IMPROVED GENERALIZATION RISK BOUNDS FOR META-LEARNING WITH PAC-BAYES-KL ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

By incorporating knowledge from observed tasks, meta-learning algorithms inspired by PAC-Bayes bounds aim to construct a hyperposterior from which an informative prior is sampled for fast adaptation to novel tasks. The goal of PAC-Bayes meta-learning theory is thus to propose an upper bound on the generalization risk over a novel task of the learned hyperposterior. In this work, we first generalize the tight PAC-Bayes-kl bound to the meta-learning setting. Based on the extended PAC-Bayes-kl bound, we further provide three improved PAC-Bayes generalization bounds for meta-learning, leading to better asymptotic behaviour than existing results. By minimizing objective functions derived from the improved bounds, we develop three PAC-Bayes meta-learning algorithms for classification. Moreover, we employ localized distribution-dependent PAC-Bayes priors for meta-learning to yield insights into the role of hyperposterior for learning a novel task. In particular, we identify that when the number of training task is large, utilizing a prior generated from an informative hyperposterior can achieve the same order of PAC-Bayes-kl bound as that obtained through setting a localized distribution-dependent prior for a novel task. Experiments with deep neural networks show that minimizing our bounds can achieve competitive performance on novel tasks w.r.t. previous PAC-Bayes meta-learning methods as well as PAC-Bayes single-task learning methods with localized distribution-dependent priors.

1 INTRODUCTION

A major limitation of deep learning techniques is the need of large amount of training data for new tasks. To address this challenge, *meta-learning* (Thrun & Pratt, 1998; Hospedales et al., 2021), also called *learning-to-learn*, has come into focus in recent years. In meta-learning, an agent extracts information from observed tasks, and aims to facilitate adapting to novel tasks with no data being observed so far. Over the last decade, meta-learning based algorithms have shown promising ability to perform well on a wide range of machine learning problems, e.g., few-shot learning (Munkhdalai & Yu, 2017; Snell et al., 2017) and reinforcement learning (Finn et al., 2017; Liu et al., 2019). However, theoretical properties about meta-learning still remain incompletely understood.

The pioneering theoretical framework for meta-learning was formulated by Baxter (2000), who first assumed that all tasks in meta-learning are independently and identically distributed (i.i.d.) from an unknown distribution, called *environment*, to ensure the relatedness of different tasks. Under this assumption, many following works proposed different meta-learning bounds with different complexity indicators, such as covering number (Baxter, 2000), Gaussian complexity (Maurer, 2009; Maurer et al., 2016; Tripuraneni et al., 2020) and generalized VC-dimension (Ben-David & Schuller, 2003). Other works include utilizing algorithmic stability (Bousquet & Elisseeff, 2002) to study the convergence rate of meta-learning algorithms (Maurer, 2005; Chen et al., 2020). Particularly, PAC-Bayes learning bounds (McAllester, 1999) are regarded as one of the tightest generalization bounds (Langford, 2005; Zhang et al., 2021). This thus causes the recent surge of interest in studying meta-learning with PAC-Bayes analysis (Pentina & Lampert, 2014; Amit & Meir, 2018).

By incorporating knowledge from n training tasks, PAC-Bayes meta-learning algorithm outputs a distribution, termed *hyperposterior*, over the set of all priors (Pentina & Lampert, 2014). When encountering a novel task from the same environment, one informative prior can be generated from the learned hyperposterior to achieve a low generalization risk on the novel task. The PAC-Bayes bound on the generalization risk of the learned hyperposterior is composed of three parts: the empirical risk on training tasks, the environment-level complexity and the task-level complexity. However, most existing PAC-Bayes meta-learning bounds suffer from a slow convergence rate. For example, the task-level complexities in these bounds are slower than $O(\frac{1}{\sqrt{nm}})$ (with m samples per training task), and always contain an extra term of $O(\frac{1}{\sqrt{m}})$ or $O(\frac{1}{\sqrt{n}})$ (see Table 1). Moreover, existing works lack theoretical explanations for why an informative hyperposterior can facilitate novel task adaptation.

