integrated with our Consistent Meta-regularization. The methods are separated into two groups: optimization-based (**O**) and metric-based (**M**). As seen, by mitigating the intra-task discrepancy, our proposed CM can improve the average performance of MAML, OptNet and R2D2, 0.30% (0.33%), 1.10% (1.90%) and 2.66% (2.59%) on miniImageNet (tieredImageNet), respectively. Particularly, ProNet achieves the best performance, which is not very surprised due to the superiority of metric-based methods for few-shot image classification on ImageNet derivatives [Chen *et al.*, 2019]. However, metric-based meta-learning approach is not convenient to be extended to other fields, *e.g.*, NLP, semantic segmentation and object detection, because it is very difficult to design an appropriate metric method.

5.3 Experiments on Office31

Office-31 is a standard benchmark for visual domain adaptation, consisting of 4,652 images in 31 classes collected from three domains: *Amazon* (**A**), which contains images downloaded from amazon.com, *Webcam* (**W**) and *DSLR* (**D**), which contain images taken by a web camera and a digital SLR camera, respectively, in an office environment. And we randomly divide office31 into three splits which contain 20, 5 and 6 classes as meta-training, meta-validation, and meta-testing, respectively. For each task T_i, we sample training dataset S_i^{tr} from two domains, which are randomly sampled from **A**, **W** and **D**, and test dataset S_i^{ts} is sampled from the rest of domain. MAML is trained in the same setting with Sec. 5.2. And the other models are trained by 30,000 tasks. The meta batch-size is set as 4 for all the methods. As for regularization parameter, we use the same set with Sec. 5.2.

Results. The results on office31 are shown in Tab. 2. All reported results are average performances with 95% confidence interval over 1000 tasks randomly sampled from meta-testing split. Because the training data S_i^{tr} and test data S_i^{ts} in each task T_i are from different domains. There is a large intra-task data-distribution discrepancy on this dataset. Our CM method can boost the average performance of MAML, OptNet, and R2D2 by 3.14%, 3.37% and 2.72% on office31. Moreover, our R2D2-CM model achieves the best performance. Compared with the results on ImageNet derivatives, ProNet cannot handle this situation of a large data-distribution discrepancy due to its limitation. As for optimization-based methods, the data-distribution discrepancy impact greatly on MAML, and we can find similar results in [Finn and Levine, 2018]. The proposed regularization can better promote optimization-based methods in this case, *i.e.*, there is a large intra-task difference.

[1] Source code: https://github.com/P1nzhuo/Consistent-MetaReg

	Models	miniImageNet 5-way 1-shot	5-shot	10-shot	Mean	tieredImageNet 5-way 1-shot	5-shot	10-shot	Mean
M	ProNet	46.66 ± 0.46	69.44 ± 0.39	74.48 ± 0.37	63.53 ± 0.41	48.99 ± 0.51	70.74 ± 0.44	74.51 ± 0.41	64.75 ± 0.45
O	MAML	**47.38 ± 0.46**	61.25 ± 0.44	65.31 ± 0.40	57.98 ± 0.43	46.28 ± 0.50	62.75 ± 0.46	**67.12 ± 0.43**	58.72 ± 0.46
	MAML-CM	46.94 ± 0.48	**61.84 ±0.42**	**66.06 ± 0.41**	**58.28 ± 0.44**	**47.31 ±0.49**	**62.78 ± 0.46**	67.06 ±0.43	**59.05 ± 0.46**
O	OptNet	42.93 ± 0.45	58.60 ± 0.41	62.83 ± 0.41	54.79 ± 0.42	43.40 ± 0.49	59.14 ± 0.45	63.40 ± 0.44	55.31± 0.46
	OptNet-CM	**44.70 ± 0.45**	**59.28 ± 0.43**	**63.66 ± 0.42**	**55.88 ± 0.43**	**45.86 ± 0.51**	**60.84 ± 0.45**	**64.92 ± 0.45**	**57.21 ± 0.47**
O	R2D2	42.80 ± 0.44	56.39 ± 0.41	60.22 ± 0.40	53.14 ± 0.42	44.65 ± 0.49	59.61 ± 0.44	63.20 ± 0.44	55.82 ± 0.46
	R2D2-CM	**44.92 ± 0.45**	**60.68 ± 0.41**	**61.81 ± 0.41**	**55.80 ± 0.42**	**47.59 ± 0.49**	**62.49 ± 0.44**	**65.16 ± 0.44**	**58.41± 0.46**

Table 1: Classification results on miniImageNet and tieredImageNet.

