Dual-Level Curriculum Meta-Learning for Noisy Few-Shot Learning Tasks

Xiaofan Que, Qi Yu

Rochester Institute of Technology
{xq5054, qi.yu}@rit.edu

Abstract

Few-shot learning (FSL) is essential in many practical applications. However, the limited training examples make the models more vulnerable to label noise, which can lead to poor generalization capability. To address this critical challenge, we propose a curriculum meta-learning model that employs a novel dual-level class-example sampling strategy to create a robust curriculum for adaptive task distribution formulation and robust model training. The dual-level framework proposes a heuristic class sampling criterion that measures pairwise class boundary complexity to form a class curriculum; it uses effective example sampling through an under-trained proxy model to form an example curriculum. By utilizing both class-level and example-level information, our approach is more robust to handle limited training data and noisy labels that commonly occur in few-shot learning tasks. The model has efficient convergence behavior, which is verified through rigorous convergence analysis. Additionally, we establish a novel error bound through a hierarchical PAC-Bayesian analysis for curriculum meta-learning under noise. We conduct extensive experiments that demonstrate the effectiveness of our framework in outperforming existing noisy few-shot learning methods under various few-shot classification benchmarks. Our code is available at https://github.com/ritmininglab/DCML.

Introduction

Meta-learning is a technique that involves training a model on multiple tasks to generalize to new, unseen tasks using only a few training examples (Finn, Abbeel, and Levine 2017). While few-shot learning (FSL) methods have achieved success on benchmark tasks, it has been shown that these methods are highly susceptible to label noise, and even a single noisy data point can significantly impact the model's overall accuracy (Liang et al. 2022; Lu et al. 2020; Mazumder, Singh, and Namboodiri 2021). Traditional techniques for incorporating noise in supervised learning involving large amounts of training data (Ren et al. 2018; Shu et al. 2019; Han et al. 2018; Yu et al. 2019) may not be applicable in FSL settings due to two main reasons. First, FSL only has access to limited labeled data, which is insufficient to support traditional techniques that require large amounts of training data. Second, FSL methods typically rely on learning from the differences between tasks, instead of directly learning from individual data points. As a result, techniques that depend on individual data points may not be as effective in FSL settings.

Prior works on noisy few-shot learning (NFSL) attempt to address this issue using techniques such as feature aggregation (Liang et al. 2022), data augmentation (Mazumder, Singh, and Namboodiri 2021), and example re-weighting (Killamsetty et al. 2020). They have achieved moderate performance in noisy few-shot learning, but are fundamentally limited from two key aspects. (i) Granularity: these approaches only consider the example-level granularity, where the model learns to distinguish between individual examples within a few-shot task. However, episodic meta-learning methods learn to generalize at both the class-level and example-level, which can lead to more robust and accurate few-shot learning. (ii) Scope: these approaches either assume that the meta-training data to be clean and only consider noisy labels in the support set during meta-testing, or the other way around, which is a limited and less realistic scenario. In practice, both can significantly affect the few-shot learning performance. Therefore, robust FSL methods should consider the prevalence of noisy labels in both the meta-training and meta-test datasets.

To overcome these limitations, we propose a novel dual-level curriculum meta learning (*i.e.,* DCML) model that can generalize at both the class-level and example-level, while also considering noisy labels in both the meta-training and meta-test datasets. It performs dual-level sampling to dynamically form a task curriculum: (i) *Class-level sampling* continuously samples a subset of classes that are suitable for meta-training at the current stage and the selected classes forms a *subject* in the overall curriculum; (ii) *Example-level sampling* further chooses a subset of clean examples from currently chosen classes to construct the support and query sets. A central ingredient to formulate a robust curriculum is the criteria to accurately determine the easiness of a class/example over the course of meta-learning. However, directly using the average loss of examples may not be accurate (Kumar, Packer, and Koller 2010; Bengio et al. 2009). This is because the network's behavior changes over time, with the network initially being under-fitted and later potentially overfitting to noisy or hard examples (Huang et al. 2019; Toneva et al. 2019). Additionally, collecting historical statistics for each task in a meta-learning context is infeasible, making methods

that rely on running averages of loss nontrivial (Zhou, Wang, and Bilmes 2020b). To address this issue, we develop novel strategies for class-level curriculum learning (*i.e.,* C-CL) and example-level curriculum learning (*i.e.,* E-CL), respectively. At C-CL, we propose novel Class Pair (CP)-metrics based on the complexity of the decision boundary for class selection. Specifically, the CP-metrics measure the pairwise class similarity in terms of the incorrect class prediction probability. We use the CP-metrics to separate noisy class pairs (NCPs), similar class pairs (SCPs) and easy class pairs (ECPs) and build a meaningful class curriculum that helps to construct more diverse and informative tasks. At E-CL, we perform clean example selection by re-weighting each sample according to its loss information. To prevent the model from overfitting to the noisy labels and losing its power of identifying clean samples, we propose to use a proxy model, which is intentionally under-trained to select clean examples.

Using our innovative hierarchical PAC-Bayesian analysis for curriculum meta-learning, we have successfully derived an error bound for the proposed model that remains assured even when confronted with noisy conditions. The hierarchical analysis decomposes the entire problem into three tiers: task, subject, and curriculum, which allows us to construct the overall curriculum bound by combining the bounds from lower tiers. In addition, our theoretical contribution also makes two novel extensions to existing PAC-Bayes literature (Amit and Meir 2018; Rothfuss et al. 2021; Ding et al. 2021), including (i) deriving a bound on noisy meta-learning tasks and (ii) tackling the non *i.i.d.* task dependencies across different subjects. Our main contributions are summarized as follows:

- We propose a curriculum meta-learning model with a novel dual-level class-example sampling strategy that formulates a robust curriculum to adaptively adjust the task distribution for robust model training.
- We propose a heuristic class sampling criterion using novel CP-metrics and an effective example sampling strategy through an under-trained proxy model.
- We provide novel theoretical contributions that include a theoretical proof on the model convergence and hierarchical PAC-Bayesian analysis of error bounds for curriculum meta-learning under noise.

Experiments conducted over multiple synthetic and real-world datasets demonstrate the superior performance on few-shot learning from noisy data.

Related Work

Noisy few-shot learning (NFSL). Currently, only sparse effort has been indulged in NFSL. RNNP (Mazumder, Singh, and Namboodiri 2021) refines the prototypes using k-means clustering, which make the model learn to better distinguish between the different classes, even when the support set contains noisy or mislabeled data. RapNets (Lu et al. 2020) addresses representation or label noise by incorporating a BiLSTM-based attentive module. This module helps the model focus on the most informative features and examples in the support set while ignoring the noisy or irrelevant ones. TraNSF (Liang et al. 2022) refines the class prototypes used by ProtoNet by aggregating the features of the support examples, and uses a Transformer-based model that employs an attention mechanism to weigh the relevance of support samples based on their correctness. This allows the model to better filter out noisy or mislabeled data during training. Other works, such RW-MAML (Killamsetty et al. 2020), AQ (Goldblum, Fowl, and Goldstein 2020), and DFSL (Li et al. 2022) tackle the out-of-distribution tasks or adversarial attacks in few-shot learning.