In this work, we focus on tackling the above two issues in PAC-Bayes meta-learning theory. Specifically, the first issue to be addressed is the lack of 'identical distribution' between samples independently drawn from different tasks, since different tasks in meta-learning are associated with different distributions. We thus can not directly apply traditional PAC-Bayes-kl bound (that holds for i.i.d. data) to meta-learning. To overcome this problem, we propose Lemma 2 to convert the independent (but non-identically distributed) setting into i.i.d. setting. Hence, we can effectively study meta-learning with traditional PAC-Bayes-kl analysis and derive a generalized PAC-Bayes-kl bound for independent data in Theorem 1. Based on the extended PAC-Bayes-kl bound, we further use its three relaxations, i.e. PAC-Bayes *classic/quadratic/*λ bounds (McAllester, 1999; Pérez-Ortiz et al., 2021; Thiemann et al., 2017), to derive three improved bounds on the generalization risk of the learned hyperposterior in Theorem 2 (which are still called PAC-Bayes *classic/quadratic/*λ meta-learning bounds for brevity). Concretely, the environment-level complexity of $O\big(\sqrt{\frac{\ln\sqrt{n}}{n}}\big)$ in our derived PAC-Bayes meta-learning bounds halves the logarithmic dependence on the number n of training tasks in the numerator. The task-level complexity of our classic bound has a rate of $O(\sqrt{\frac{\ln\sqrt{nm}}{nm}})$, without containing the extra term $O(\frac{1}{\sqrt{m}})$ or $O(\frac{1}{\sqrt{n}})$. In particular, the task-level complexities of our quadratic bound and λ bound have a fast convergence rate of $O(\frac{\ln(nm)}{nm})$ (see Table 1). Moreover, to address the second issue, we employ localized distribution-dependent priors (Catoni, 2007) to meta-learning for explaining the power of hyperposterior. 'Localized' means that the prior is dependent on the data distribution of a particular task and is informative enough for fast adaptation. Hence there is a recent surge of interest in revealing the similarities between the localized prior and the meta-learned prior (Pentina & Lampert, 2014). We demonstrate that, sampling an informative prior from hyperposterior can reach the same tight PAC-Bayes-kl bound as setting a localized prior on a novel task, validating the ability of meta-learning models to adapt to a new task.

Overall, our main contributions are four-fold: **(1)** By reducing the independent setting to the i.i.d. setting with the proposed lemma, we utilize traditional PAC-Bayes-kl analysis in i.i.d. case to derive a generalized PAC-Bayes-kl bound for independent data. **(2)** Based on the relaxations of the extended PAC-Bayes-kl bound, we further obtain three improved PAC-Bayes meta-learning bounds and three bound-minimization meta-learning algorithms. **(3)** For the first time, we demonstrate that when the number of training tasks is large, sampling a prior from an informative hyperposterior with a specific distribution can achieve the same tight PAC-Bayes-kl bound as setting a localized distribution-dependent prior for a novel task. **(4)** Experiments with deep neural networks show that minimizing our derived bounds yields competitive results w.r.t. previous PAC-Bayes meta-learning methods and single-task learning methods that use a localized distribution-dependent prior.

2 Related Work

PAC-Bayes Theory. The goal of PAC-Bayes theory is to connect the generalization error of the learned posterior with its empirical error and its KL-divergence with respect to some prior distribution. Since its first appearance (McAllester, 1998; 1999), PAC-Bayes theory has been developed by (Seeger, 2002; Maurer, 2004; Catoni, 2007) and extended to more general cases, where the data is non-i.i.d. (Ralaivola et al., 2010; Seldin et al., 2012) or the KL-divergence is replaced by other divergence measures (Alquier & Guedj, 2018). The subsequent works connected PAC-Bayes theory to other research areas, such as statistical estimation problems (Zhang, 2006; Germain et al., 2016), kernel methods like SVM (Langford & Shawe-Taylor, 2002) or Gaussian process (Germain et al., 2009; Reeb et al., 2018), and theoretical properties of deep neural networks (Neyshabur et al., 2017; 2018). Another important branch of PAC-Bayes theory focuses on bounding the KL-divergence with an elaborate choice of distribution/data-dependent prior, which was first proposed as *localized PAC-Bayes analysis* in (Ambroladze et al., 2006; Catoni, 2007) and further developed by (Parrado-Hernández et al., 2012; Lever et al., 2013; Rivasplata et al., 2020). In this work, we utilize PAC-Bayes-kl analysis to explore the generalization ability of meta-learning models. Moreover, we provide a case in which sampling the prior from the learned hyperposterior can achieve an equivalent PAC-Bayes-kl upper bound as setting a localized distribution-dependent prior on the novel task.