	Models	office-31 5-way 1-shot	5-shot	10-shot	Mean
M	ProNet	37.62 ± 0.97	55.89 ± 1.36	59.87 ± 1.54	51.13 ± 1.29
O	MAML	30.77 ± 0.78	44.50 ± 1.02	49.25 ± 1.19	41.51 ± 1.00
	MAML-CM	**32.41 ± 0.85**	**47.94 ± 1.13**	**53.59 ± 1.37**	**44.65 ± 1.12**
O	OptNet	34.57 ± 0.92	48.39 ± 1.21	52.25 ± 1.41	45.07 ± 1.18
	OptNet-CM	**37.42 ± 0.97**	**52.19 ± 1.33**	**55.70 ± 1.51**	**48.44± 1.27**
O	R2D2	37.45 ± 1.04	55.95 ± 1.48	59.80 ± 1.67	51.07 ± 1.40
	R2D2-CM	**39.98 ± 1.01**	**58.95 ± 1.46**	**62.45 ± 1.64**	**53.79± 1.37**

Table 2: Classification results on office-31.

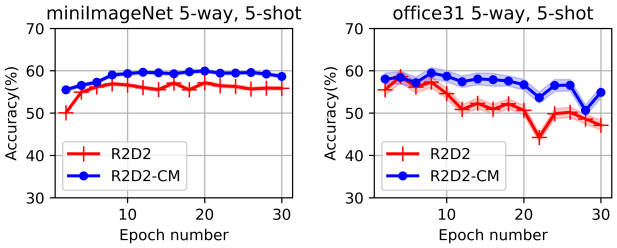

Figure 2: Accuracy of R2D2 and R2D2-CM (ours) in different epochs on the meta-testing sets of miniImageNet and office31.

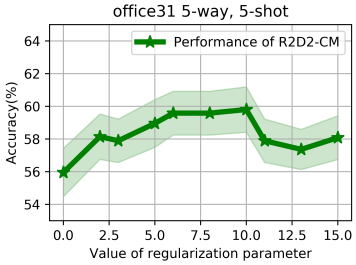

Figure 3: The performance of R2D2-CM with different regularization parameter on the meta-testing set of office31.

5.4 Learning Efficiency Analysis

In this section, we reveal that with our proposed meta-regularization, the traditional gradient-based meta-learning approach can converge faster and achieve better performance.

We report the results of R2D2 and R2D2-CM on meta-testing split of miniImageNet and office31 at different training epochs, respectively, as shown in Fig. 2. Shaded region denotes 95% confidence interval. When conducting the comparison between R2D2 and R2D2-CM (ours) on 5-way, 5-shot classification task on two datasets, we can observe that our method outperforms the R2D2 at the beginning of training (about 6th epoch), and maintain higher performance until convergence. In general, our method can help model converge faster and achieve better performance. Because office31 is smaller than miniImageNet (the number of images is just 7.7% of miniImageNet), with training epoch increasing, meta-overfitting happens in both two models.

5.5 Parameter Sensitivity Analysis

In order to evaluate the influence of meta-regularization parameter on our method, we train our R2D2-CM with different values of regularization parameter δ on office31. And the performance of these models on meta-testing split is in Fig. 3. Shaded region denotes 95% confidence interval.

Our method outperforms the traditional R2D2 approach (regularization parameter $= 0$) with different values of regularization parameter. And the performance of our method doesn't change much from $\delta = 6$ to $\delta = 10$. It can confirm that our model has a lager robust interval on δ.