Curriculum meta-learning (CML). The combination of meta-learning and curriculum learning has received increasing attentions (Zhang et al. 2021; Liu and Fu 2021; Shevchuk 2019; Cioba et al. 2022). For example, (Zhang et al. 2022a) uses self-paced learning to decide the hardness of tasks in a given batch adaptively according to the learned model. The task batch is first randomly sampled from the base classes then ordered by the model, therefore, the hardness of each task is not global and may not be optimal. (Agrawal, Squire et al. 2021) defines the curriculum schedule for meta-learning by increasing the support size for each task at the beginning then reducing it to the pre-defined shot-size. CML has also been applied to different domains such as recommender systems (Chen et al. 2021), NLP (Wu et al. 2021; Zhan et al. 2021), medical data analysis (Li and Lovell 2022; Zhang et al. 2022b), long-tailed recognition (Sinha and Ohashi 2023) and reinforcement learning (Mehta et al. 2020; Portelas et al. 2020), to name a few. Specifically, (Chen et al. 2021) considers the hardness for both cities and users and samples the hard ones to update the meta-learner for better convergence. Our proposed model is designed specifically for tackling a noisy task environment with a unique dual-level sampling strategy. It also offers a rigorous analysis on both model convergence and error bound, which are missing from most existing works.

Preliminaries

Meta-learning. Given the data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the meta-model θ is learned using batches of episodes (*i.e.,* tasks) $\mathcal{T}$ sampled from $\mathcal{D}$ in an episodic way. The episodic sampling process of a N-way K-shot classification task $\mathcal{T}_i$ includes two steps: first randomly samples N classes from the base class set $\mathcal{C}_{\mathcal{B}}$; then for each class randomly sample K images as the support set $\mathcal{S}_i^{sup} = \{(\mathbf{x}_i^j, y_i^j)\}_{j=1}^{NK}$, and another Q images as the query set $\mathcal{S}_i^{que} = \{(\mathbf{x}_i^j, y_i^j)\}_{j=1}^{NQ}$.

Curriculum learning. In curriculum (or self-paced) learning, a weight vector $\mathbf{v}$ is introduced and jointly optimized with the model parameter $\mathbf{w}$ in an alternative way. The objective is formulated as follows:

$$\min_{\mathbf{w}, \mathbf{v} \in [0,1]^n} \sum_{i=1}^{n} v_i \mathcal{L}(y_i, f_{\mathbf{w}}(\mathbf{x}_i)) - \lambda R(\mathbf{v}), \tag{1}$$

where $\mathcal{L}(y_i, f_{\mathbf{w}}(\mathbf{x}_i))$ is the training loss between the predicted label $f_{\mathbf{w}}(\mathbf{x}_i)$ and the ground truth y_i and $R(\mathbf{v})$ is a regularizer to achieve the desired curriculum design. Intuitively, v_i indicates the easiness of sample $\mathbf{x}_i$ for the current model and the goal is to gradually train the model by following easy to difficult examples (to mimic how humans learn). There are different forms of the regularizer. For hard weighting, it

leverages a negative l_1-norm: $R(\mathbf{v}) = \sum_{i=1}^{n} |v_i|$. For *soft weighting*, linear, logarithmic, and mixture forms are popular choices (Jiang et al. 2014a). The hyperparameter λ is a general threshold for selecting examples.

Methodology

Overview. DCML contains two levels: C-CL and E-CL. In C-CL, a class curriculum is progressively built with a series of subjects $\tau_1, ..., \tau_C$ using the CP-metrics μ and σ. Based on the easy-to-complex curriculum learning intuition, ECPs are first fed into the target model using a small initial μ_{th}^0. Then the values of μ_{th}, σ_{th} are gradually increased so that the similar but difficult SCPs are included to provide fine-grained classification tasks for training. The NCPs are put in the end of the curriculum and trained insufficiently to avoid overfitting. Some are removed depending on the corruption level. In E-CL, a proxy is used to filter out noise in the example-level by sampling small loss examples of each task. This is achieved by assigning proper weights to examples sampled within each task. The detailed training process is summarized in Algorithm 1 of the Appendix (Que and Yu 2024).

Proxy models. Training proxy models is widely utilized in active learning and core-set selection (Coleman et al. 2019). Usually, the proxies are designed similar to the target model but with fewer hidden layers and trained with fewer epochs for efficiency consideration, which may lead to inevitable performance compromise. In our framework, instead of training extra proxies, we repeatedly use the model itself trained in different time steps as proxies for class selection (*i.e.,* C-CL) and use the model trained in the early period as the proxy for example selection (*i.e.,* E-CL). It has the following merits: (i) the proxies are kept fixed during training so only a small amount memory is required for storage and there is no dynamic memory concern; (ii) the proxies are exactly the same as the target model so the performance won't be compromised; (iii) the proxies for sample selection is trained with fewer epochs to be under-fitted for identifying the clean samples from the noise.

CP-metrics for Class-Level Sampling

Given an image $\mathbf{x}$, the embedding network f_{θ^t} with a classifier head g_t at time step t outputs the the pre-softmax logit $p_t(\mathbf{x}) = g_t(f_{\theta^t}(\mathbf{x}))$, which is a $N \times 1$ vector: $p_t(\mathbf{x}) = [p_t^{cp_{j1}}(\mathbf{x})), ..., p_t^{cp_{jN}}(\mathbf{x}))]$, where N is the number of classes, and the element $p_t^{cp_{jk}}(\mathbf{x})$ indicates the probability of assigning the example $\mathbf{x}$ with label k to class j. In a N-way K-shot classification task, k ranges from 1 to N. We eliminate the matched class logit $p_t^{cp_{jj}}$, and collect the mismatched class logits where $j \neq k$ and called them the incorrect prediction probabilities of example $\mathbf{x}$ at time step t.