PAC-Bayes Meta-Learning Theory. Under the i.i.d. task environment framework proposed by Baxter (2000), Pentina & Lampert (2014) first introduced the notation of 'hyperposterior' and provided a PAC-Bayes bound for meta-learning. Following this work, Pentina & Lampert (2015) extended PAC-Bayes meta-learning theory to the general case of non-i.i.d. tasks. Amit & Meir (2018) proposed a demonstration framework that can incorporate existing PAC-Bayes theories in single-task learning to derive new PAC-Bayes bounds for meta-learning. Rothfuss et al. (2021) gave a PAC-bayes meta-learning bound for unbounded loss with exponential moment assumption. Farid & Majumdar (2021) utilized both algorithm stability analysis and PAC-Bayes techniques to yield task-level and environment-level complexities respectively. However, the task-level complexities in most these bounds have a slow convergence rate, such as $O\big(\frac{1}{n\sqrt{m}} + \frac{1}{\sqrt{m}}\big)$ (we suppress the complexity factors in the numerator for clarity, hereafter) in Pentina & Lampert (2014), $O\big(\frac{1}{n\sqrt{m}} + \frac{1}{\sqrt{n}}\big)$ in Rothfuss et al. (2021), or $O(\frac{\ln(n\sqrt{m})}{m})$ in Amit & Meir (2018); Liu et al. (2021) (see Table 1). In contrast, the task-level complexity in our derived classic bound has a convergence rate of $O(\sqrt{\frac{\ln\sqrt{nm}}{nm}})$, and even can achieve a fast rate of $O(\frac{\ln(nm)}{nm})$ in our quadratic/λ bounds (see Theorem 2). Detailed comparisons between these bounds can be found in Table 1 and the discussions below Theorem 2.

3 Preliminary

3.1 Notations for PAC-Bayes Single-Task Learning

For supervised learning problems, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is the product of input space $\mathcal{X}$ and output space $\mathcal{Y}$, $\mathcal{H}$ is a class of hypotheses, and the loss function $l : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ is assumed to be bounded in the interval $[0, 1]$. In single-task learning, an algorithm is given a size-m training sample $S = \{z_i\}_{i=1}^m$, with each z_i drawn i.i.d. from an unknown distribution D over $\mathcal{Z}$. Let $\mathcal{M}_1(A)$ define the set of probability measures over the set A. In PAC-Bayes learning, the agent aims to output a posterior $Q = Q(S, P)$ over $\mathcal{H}$ by training an algorithm with sample S and prior P over $\mathcal{H}$. For any hypothesis $h \in \mathcal{H}$, its expected error over D is denoted by $er(h, D) \triangleq \mathbf{E}_{z\sim D} l(h, z)$ and its empirical error over S is $\widehat{er}(h, S) = \frac{1}{m}\sum_{i=1}^m l(h, z_i)$. Then the expected error $er(Q, D)$ with respect to (w.r.t.) the distribution Q and its empirical counterpart $\widehat{er}(Q, S)$ are defined as:

$$er(Q, D) \triangleq \mathbf{E}_{h\sim Q} er(h, D), \quad \widehat{er}(Q, S) \triangleq \mathbf{E}_{h\sim Q} \widehat{er}(h, S). \tag{1}$$