6 Conclusion

In this paper, we consider the intra-task discrepancy issue in the traditional meta-learning models, which is usually ignored in previous works. To handle this issue, we introduce Consistent Meta-regularization to alleviate the discrepancy for gradient-based meta-learning approaches. Moreover, we prove that with our proposed meta-regularization, the traditional meta-learning approach can achieve lower regret bound in the convex setting. Furthermore, extensive experiments on three datasets reveal the superiority of our method in the non-convex setting. Particularly, in this paper, we follow recent meta-learning approaches, which use deep models as meta-learner in the non-convex setting. In the future work, we will further give the theoretical analysis in the non-convex setting.

Acknowledgments

The work is supported by Science and Technology Innovation 2030-"New Generation Artificial Intelligence" Major Project (No. 2018AAA0100905) and NSFC (No. 1673203).

References

[Al-Shedivat *et al.*, 2018] Maruan Al-Shedivat, Trapit Bansal, Yura Burda, Ilya Sutskever, Igor Mordatch, and Pieter Abbeel. Continuous adaptation via meta-learning in nonstationary and competitive environments. In *ICLR*, 2018.

[Amit and Meir, 2018] Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended pac-bayes theory. In *ICML*, pages 205–214, 2018.

[Bengio *et al.*, 1992] Samy Bengio, Yoshua Bengio, Jocelyn Cloutier, and Jan Gecsei. On the optimization of a synaptic learning rule. In *Preprints Conf. Optimality in Artificial and Biological Neural Networks*, volume 2, 1992.

[Bertinetto *et al.*, 2019] Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *ICLR*, 2019.

[Chen *et al.*, 2019] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019.

[Finn and Levine, 2018] Chelsea Finn and Sergey Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. In *ICLR*, 2018.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017.

[Franceschi *et al.*, 2018] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *ICML*, pages 1563–1572, 2018.

[Hazan *et al.*, 2007] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *ML*, 69(2-3):169–192, 2007.

[Javed and White, 2019] Khurram Javed and Martha White. Meta-learning representations for continual learning. In *NeurIPS*, pages 1818–1828, 2019.

[Lee *et al.*, 2019] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, pages 10657–10665, 2019.

[Li *et al.*, 2017] 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.

[Munkhdalai and Yu, 2017] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *ICML*, pages 2554–2563, 2017.

[Na *et al.*, 2019] Donghyun Na, Haebeom Lee, Saehoon Kim, Minseop Park, Eunho Yang, and Sung Ju Hwang. Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks. *arXiv preprint arXiv:1905.12917*, 2019.

[Rajeswaran *et al.*, 2019] Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *NeurIPS*, pages 113–124, 2019.

[Rakelly *et al.*, 2019] Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *ICML*, pages 5331–5340, 2019.

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

[Ren *et al.*, 2018] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *ICLR*, 2018.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[Saenko *et al.*, 2010] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010.

[Schmidhuber, 1987] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4077–4087, 2017.

[Thrun and Pratt, 1998] Sebastian Thrun and Lorien Y. Pratt, editors. *Learning to Learn*. Springer, 1998.

[Tian *et al.*, 2019] Pinzhuo Tian, Zhangkai Wu, Lei Qi, Lei Wang, Yinghuan Shi, and Yang Gao. Differentiable meta-learning model for few-shot semantic segmentation. *arXiv preprint arXiv:1911.10371*, 2019.

[Utgoff, 1986] Paul E Utgoff. Shift of bias for inductive concept learning. *Machine learning: An artificial intelligence approach*, 1986.

[Vilalta and Drissi, 2002] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95, 2002.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016.

[Xie *et al.*, 2019] Yujia Xie, Haoming Jiang, Feng Liu, Tuo Zhao, and Hongyuan Zha. Meta learning with relational information for short sequences. In *NeurIPS*, pages 9901–9912, 2019.

[Zinkevich, 2003] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.