According to the inter-class variance, it's reasonable to assume that NCPs and SCPs will in expectation have a larger incorrect prediction probability than clean and easily classified ones. In addition, NCPs should not consistently recur due to their randomness nature while inherently SCPs may lead to more consistent incorrect predictions by DNNs. Intuitively, the variance of the CP-statistics should be higher for

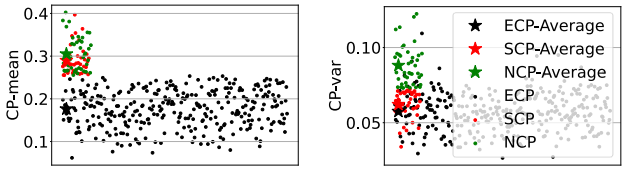

Figure 1: Illustration of the class pairs separated by CP-metrics. The points denoting different pairs are scattered in the figure with different colors. Averaged values of the class pairs in different groups are shown in lines with different colors. Left: CP-mean separates ECPs from NCPs and SCPs; Right: CP-var separates SCPs from NCPs.

the NCPs than the similar but useful ones. Our later quantitative and qualitative results confirm this assumption. Based on this intuition, we incorporate the historical statistics to formulate the metrics *CP-mean* $\mu^{\text{cp}_{jk}}$ and *CP-var* $\sigma^{\text{cp}_{jk}}$ to separate easy, similar and noisy class pair cp_{jk}, which are defined as follows:

$$\mu^{\text{cp}_{jk}}(\mathbf{x}) = \frac{1}{TN_j} \sum_{t=1}^{T} \sum_{i}^{N_j} p_t^{\text{cp}_{jk}}(\mathbf{x}_i), \tag{2}$$

$$\sigma^{\text{cp}_{jk}}(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} \left[\frac{\sum_i^{N_j} (p_t^{\text{cp}_{jk}}(\mathbf{x}_i) - \mu^{\text{cp}_{jk}}(\mathbf{x}))^2}{N_j} \right]^{\frac{1}{2}}, \tag{3}$$

where T is the total number of current model checkpoints that keeps increasing as the training proceeds and N_j is the total number of sampled instances that belong to class j. The value of μ^{cp} measures the misclassified probability of a class pair, which is used to separate the ECPs from the SCPs and NCPs. On the other hand, σ^{cp} measures the variation of a class being misclassified into another class that can help to discriminate SCPs from NCPs. Through the predefined thresholds: μ_{th}, σ_{th}, we can separate the class pairs into easy, similar and noisy groups given below,

$$\begin{aligned} C_{easy} &= \{cp_{jk} : \mu^{cp_{jk}}(\mathbf{x}) < \mu_{th}\}, \\ C_{similar} &= \{cp_{jk} : \mu^{cp_{jk}}(\mathbf{x}) > \mu_{th} \ \& \ \sigma^{cp_{jk}}(\mathbf{x}) < \sigma_{th}\}, \\ C_{noisy} &= \{cp_{jk} : \mu^{cp_{jk}}(\mathbf{x}) > \mu_{th}) \ \& \ \sigma^{cp_{jk}}(\mathbf{x}) > \sigma_{th}\}. \end{aligned} \tag{4}$$

The choice of these thresholds follows some intuitive guidelines to ensure good performance without a costly grid search. Like other curriculum learning models (Bengio et al. 2009; Kumar, Packer, and Koller 2010), the setting of thresholds usually depends on the noise ratio, which can be determined by collecting a subset of *i.i.d.* samples, verifying their labels, and identifying the wrong ones. After the noise ratio is estimated, we can directly set the thresholds that aim to eliminate the same percentage of noisy labels. In Fig.1, we show that different class pairs are clearly separated using CP-metrics with examples from Omniglot. In the left plot, the average of CP-mean μ^{cp} of ECPs are smaller than the NCPs and SCPs by a large margin; in the right plot, the SCPs and NCPs are further discriminated by the separation of their CP-var σ^{cp}. We observe more black dots in the figure since

the number of easy-pairs are way more than the noisy and similar ones in practice.

Example-Level Curriculum Learning

Given a meta-model f_θ parameterized by θ, for a new task $\mathcal{T}_j$, the meta-model is adapted to the task-specific model θ_j using one gradient update with support set $\mathcal{D}_j^{\text{sup}}$:

$$\theta_j = \theta - \alpha \nabla_\theta \mathbf{v}_j^{\text{sup}} \mathcal{L}_{\mathcal{T}_j}(\theta, \mathcal{D}_j^{\text{sup}}), \tag{5}$$

where $\mathbf{v}_j^{\text{sup}}$ is the support set weight vector for task $\mathcal{T}_j$, with its k-th element $v_{j,k}^{\text{sup}}$ computed as:

$$v_{j,k}^{\text{sup}} = \mathbb{1}(\ell(f_{\theta^p}(\mathbf{x}_{j,k}), y_{j,k}) < \lambda), \tag{6}$$

where $\mathbb{1}(\cdot)$ is an indicator function, ℓ is cross-entropy loss, θ^p is the proxy model parameter, λ is a predefinded hyperparameter, $\mathbf{v}_j^{\text{sup}} \mathcal{L}_{\mathcal{T}_j}(\theta, \mathcal{D}_j^{\text{sup}}) = \frac{1}{B} \sum_k v_{j,k}^{\text{sup}} \ell(f_\theta(\mathbf{x}_j^k), y_j^k)$ with B denoting the number of non-zero weights for the support set images, and $(\mathbf{x}_j^k, y_j^k)$ is a image-label pair in support set $\mathcal{D}_j^{\text{sup}}$ (*i.e.,* $\mathcal{S}_j$). As in (Finn, Abbeel, and Levine 2017), the optimization of the meta-model across tasks is performed via stochastic gradient descent (SGD), such that the model parameters θ are updated as follows:

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_j \sim \tau_i} \mathbf{v}_j^{\text{que}} \mathcal{L}_{\mathcal{T}_j}(\theta_j, \mathcal{D}_j^{\text{que}}), \tag{7}$$

where $\mathbf{v}_j^{\text{que}}$ is the query set weight vectors designed similar to that of the support set except that they use a task-specific model to compute the loss: $v_{j,k}^{\text{que}} = \mathbb{1}(\ell(f_{\theta_j}(\mathbf{x}_{j,k}), y_{j,k}) < \lambda)$, where τ_i is the current subjects that the classes are sampled from and α, β are the corresponding step sizes.

Hierarchical PAC-Bayesian Analysis

In the context of meta-learning, PAC-Bayesian theory is extensively studied to provide guarantees for generalization errors (Ding et al. 2021; Farid and Majumdar 2021; Liu et al. 2021).