Denote the KL-divergence, or relative entropy, between two distributions Q and P by $\mathrm{KL}(Q||P) = \mathbf{E}_{h\sim Q} \ln \frac{\mathrm{d}Q}{\mathrm{d}P}$, where $\frac{\mathrm{d}Q}{\mathrm{d}P}$ is the Radon-Nikodym derivative of Q with respect to P. For Bernoulli distributions with biases p and q we use $\mathrm{kl}(p||q)$ as a shorthand for $\mathrm{KL}([p, 1-p]||[q, 1-q])$, the KL-divergence between the two distributions. Then, the classical PAC-Bayes-kl bound in single-task learning (see e.g. Maurer 2004, Thm 5) states that, for any $\delta \in (0, 1)$, any predefined prior P, with probability at least $1 - \delta$ over the draw of i.i.d. sample $S = \{z_i\}_{i=1}^m$, the binary relative entropy between the expected error of any posterior Q and its empirical error can be bounded by :

$$\mathrm{kl}(\widehat{er}(Q, S)||er(Q, D)) \leq \frac{\mathrm{KL}(Q||P) + \ln \frac{2\sqrt{m}}{\delta}}{m}. \tag{2}$$

3.2 Notations for PAC-Bayes Meta-Learning

In meta-learning, the agent observes n training datasets $\{S_i\}_{i=1}^n$, where the samples in each S_i are i.i.d. generated from distribution D_i ($i = 1, \ldots, n$) over the sample space $\mathcal{Z}$ (i.e., $D_i \in \mathcal{M}_1(\mathcal{Z})$). Within the meta-learning framework of Baxter (2000), these n *different* data distributions $\{D_i\}_{i=1}^n$ are assumed to be i.i.d. sampled from the same *environment* τ (i.e., $\tau \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{Z}))$). Therefore, the whole training data $\{S_i\}_{i=1}^n$ in meta-learning are independently but non-identically distributed. Most existing theoretical works take such independent meta-learning setting for analysis (Maurer, 2005; Pentina & Lampert, 2014; Maurer et al., 2016; Amit & Meir, 2018; Chen et al., 2020; Rothfuss et al., 2021) and our work follows this line. We also assume the size of each observed dataset S_i as m to facilitate comparisons among different meta-learning bounds. Then, under the PAC-Bayes meta-learning framework proposed by Pentina & Lampert (2014), the prior P is regarded as a random variable sampled from a predefined distribution $\mathcal{P}$ over priors, called a *hyperprior* (i.e.

Table 1: Different PAC-Bayes bounds on $er(\mathcal{Q})$. **Meta-Learning Bound = Empirical Error + Environment-Level Complexity + Task-Level Complexity**. n is the number of observed tasks, m is the number of samples in S_i ($i \in [n]$). $\mathcal{P}$ and $\mathcal{Q}$ represent hyperprior and hyperposterior respectively, both of which are probability measures over the set of all priors. P is the prior sampled randomly from $\mathcal{P}$, and $Q_i = Q(S_i, P)$ is the posterior for the i-th training task obtained by training PAC-Bayes single-task algorithm with the data S_i and the prior P. $C_1, C_2 > 1$ are two constants.

Different Bounds	**Empirical Error**	**Environment-Level Complexity**	**Task-Level Complexity**
(Pentina & Lampert, 2014)	$\widehat{er}(\mathcal{Q})$	$O(\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})}{\sqrt{n}})$	$O(\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\sum_{i=1}^{n}\mathbf{E}_{P\sim\mathcal{Q}}\mathrm{KL}(Q_i\vert\vert P)}{n\sqrt{m}} + \frac{1}{\sqrt{m}})$
MLAP-M (Amit & Meir, 2018)	$\widehat{er}(\mathcal{Q})$	$O(\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\ln n}{n}})$	$O(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\mathrm{KL}(Q_i\vert\vert P)+\ln(nm)}{m}})$
MLAP-S (Amit & Meir, 2018)	$\widehat{er}(\mathcal{Q})$	$O(\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\ln n}{n}})$	$O(\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\mathrm{KL}(Q_i\vert\vert P)+\ln(n\sqrt{m})}{m})$
PACOH (Rothfuss et al., 2021)	$\widehat{er}(\mathcal{Q})$	$O(\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})}{\sqrt{n}})$	$O(\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\sum_{i=1}^{n}\mathbf{E}_{P\sim\mathcal{Q}}\mathrm{KL}(Q_i\vert\vert P)}{n\sqrt{m}} + \frac{1}{\sqrt{n}})$
λ bound (Liu et al., 2021)	$C_1\widehat{er}(\mathcal{Q})$	$O(\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\ln n}{n}})$	$O(\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\mathrm{KL}(Q_i\vert\vert P)+\ln(n\sqrt{m})}{m})$
classic bound (ours)	$\widehat{er}(\mathcal{Q})$	$O(\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\ln\sqrt{n}}{n}})$	$O(\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^{n}\mathrm{KL}(Q_i\vert\vert P)+\ln\sqrt{nm}}{nm}})$
quadratic bound (ours)	$\widehat{er}(\mathcal{Q})$	$O(\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\ln\sqrt{n}}{n}})$	$O(\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^{n}\mathrm{KL}(Q_i\vert\vert P)+\ln\sqrt{nm}}{nm})$
λ bound (ours)	$C_2\widehat{er}(\mathcal{Q})$	$O(\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\ln\sqrt{n}}{n}})$	$O(\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^{n}\mathrm{KL}(Q_i\vert\vert P)+\ln\sqrt{nm}}{nm})$

$\mathcal{P} \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{H})))$. During the meta-training phase, with the prior P sampled from hyperprior $\mathcal{P}$ for each observed task, the agent further incorporates the information from the n training tasks and computes a *hyperposterior* $\mathcal{Q} \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))$. During the meta-test phase, an informative prior is sampled from the learned hyperposterior $\mathcal{Q}$ to adapt to the novel task. The goal of PAC-Bayes meta-learning algorithms is thus to learn an informative hyperposterior $\mathcal{Q}$ with a hyperprior $\mathcal{P}$ and the training datasets $\{S_i\}_{i=1}^{n}$ as input. Formally, the quality of $\mathcal{Q}$ can be measured by the *transfer risk* (Maurer, 2009) on the data distribution D randomly sampled from the environment τ:

$$er(\mathcal{Q}) \triangleq \mathbf{E}_{P\sim\mathcal{Q}}\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m}er(Q(S,P),D). \tag{3}$$

Due to the difficulty of minimizing $er(\mathcal{Q})$ directly, we consider the average empirical error over the n training tasks, which is called *empirical multi-task risk* $\widehat{er}(\mathcal{Q})$ and can be minimized during meta-training phase, and the *expected multi-task risk* $\widetilde{er}(\mathcal{Q})$ as the empirical proxies for $er(\mathcal{Q})$:

$$\widehat{er}(\mathcal{Q}) \triangleq \mathbf{E}_{P\sim\mathcal{Q}}1/n\sum_{i=1}^{n}\widehat{er}(Q(S_i,P),S_i), \quad \widetilde{er}(\mathcal{Q}) \triangleq \mathbf{E}_{P\sim\mathcal{Q}}1/n\sum_{i=1}^{n}er(Q(S_i,P),D_i). \tag{4}$$

4 Theoretical Results

In Section 4.1, we extend the PAC-Bayes-kl bound from i.i.d. setting to independent meta-learning setting. In Section 4.2 we use the extended PAC-Bayes-kl bound to obtain three improved PAC-Bayes bounds for meta-learning. In Section 4.3, we demonstrate that on a novel task, utilizing the prior sampled from an informative hyperposterior can achieve the same tight PAC-Bayes-kl bound as that obtained by setting a localized prior. Section 4.4 details how to develop PAC-Bayes meta-learning algorithms with our three improved bounds. All proofs are deferred to the Appendix.

4.1 Generalizing PAC-Bayes-kl Bound from i.i.d. Setting to Independent Meta-Learning Setting

We first provide a PAC-Bayesian demonstration template that can cope with independent meta-learning setting. Such demonstration strategy is originated from (Germain et al., 2009; Lever et al., 2013) for i.i.d. case. We simply generalize such framework to the independent case and use it to analyze how to extend the PAC-Bayes-kl bound for independent but non-identically distributed data.