PAC-Bayesian for supervised tasks. Let $\mathcal{Z} : \mathcal{X} \times \mathcal{Y}$ and $\mathcal{H}$ denote an input-output space and a hypothesis space, respectively. For a supervised task, a hypothesis h is sampled from $\mathcal{H}$ to make predictions on input $z \sim \mathcal{Z}$, whose performance is measured by a loss function $\ell(h, z) : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$. Given a set of m observations $S = \{z\}_{i=1}^m \sim \mathcal{D}^m$, where $\mathcal{D}$ is a data distribution over $\mathcal{Z}$. The superscript m is used to denote that m examples are sampled *i.i.d.* from $\mathcal{D}$. The goal is to minimize the expected error $er(h, \mathcal{D}) = \mathbb{E}_{z\sim\mathcal{D}} \ell(h, z)$. Since $\mathcal{D}$ is usually unknown, the empirical error $\hat{er}(h, S) = \frac{1}{m} \sum_{i=1}^m \ell(h, z_i)$ is utilized in practice. PAC-Bayesian setting assumes that there exists a prior distribution of the hypothesis space in the form of $P(h) \sim \mathcal{M}(\mathcal{H})$, where $\mathcal{M}(\mathcal{H})$ is the set of all possible probability measures in $\mathcal{H}$. Upon observing the training dataset S, the base learner $Q(S, P)$ updates the prior into a posterior $Q(h)$. The base learner $Q(S, P)$ is a mapping: $Q : \mathcal{Z}^m \times \mathcal{M} \to \mathcal{M}$. Formally, for any probability distribution Q over the hypothesis set, the corresponding Gibbs predictor for every point $z \in \mathcal{Z}$ randomly samples $h \sim Q$ and returns $h(z)$. The expected loss of such Gibbs predictor on a task corresponding to a data distribution $\mathcal{D}$ is given by: $er(Q, \mathcal{D}) = \mathbb{E}_{h\sim Q} er(h, \mathcal{D})$. It's empirical counterpart is defined as $\hat{er}(Q, S) = \mathbb{E}_{h\sim Q} \hat{er}(h, S)$.

PAC-Bayesian for meta-learning. During meta-training, a series of tasks are sampled from the task distribution τ. Upon observing datasets $S_1, ..., S_n$ from the tasks, the meta-learner presumes a hyper-prior $\mathcal{P}(P) \sim \mathcal{M}(\mathcal{M}(\mathcal{H}))$ as a distribution over priors P, and updates it to a hyper-posterior $\mathcal{Q}(P)$. The performance of the hyper-posterior is measured by the so-called transfer error $er(\mathcal{Q}, \tau) = \mathbb{E}_{P\sim\mathcal{Q}} \mathbb{E}_{\mathcal{D}\sim\tau} \mathbb{E}_{S\sim\mathcal{D}} er(Q, \mathcal{D})$. Its empirical counterpart is defined as $\hat{er}(\mathcal{Q}, S_1, ..., S_n) = \mathbb{E}_{P\sim\mathcal{Q}} \frac{1}{n} \sum_{i=1}^n \hat{er}(Q, S_i)$.

Hierarchical PAC-Bayesian analysis for DCML. We propose to conduct PAC-Bayesian analysis for curriculum meta-learning under noise through a three-tier hierarchy: task, subject and curriculum. Specifically, a curriculum is divided into C subjects, the task distribution of each subject is denoted as τ_i, where a batch of tasks S_i are sampled.

PAC-Bayesian bound for a singe task in meta-learning. For each task $\mathcal{T}_j^i$ in the i-th subject, the base learner Q_j^i is updated with a prior P_j^i, which is sampled from a hyper-posterior $\mathcal{Q}(P(h))$ (*i.e.,* the meta-model), and dataset $S_j^i = \{z\}_{k=1}^m$ sampled *i.i.d.* from the task distribution τ_i with data distribution $\mathcal{D}_i$. The expected and empirical errors are defined as $er(Q, \mathcal{D}_i) = \mathbb{E}_{h\sim Q} \mathbb{E}_{z\sim\mathcal{D}} \ell(h, z)$, $\hat{er}(Q, S_j^i) = \mathbb{E}_{h\sim Q} \frac{1}{m} \sum_{k=1}^m \ell(h, z_k)$. We have the following theorem:

Theorem 1 (Noisy meta-learning task bound)**.** *Let Q be a base learner and $\mathcal{P}$ be some pre-defined hyper-prior distribution over prior P. Then for any $\delta \in (0, 1]$, the following inequality holds uniformly for all hyper-posterior distribution $\mathcal{Q}$ with probability at least $1 - \delta$,*

$$\mathbb{E}_{P\sim\mathcal{Q}} er(Q, \mathcal{D}) \le \mathbb{E}_{P\sim\mathcal{Q}} \hat{er}(Q, S) + \sqrt{\frac{D(\mathcal{Q}||\mathcal{P}) + \mathbb{E}_{P\sim\mathcal{Q}} D(Q(S, P)||P) + \log \frac{m(1-\hat{r})}{\delta}}{2(m(1-\hat{r}) - 1)}}, \tag{8}$$

where $\hat{r}$ is the empirical noise rate.

Proof Sketch. The proof consists of three main steps: 1) we apply the change of measure of KL divergence between measurement spaces to distributions; 2) we apply Theorem 6 (as shown in the Appendix) to bound the expected task error with empirical task error; 3) we reduce the number of training samples in each task according to the noise ratio to finalize the proof.

According to Eqn. (6), the expected and empirical risks becomes $er(Q, \mathcal{D}_j^i) = \mathbb{E}_{h\sim Q} \mathbb{E}_{z\sim\mathcal{D}} v_z \ell(h, z)$, $\hat{er}(Q, S_j^i) = \mathbb{E}_{h\sim Q} \frac{1}{m} \sum_{k=1}^m v_{z_k} \ell(h, z_k)$, where $v_{z_k} \in \{0, 1\}$ is the weight index of the k-th example in task $\mathcal{T}_j^i$. Assuming the true noise ratio of each task is r, we can rewrite the loss term as $\mathcal{L}_{\mathcal{T}_j^i}(\theta_j^i, \mathcal{D}_i^{\text{que}}) = (1 - r) \sum_{(x_{j,k}^i, y_{j,k}^i) \in \mathcal{D}_{i,j}^{\text{que}}} \ell(\theta_j^i, (x_{j,k}^i, y_{j,k}^i))$. We then have $m(1 - \hat{r})$ clean examples remained in each task empirically and the proof of the theorem can follow.

Theorem 1 shows that the PAC-Bayesian bound is compromised due to the existence of the noisy labels. In practice, the

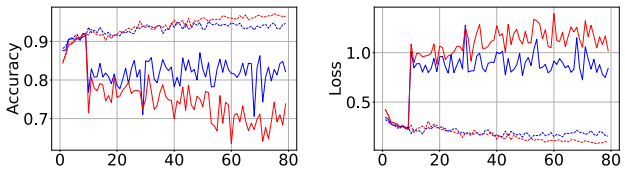

Figure 2: Example of unstable performance under noise. Left: Training and testing accuracy; Right: Training and testing loss.

noise ratio r usually remains unknown. Existing data cleansing techniques usually end up with accidentally removing some clean but difficult examples, which results in a larger $\hat{r}$ and hence enlarges the bound. Our dual-level design aims to separate the similar from the noisy ones and eliminate the least useful examples to keep the empirical noise ratio $\hat{r}$ close to the true one. This effectively leads to a tighter bound under a noisy data distribution.