Lemma 1 *Let $\mathcal{S} = (\xi_1, \ldots, \xi_K)$ be a size-K random vector, with each component $\xi_k (k \in [K])$ drawn independently (not necessarily identically distributional) according to the measure ν_k over the set A_k. Let $\mathcal{F}$ be a set of random variables f, π be a fixed measure over $\mathcal{F}$ that does not depend on $\mathcal{S}$. Let $R(f) = \frac{1}{K}\sum_{k=1}^{K}\mathbf{E}_{\xi_k}g_k(f,\xi_k)$, $r(f) = \frac{1}{K}\sum_{k=1}^{K}g_k(f,\xi_k)$, where $g_k : \mathcal{F} \times A_k \to [0,1]$ is a bounded function. Let $\Phi : [0,1] \times [0,1] \to \mathbb{R}$ be a convex function. Then for any $\delta \in (0,1)$, any $t > 0$, with probability at least $1-\delta$ over the draw of $\mathcal{S}$, for any distributions ρ over $\mathcal{F}$, we have*

$$\Phi(\mathbf{E}_{f\sim\rho}R(f), \mathbf{E}_{f\sim\rho}r(f)) \leq \frac{1}{t}\big[\,\mathrm{KL}(\rho||\pi) + \ln\frac{\mathbf{E}_{f\sim\pi}\mathbf{E}_{\mathcal{S}}e^{t\Phi(R(f),r(f))}}{\delta}\big].$$

To derive PAC-Bayes bounds on $er(\mathcal{Q})$ for meta-learning, some existing works (Pentina & Lampert, 2014; 2015; Rothfuss et al., 2021) can be considered to set the convex function $\Phi(R(f), r(f)) = R(f) - r(f)$ in Lemma 1 and use this lemma to bound $er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})$ and $\widetilde{er}(\mathcal{Q}) - \widehat{er}(\mathcal{Q})$ respectively. Concretely, using Hoeffding's Lemma (Hoeffding, 1963) to bound $\mathbf{E}\mathrm{e}^{t\Phi(R(f),r(f))}$ with $\mathrm{e}^{t^2/8K}$, the right-hand-side (RHS) in the inequality of Lemma 1 can be written in the form of $A/t + tB/K$, where $A, B > 0$. Setting $t = \sqrt{K}$ obtains a bound of slow rate $O(1/\sqrt{K})$ on $\mathbf{E}_{f\sim\rho}R(f)$. Hence, we can see that, setting $\Phi(R(f), r(f)) = R(f) - r(f)$ just yields a bound of slow convergence rate. To obtain an improved bound for meta-learning setting, we set $\Phi(R(f), r(f)) = \mathrm{kl}(r(f)||R(f))$ and expect to derive generalized PAC-Bayes-kl bound that can be converted to an explicit bound of fast rate $O(\ln K/K)$ on $\mathbf{E}_{f\sim\rho}R(f)$. Recall that for single-task learning, setting all measures ν_k's in Lemma 1 as the same, setting $\Phi(R(f), r(f))$ as $\mathrm{kl}(r(f)||R(f))$ and bounding $\mathbf{E}_{\mathcal{S}}e^{K\,\mathrm{kl}(r(f),R(f))}$ with $2\sqrt{K}$, we can recover the PAC-Bayes-kl bound for the i.i.d. case in Eq. (2). However, the i.i.d. assumption is a necessary condition to bound $\mathbf{E}_{\mathcal{S}}e^{K\,\mathrm{kl}(r(f),R(f))}$ with $2\sqrt{K}$ (see Maurer 2004, Eq.(1)). Hence, traditional PAC-Bayes-kl bound can not be directly applied to meta-learning setting where there is no i.i.d. data assumption. To address this issue, we use the following lemma to convert the independent setting into the i.i.d. setting, thus giving an upper bound on $\mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(r(f)||R(f))}$ with $\mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{1}{K}\sum_{k=1}^{K}\bar{Y}_k||\bar{\mu})}$ of i.i.d. random variables $\{\bar{Y}_k\}_{k=1}^{K}$.

Lemma 2 *Let $X_1, ..., X_K$ be a sequence of independent random variables, such that $X_k \in [0, 1]$ almost surely and $\mathbf{E}X_k = \mu_k$, for $k = 1, ..., K$. Let $\bar{Y}_1, ..., \bar{Y}_K$ be i.i.d. Bernoulli random variables with $\mathbf{E}\bar{Y}_k = \frac{1}{K}\sum_{k=1}^{K}\mu_k = \bar{\mu}$. Let $X = \sum_{k=1}^{K} X_k, \bar{Y} = \sum_{k=1}^{K}\bar{Y}_k$. Then we have*

$$\mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{1}{K}\sum_{k=1}^{K}X_k||\bar{\mu})} \le \mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{1}{K}\sum_{k=1}^{K}\bar{Y}_k||\bar{\mu})}.$$

A similar result can be found in (Maurer, 2004, Thm 1) where $\{X_k\}_{k=1}^{K}$ are i.i.d. data. Our derived Lemma 2 can be considered as the generalized version where $\{X_k\}_{k=1}^{K}$ are just independent random variables. The proof of Lemma 2 can be found in Corollary 2 of Appendix B.1. Setting X_k as $g_k(f, \xi_k)$ ($k \in [K]$), we can apply Lemma 2 to show that $\mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(r(f)||R(f))} \le \mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{1}{K}\sum_{k=1}^{K}\bar{Y}_k||\bar{\mu})} \le 2\sqrt{K}$, leading to the following generalized PAC-Bayes-kl bound.

Theorem 1 *In the setting of Lemma 1, set $\Phi(R(f), r(f)) = \mathrm{kl}(r(f)||R(f))$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the draw of $\mathcal{S}$, we have for any $\rho \in \mathcal{M}_1(\mathcal{F})$,*

$$\mathrm{kl}(\mathbf{E}_{f\sim\rho}r(f)||\mathbf{E}_{f\sim\rho}R(f)) \le \frac{\mathrm{KL}(\rho||\pi) + \ln(2\sqrt{K}/\delta)}{K}.$$

There are two points to note about the above Theorem 1: (1) Setting all measures ν_k's as the same, we can obtain the traditional PAC-Bayes-kl bound for i.i.d. case in Eq. (2). (2) Actually, we can also use the generalized chromatic PAC-Bayes bound in (Ralaivola et al., 2010, Thm 28) (just set the fractional chromatic number in it as 1) to obtain an upper bound of slow convergence rate $O(\sqrt{\ln K/K})$ on $\mathbf{E}_{f\sim\rho}R(f)$ of independent data. In the next section, by Pinsker's inequalities $\mathrm{kl}(p||q) \ge 2(p-q)^2$ and $\mathrm{kl}(p||q) \ge (p-q)^2/(2q)$, we will use the relaxations of the extended PAC-Bayes-kl bound (i.e. the three explicit bounds on $\mathbf{E}_{f\sim\rho}R(f)$ in Corollary 3 of Appendix B.2, two of which have a fast convergence rate $O(\ln K/K)$.) to derive three novel PAC-Bayes meta-learning bounds on $er(\mathcal{Q})$.