PAC-Bayesian bound within a subject. The dual-level curriculum meta-learning forms a series of subjects dynamically. For a batch of tasks sampled from a fixed subject with task environment τ_i, the expected error of a fixed prior P_i and its empirical counterpart are defined as $er_i(P,\tau_i) = \mathbb{E}_{\mathcal{D}^m\sim\tau_i}\mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{h\sim Q}\mathbb{E}_{z\sim\mathcal{D}}\ell(h,z)$, $\hat{er}_i(P,S_1,\ldots,S_{n_i}) = \mathbb{E}_{h\sim Q}\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{1}{m}\sum_{k=1}^{m}\ell(h,z_k^j)$, where n_i is the number of tasks sampled from the i-th subject. We derive the PAC-Bayesian bound for a batch of tasks sampled from a subject:

Theorem 2 (Curriculum meta-learning subject bound)**.** *Let Q be a base learner and $\mathcal{P}$ be some pre-defined hyper-prior distribution over prior P. Then for any $\delta \in (0,1]$, the following inequality holds uniformly for all hyper-posterior distribution $\mathcal{Q}$ with probability at least $1-\delta$,*

$$\mathbb{E}_{P\sim\mathcal{Q}}er_i(P,\tau_i) \le \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{E}_{P\sim\mathcal{Q}}\hat{er}_i(P,S_1,\ldots,S_{n_i}) + \sqrt{\frac{1}{2(n_i-1)}(D(\mathcal{Q}||\mathcal{P}) + \log\frac{2n_i}{\delta})} + \frac{1}{n_i}\sum_{j=1}^{n_i}\sqrt{\frac{D(\mathcal{Q}||\mathcal{P}) + \mathbb{E}_{P\sim\mathcal{Q}}D(Q(S_i,P)||P) + \log\frac{\kappa}{\delta}}{2(m(1-\hat{r})-1)}}. \quad (9)$$

where $\kappa = 2n_i m(1-\hat{r})$.

PAC-Bayesian bound for curriculum meta-learning. While tasks sampled from the same subjects can still be considered as *i.i.d.*, it is no longer the case when the entire curriculum is considered. This is because classes are organized into different subjects based on their difficulty levels and tasks from the adjacent subjects have similar difficulty levels (and hence correlated). Non *i.i.d.* tasks have been investigated under the life-long learning setting (Pentina and Lampert 2015) and a bound can be derived with the key assumption that the model should perform stable across different tasks. However, no concrete strategies have been developed to make sure the assumption holds. However, under the noisy label setting, this assumption might collapse. Fig. 2 shows the training and testing processes using both accuracy and loss of a meta-learning model trained with a noisy Omniglot dataset (the solid lines). We can see that the model presents a significant oscillating performance under the influence of noisy labels, making the stability assumption not hold anymore. However, by gradually exposing the model to the noisy environment in an easy-to-complex order (*e.g.,* through our design), the performance is much more stable (the dash lines). Therefore, we are assured that the expected performance of the meta-learner does not change over subjects as the task distributions gradually switch from easy to complex. More formally, for each subject τ_i, the quality of prior P measured by the expected loss when using it to learn new tasks, as defined by $er_i(P,\tau_i)$ won't change over time: $\mathbb{E}_{\mathbb{E}_1,\ldots,\mathbb{E}_C}[er_i(P,\tau_i)] = er(P,\tau)$, where $\mathbb{E}_i = (t_j^i,\tau_i,S_j^i)$.

For any hypothesis sampled from the adapted posterior Q: $h \sim Q(S,P)$, we have the empirical loss $\hat{er}_i(Q(S_i,P),S_i) = \mathbb{E}_{h\sim Q(S_i,P)}\frac{1}{m}\sum_{j=1}^{m}\ell(h,z_j)$ over the observed training data S_i. The expected multi-task error $\tilde{er}_i(Q(S_i,P),\mathcal{D}_i) = \mathbb{E}_{h\sim Q(S,P)}\mathbb{E}_{z\sim\mathcal{D}_i}\ell(h,z)$ is evaluated on the corresponding unknown task distribution $\mathcal{D}_i$, where the training data are sampled from $S_i \sim \mathcal{D}_i$. The performance of the hyper-posterior $\mathcal{Q}$ is measured by expected loss of learning new tasks using priors drawn from $\mathcal{Q}$, which is defined as $er(\mathcal{Q},\tau) = \mathbb{E}_{P\sim\mathcal{Q}}\sum_{i=1}^{C}er_i(P_i,\tau_i)$ and its empirical counterpart $\hat{er}(\mathcal{Q},S_1,...,S_C) = \mathbb{E}_{P\sim\mathcal{Q}}\frac{1}{C}\sum_{i=1}^{C}\hat{er}_i(Q(S_i,P),S_i)$. Below, we present the bound for curriculum meta-learning with an easy-to-complex order.

Theorem 3 (Curriculum meta-leaning PAC-Bayes bound)**.** *Let Q be a base learner and $\mathcal{P}$ be some pre-defined hyper-prior distribution over prior P. Then for any $\delta \in (0,1]$, the following inequality holds uniformly for all hyper-posterior distribution $\mathcal{Q}$ with probability at least $1-\delta$,*

$$er(\mathcal{Q}) \le \hat{er}(\mathcal{Q}) + \frac{1}{\sqrt{C}}(D(\mathcal{Q}||\mathcal{P}) + \frac{1}{2} - \log\frac{\delta}{2}) + \frac{1}{C}\sum_{i=1}^{C}\sqrt{\frac{1}{2(n_i-1)}(D(\mathcal{Q}||\mathcal{P}) + \log\frac{4Cn_i}{\delta})} + \frac{1}{C}\sum_{i=1}^{C}\frac{1}{n_i}\sum_{j=1}^{n_i}\sqrt{\frac{D(\mathcal{Q}||\mathcal{P}) + \mathbb{E}_{P\sim\mathcal{Q}}D(Q(S_i,P)||P) + \log\frac{\hat{n}}{\delta}}{2(m(1-\hat{r})-1)}},$$

where $\hat{n} = 4Cn_i m(1-\hat{r})$.

Proof Sketch. The proof consists of three main steps: 1) we apply the Donsker-Varadhan's variational formula (Seldin et al. 2012) to bound the expected risk of multiple subsets with their empirical counterparts; 2) we utilize the result of subset bound to obtain intermediate result of the curriculum meta-learning; 3) we use the union bound argument (Amit and Meir 2018) to finalize the proof.

Experiments

In this section, datasets, the types of synthetic label noise, training process, and hyperparameter settings are first explained. Then the comparison results on different datasets

and various ablation studies are designed evaluate the effectiveness of DCML.

Datasets. We evaluate the effectiveness of the proposed dual-level curriculum meta-learning framework (*i.e.,* DCML) using three benchmark datasets: miniImageNet (Ravi and Larochelle 2017), FC100 (Oreshkin, López, and Lacoste 2018), Omniglot (Lake et al. 2011) for few-shot learning along with three real-world noisy datasets: miniWV (Li et al. 2017), Food101 (Bossard, Guillaumin, and Van Gool 2014) and CIFAR-100N (Wei et al. 2022). The details of the datasets are given in the Appendix.

Baselines. We compare our method with following methods: we improve MAML (Finn, Abbeel, and Levine 2017) with existing robust deep learning methods that achieve competitive performance on noisy data, including MAML+spld (Jiang et al. 2014b), MAML+focal (Lin et al. 2017), MAML+dih (Zhou, Wang, and Bilmes 2020b), curriculum meta-learning method for few-shot classification, Curriculum MAML (CMAML) (Agrawal, Squire et al. 2021), state of the art meta-learning method CT (Luo, Xu, and Xu 2022), and robust few-shot learning methods RNNP (Mazumder, Singh, and Namboodiri 2021), RapNets (Lu et al. 2020) TraNFS (Liang et al. 2022) and IDEAL (An et al. 2023). We discuss the details in the Appendix. All experiments are conducted on an NVIDIA A100 GPU with three runs (RIT Research Computing 2019).

Synthetic label noise. We study four settings of synthetic label noises: (i) Symmetric: it is also known as UniformFlip (Ren et al. 2018), where all label classes can uniformly flip to any other label classes; (ii) Asymmetric: the label classes can only flip to other similar label classes; (iii) BackgroundFlip (Ren et al. 2018): all label classes can flip to a single background class; (iv) Mixture: it includes symmetric, asymmetric and BackgroundFlip label noise, by mimicking the real-world noise. For detailed noise and training hyperparameter settings, please refer to the Appendix.

Performance Comparison

Tab. 1 and Tab. 3 present 5-way 1-shot test accuracy with synthetic label noises and real-world label noises, respectively. The corresponding standard deviations are provided in Tab. 6 and Tab. 8 in the Appendix. In the Appendix, we also present the results of mixture noise (*i.e.,* Tab. 11), and results of 5-way 5-shot tasks (*i.e.,* Tab. 13 and Tab. 14). All results are the average of 3 runs. Regardless of the various strong baselines, our method ranks at the top for both synthetic label noises and real-world label noises, showing that our method is less affected by different noise types. For different types of noises, we have applied our CP-metrics to arrange task distributions differently. For symmetric noise, the noise labels are randomly flipped to different classes. Therefore, we propose to use CP-metrics to classify the noisy class pairs and the similar class pairs. During training, the noisy ones are discarded and the similar ones are included during the later stage of training as difficult tasks. For asymmetric ones, the similar class pairs are the ones containing higher ratios of label noise, therefore, they are discretely eliminated during

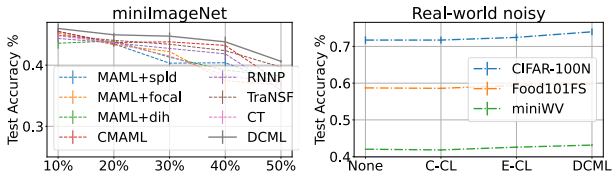

Figure 3: Left: Test accuracy with different ratios of symmetric label noise. Right: Test accuracy with different DCML components.

early training and insufficiently trained during the later training stage so that the model won't overfit the noisy labels. For the background noise, the threshold of the eliminated class is fixed as one during the entire training so that the model won't be affected by the background noise. For the mixture noise, combined strategies are used to divide the training into different stages: during the early training stage, the easy classes are utilized; during the later training stage, the similar classes are incorporated and trained insufficiently; lastly, a portion of noisy classes such as background noisy classes are completely banned during entire training. Considering the fact that the mixture noise is created by mimicking the real-world noise, the same strategy dealing with the mixture is applied the training of the real-world noisy datasets. By using different strategies, we make sure the data is sufficiently utilized and the impact of noise is minimized. From the above results, we observe that our method ranks the top across different datasets with different noisy types, thanks to the flexible training strategies design.

Ablation Study

In this section, we study the noisy few-shot task evaluation setting, and investigate the impacts of different components via ablation studies, including noise ratio, C-CL, E-CL. More ablations are presented in the Appendix.

Noisy few-shot tasks evaluation. To verify the effectiveness of the label noise to each few-shot task, we manually add 40% of synthetic symmetric label noise to the support set of each task during evaluation. In Tab. 2 and Tab. 4, we observe that the performance drops drastically up to 10% (compared to Tab. 1 and Tab. 3) when the symmetric noise is applied to each task during evaluation, indicating that FSL is extraordinary vulnerable to noise. Our method manages to rank the top even in such crucial situation, showing its superior robustness. The corresponding standard deviation and results of 5-way 5-shot cases are shown in the Appendix.

Impact of noise ratio. In this study, we conduct experiments on miniImageNet with 10%, 20%, 30%, 40%, 50% symmetric label noise, and 30%, 50% mixture of symmetric and asymmetric label noise. As shown in Fig. 3 (left), with the increase of noise ratio, the performance of different baselines drop rapidly while our method drops the least. Specifically, with 10% of label noise, our method only slightly outperforms other baselines. However, the performance margin of our method and the others increase as the noise ratio

Method	miniImageNet	FC100	Omniglot	miniImageNet	FC100	Omniglot	miniImageNet	FC100	Omniglot
Noise Type	Asymetric			Symmetric			Background		
MAML	0.4612	0.3636	0.9408	0.3754	0.3428	0.9166	0.4676	0.3433	0.9585
MAML+spld	0.4605	0.3563	0.9379	0.3602	0.3236	0.9037	0.4478	0.3503	0.9571
MAML+focal	0.4512	0.3502	0.9409	0.3852	0.3422	0.9041	0.4562	0.3545	0.9592
MAML+dih	0.4617	0.3577	0.9381	0.3749	0.3418	0.9046	0.4594	0.3477	0.9556
CMAML	0.4533	0.3573	0.9206	0.4134	0.3389	0.8994	0.4618	0.3463	0.9402
RNNP	0.4622	0.3973	0.9506	0.4511	0.3889	0.9394	0.4650	0.3963	0.9602
TraNFS	0.5100	0.4152	0.9643	0.4080	0.4119	0.9347	0.5100	0.4129	0.9787
CT	0.4642	0.3589	0.9429	0.3711	0.3418	0.9140	0.4650	0.3437	0.9571
IDEAL	0.5201	0.4410	0.9566	0.4440	0.4606	0.9200	0.4800	0.4278	0.9766
DCML	**0.5315**	**0.4671**	**0.9677**	**0.4627**	**0.4822**	**0.9543**	**0.5193**	**0.4448**	**0.9843**

Table 1: 5-way 1-shot on different types of noises

Method	miniImageNet	FC100	Omniglot	miniImageNet	FC100	Omniglot	miniImageNet	FC100	Omniglot
Noise Type	Asymetric			Symmetric			Background		
MAML	0.3176	0.3045	0.6712	0.3174	0.2980	0.6557	0.3530	0.2883	0.6660
MAML+spld	0.3462	0.2967	0.6685	0.2929	0.2776	0.6411	0.3414	0.2895	0.6864
MAML+focal	0.3493	0.3003	0.6710	0.3028	0.2861	0.6400	0.3524	0.2935	0.6834
MAML+dih	0.3530	0.2975	0.6681	0.3035	0.2863	0.6425	0.3451	0.2819	0.6810
CMAML	0.3427	0.2917	0.6580	0.3251	0.2917	0.6457	0.3451	0.2819	0.6810
RNNP	0.3487	0.2863	0.6780	0.3151	0.2969	0.6572	0.3631	0.2919	0.6805
TraNFS	0.3562	0.3011	0.6702	0.3103	0.2932	0.6651	0.3667	0.2906	0.6823
CT	0.3471	0.2963	0.6680	0.3005	0.2908	0.6431	0.3544	0.2844	0.6805
IDEAL	0.3488	0.3009	0.6822	0.3200	0.3028	0.6691	0.3500	0.3001	0.6650
DCML	**0.3573**	**0.3144**	**0.6833**	**0.3332**	**0.3142**	**0.6740**	**0.3558**	**0.3149**	**0.6716**

Table 2: 5-way 1-shot test accuracy on few-shot classification task with symmetric noise on support set during meta-test

Datasets	CIFAR-100N	Food101FS	miniWV
MAML	0.5446	0.3927	0.3274
MAML+spld	0.5266	0.3947	0.3221
MAML+focal	0.5202	0.399	0.3198
MAML+dih	0.5377	0.3949	0.3282
CMAML	0.5528	0.4144	0.3382
RNNP	0.5928	0.4844	0.3815
TraNFS	0.5853	0.3957	0.3284
CT	0.5460	0.3957	0.3263
IDEAL	0.5890	0.4511	0.3726
DCML	**0.6059**	**0.5165**	**0.3986**

Table 3: 5-way 1-shot accuracy on real-world noisy datasets

Datasets	CIFAR-100N	Food101FS	miniWV
MAML	0.3908	0.3233	0.2760
MAML+spld	0.3901	0.3160	0.2767
MAML+focal	0.3914	0.3144	0.2744
MAML+dih	0.3913	0.3127	0.2751
CMAML	0.3977	0.3116	0.2772
RNNP	0.3775	0.3267	0.2760
TraNFS	0.4146	0.3285	0.2800
CT	0.4015	0.3249	0.2848
IDEAL	0.4208	0.3188	0.2789
DCML	**0.4239**	**0.3343**	**0.2942**

Table 4: 5-way 1-shot few-shot classification test accuracy with symmetric noise on support set during meta-test

increase. Similarly, as shown in Tab. 10, when the noise ratio increases from 30% to 50%, the test accuracy of our method only drops 0.04%, while some others drops nearly 10%.

Impact of C-CL and E-CL. The right plot of Fig. 3 shows the 5-way 5-shot result of DCML with only one level of curriculum applied (*i.e.,* C-CL or E-CL), and no curriculum applied (*i.e.,* None) on the three real-world noisy datasets CIFAR-100N, Food101 and miniWV, and compared to both level of curriculum applied (*i.e.,* DCML). We observe that applying a singe level of curriculum learning (C-CL or E-CL) performs inferior to the dual-level design but superior to none curriculum learning applied, indicating the combination of C-CL and E-CL benefits model training the most. Without E-CL, C-CL is misled and a sub-optimal curriculum is formed which deteriorates the performance due to the noise in each task. When E-CL is applied, the noise in each task is removed so that C-CL formulates an optimal curriculum using clean tasks, hence DCML gives the highest performance.

Conclusion

In this paper, we develop a dual-level curriculum meta-learning framework for robust few-shot learning. In the proposed framework, class-level sampling formulates an easy-to-complex curriculum by identifying difficult classes from the noisy ones using our proposed CP-metrics whereas a proxy model is leveraged at the example level to choose clean examples. We provide convergence analysis and a novel hierarchical PAC-Bayes analysis for the dual-level framework under the noisy setting. The performance is validated on benchmark and real-world datasets with diverse label noises.

Acknowledgments

This research was partially supported by NSF IIS award IIS1814450 and ONR award N00014-18-1-2875. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the official views of any funding agency. We thank the anonymous reviewers for their constructive comments.

References

Agrawal, P.; Squire, O.; et al. 2021. Curriculum Meta-Learning for Few-shot Classification. In *Proc. of NeurIPS*.

Amit, R.; and Meir, R. 2018. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *Proc. of ICML*.

An, Y.; Xue, H.; Zhao, X.; and Wang, J. 2023. From Instance to Metric Calibration: A Unified Framework for Open-World Few-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proc. of ICML*.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Proc. of ECCV*.

Chen, Y.; Wang, X.; Fan, M.; Huang, J.; Yang, S.; and Zhu, W. 2021. Curriculum meta-learning for next POI recommendation. In *Proc. of KDD*.

Cioba, A.; Bromberg, M.; Wang, Q.; Niyogi, R.; Batzolis, G.; Garcia, J.; Shiu, D.-s.; and Bernacchia, A. 2022. How to distribute data across tasks for meta-learning? In *Proc. of AAAI*.

Coleman, C.; Yeh, C.; Mussmann, S.; Mirzasoleiman, B.; Bailis, P.; Liang, P.; Leskovec, J.; and Zaharia, M. 2019. Selection via proxy: Efficient data selection for deep learning. In *Proc. of ICLR*.

Ding, N.; Chen, X.; Levinboim, T.; Goodman, S.; and Soricut, R. 2021. Bridging the Gap Between Practice and PAC-Bayes Theory in Few-Shot Meta-Learning. *Proc. of NeurIPS*.

Farid, A.; and Majumdar, A. 2021. PAC-bus: Meta-learning bounds via PAC-Bayes and uniform stability. In *Proc. of NeurIPS*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. of ICML*.

Goldblum, M.; Fowl, L.; and Goldstein, T. 2020. Adversarially robust few-shot learning: A meta-learning approach. *Proc. of NeurIPS*.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proc. of NeurIPS*.

Huang, J.; Qu, L.; Jia, R.; and Zhao, B. 2019. O2u-net: A simple noisy label detection approach for deep neural networks. In *Proc. of ICCV*.

Jiang, L.; Meng, D.; Mitamura, T.; and Hauptmann, A. G. 2014a. Easy samples first: Self-paced reranking for zero-example multimedia search. In *Proc. of ACM MM*.

Jiang, L.; Meng, D.; Yu, S.-I.; Lan, Z.; Shan, S.; and Hauptmann, A. 2014b. Self-paced learning with diversity. In *Proc. of NeurIPS*.

Killamsetty, K.; Li, C.; Zhao, C.; Iyer, R.; and Chen, F. 2020. A reweighted meta learning framework for robust few shot learning. *arXiv preprint arXiv:2011.06782*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Kumar, M. P.; Packer, B.; and Koller, D. 2010. Self-Paced Learning for Latent Variable Models. In *Proc. of NeurIPS*.

Lake, B.; Salakhutdinov, R.; Gross, J.; and Tenenbaum, J. 2011. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*.

Li, M.; and Lovell, B. 2022. End to End Generative Meta Curriculum Learning For Medical Data Augmentation. *arXiv preprint arXiv:2212.10086*.

Li, W.; Wang, L.; Li, W.; Agustsson, E.; and Van Gool, L. 2017. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*.

Li, W.; Wang, L.; Zhang, X.; Qi, L.; Huo, J.; Gao, Y.; and Luo, J. 2022. Defensive Few-shot Learning. *IEEE transactions on pattern analysis and machine intelligence*.

Liang, K. J.; Rangrej, S. B.; Petrovic, V.; and Hassner, T. 2022. Few-shot learning with noisy labels. In *Proc. of CVPR*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proc. of ICCV*.

Liu, J.; and Fu, Z. 2021. Curriculum Meta Learning: Learning to Learn from Easy to Hard. In *Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering*.

Liu, T.; Lu, J.; Yan, Z.; and Zhang, G. 2021. PAC-Bayes bounds for meta-learning with data-dependent prior. *arXiv preprint arXiv:2102.03748*.

Lu, J.; Jin, S.; Liang, J.; and Zhang, C. 2020. Robust few-shot learning for user-provided data. *IEEE transactions on neural networks and learning systems*.

Luo, X.; Xu, J.; and Xu, Z. 2022. Channel importance matters in few-shot image classification. In *Proc. of ICML*.

Mairal, J. 2013. Stochastic majorization-minimization algorithms for large-scale optimization. In *Proc. of NeurIPS*.

Mazumder, P.; Singh, P.; and Namboodiri, V. P. 2021. RNNP: A Robust Few-Shot Learning Approach. In *Proc. of WACV*.

McAllester, D. A. 1999. PAC-Bayesian model averaging. In *Proc. of COLT*.

Mehta, B.; Deleu, T.; Raparthy, S. C.; Pal, C. J.; and Paull, L. 2020. Curriculum in gradient-based meta-reinforcement learning. *arXiv preprint arXiv:2002.07956*.

Oreshkin, B.; López, P. R.; and Lacoste, A. 2018. TADAM: Task dependent adaptive metric for improved few-shot learning. In *Proc. of NeurIPS*.

Pentina, A.; and Lampert, C. H. 2015. Lifelong learning with non-iid tasks. *Proc. of NeurIPS*.

Portelas, R.; Romac, C.; Hofmann, K.; and Oudeyer, P.-Y. 2020. Meta automatic curriculum learning. *arXiv preprint arXiv:2011.08463*.

Que, X.; and Yu, Q. 2024. Appendix: Dual-Level Curriculum Meta-Learning for Noisy Few-Shot Learning Tasks. https://github.com/ritminginglab/DCML/tree/main. Accessed: 2024-02-13.

Ravi, S.; and Larochelle, H. 2017. Optimization as a model for few-shot learning. In *Proc. of ICLR*.

Ren, M.; Zeng, W.; Yang, B.; and Urtasun, R. 2018. Learning to reweight examples for robust deep learning. In *Proc. of ICML*.

RIT Research Computing. 2019. Research Computing Services.

Rothfuss, J.; Fortuin, V.; Josifoski, M.; and Krause, A. 2021. PACOH: Bayes-optimal meta-learning with PAC-guarantees. In *Proc. of ICML*.

Seldin, Y.; Laviolette, F.; Cesa-Bianchi, N.; Shawe-Taylor, J.; and Auer, P. 2012. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*.

Shevchuk, G. 2019. Meta-Teaching: Curriculum Generation for Lifelong Learning. In *Proc. of ICML*.

Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Zhou, S.; Xu, Z.; and Meng, D. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Proc. of NeurIPS*.

Sinha, S.; and Ohashi, H. 2023. Difficulty-Net: Learning to Predict Difficulty for Long-Tailed Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Toneva, M.; Sordoni, A.; Combes, R. T. d.; Trischler, A.; Bengio, Y.; and Gordon, G. J. 2019. An empirical study of example forgetting during deep neural network learning. In *Proc. of ICLR*.

Wei, J.; Zhu, Z.; Cheng, H.; Liu, T.; Niu, G.; and Liu, Y. 2022. Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. In *Proc. of ICLR*.

Wu, T.; Li, X.; Li, Y.-F.; Haffari, R.; Qi, G.; Zhu, Y.; and Xu, G. 2021. Curriculum-meta learning for order-robust continual relation extraction. In *Proc. of AAAI*.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *Proc. of ICML*.

Zhan, R.; Liu, X.; Wong, D. F.; and Chao, L. S. 2021. Meta-curriculum learning for domain adaptation in neural machine translation. In *Proc. of AAAI*.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2020. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proc. of CVPR*.

Zhang, J.; Song, J.; Gao, L.; Liu, Y.; and Shen, H. T. 2022a. Progressive Meta-learning with Curriculum. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, J.; Song, J.; Yao, Y.; and Gao, L. 2021. Curriculum-Based Meta-learning. In *Proc. of ACM MM*.

Zhang, W.; Geng, S.; Fu, Z.; Zheng, L.; Jiang, C.; and Hong, S. 2022b. MetaVA: Curriculum Meta-learning and Pre-fine-tuning of Deep Neural Networks for Detecting Ventricular Arrhythmias based on ECGs. *arXiv preprint arXiv:2202.12450*.

Zhou, T.; Wang, S.; and Bilmes, J. 2020a. Robust Curriculum Learning: From clean label detection to noisy label self-correction. In *Proc. of ICLR*.

Zhou, T.; Wang, S.; and Bilmes, J. A. 2020b. Curriculum Learning by Dynamic Instance Hardness. *Proc. of NeurIPS*.