4.2 Three Improved PAC-Bayes Bounds for Meta-Learning Based on Extended PAC-Bayes-kl Bound

To give a bound on $er(\mathcal{Q})$, we choose to bound $er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})$ and $\widetilde{er}(\mathcal{Q}) - \widehat{er}(\mathcal{Q})$. The derived two bounds are called **Environment-Level Complexity** and **Task-Level Complexity**, respectively. In this work, we first bound $\mathrm{kl}(\widetilde{er}(\mathcal{Q})||er(\mathcal{Q}))$ and $\mathrm{kl}(\widehat{er}(\mathcal{Q})||\widetilde{er}(\mathcal{Q}))$ with the extended PAC-Bayes-kl bound in Theorem 1, and then give the explicit bounds on $er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})$ and $\widetilde{er}(\mathcal{Q}) - \widehat{er}(\mathcal{Q})$. Concretely, the explicit bounds are derived as: **(1)** After obtaining a bound on $\mathrm{kl}(\widetilde{er}(\mathcal{Q})||er(\mathcal{Q}))$, we can only use Pinsker's inequality $\mathrm{kl}(p||q) \ge 2(p-q)^2$ to bound $|\widetilde{er}(\mathcal{Q}) - er(\mathcal{Q})|$. **(2)** After deriving the bound on $\mathrm{kl}(\widehat{er}(\mathcal{Q})||\widetilde{er}(\mathcal{Q}))$, we can still use the inequality $\mathrm{kl}(p||q) \ge 2(p-q)^2$ to bound $|\widehat{er}(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})|$. Besides, since $\widehat{er}(\mathcal{Q}) < \widetilde{er}(\mathcal{Q})$, we can use the stronger inequality $\mathrm{kl}(p||q) \ge \frac{(p-q)^2}{2q}$ to give a sharper bound on $\widetilde{er}(\mathcal{Q})$. Combing the derived environment-level complexity and task-level complexity with union bound yields three improved meta-learning bounds below.

Theorem 2 *Denote $Q(S_i, P)$ by Q_i for brevity, $\forall i \in [n]$. For any predefined hyperprior $\mathcal{P} \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))$, any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draw of samples $\{S_1, \dots, S_n\}$, simultaneously for all hyperposterior $\mathcal{Q} \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))$, we have*
(i) PAC-Bayes-classic meta-learning bound:

$$er(\mathcal{Q}) \leq \widehat{er}(\mathcal{Q}) + \sqrt{\frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \ln \frac{2\sqrt{n}}{\delta}}{2n}} + \sqrt{\frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}} \sum_{i=1}^{n} \mathrm{KL}(Q_i||P) + \ln \frac{2\sqrt{nm}}{\delta}}{2nm}}.$$

(ii) PAC-Bayes-quadratic meta-learning bound:

$$er(\mathcal{Q}) \leq \sqrt{\frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \ln \frac{2\sqrt{n}}{\delta}}{2n}} + \Big(\sqrt{\widehat{er}(\mathcal{Q}) + \frac{\Delta + \ln \frac{4\sqrt{nm}}{\delta}}{2nm}} + \sqrt{\frac{\Delta + \ln \frac{4\sqrt{nm}}{\delta}}{2nm}}\Big)^2.$$

where $\Delta = \mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}} \sum_{i=1}^{n} \mathrm{KL}(Q_i||P)$.
(iii) PAC-Bayes-λ meta-learning bound, for any $\lambda \in (0,2)$:

$$er(\mathcal{Q}) \leq \sqrt{\frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \ln \frac{2\sqrt{n}}{\delta}}{2n}} + \frac{\widehat{er}(\mathcal{Q})}{1-\lambda/2} + \frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}} \sum_{i=1}^{n} \mathrm{KL}(Q_i||P) + \ln \frac{4\sqrt{nm}}{\delta}}{nm\lambda(1-\lambda/2)}.$$

The detailed comparisons between our derived three meta-learning bounds and existing bounds can be found in Table 1. We can see that: **(1)** Compared with the environment-level complexities of $O\big(\sqrt{\frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P})+\ln n}{n}}\big)$ in Amit & Meir (2018); Liu et al. (2021), the environment-level complexities of $O\big(\sqrt{\frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P})+\ln\sqrt{n}}{n}}\big)$ in our derived three bounds halve the logarithmic dependence on the number n of training tasks in the numerator. **(2)** The task-level complexities in existing bounds can not fully utilize the whole nm training samples (e.g., $O\big(\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta+\ln(n\sqrt{m})}{m}\big)$ in Amit & Meir (2018) and $O\big(\frac{\Delta}{n\sqrt{m}} + \frac{1}{\sqrt{n}}\big)$ in Rothfuss et al. (2021), where the definition of complexity Δ can be found in Theorem 2 (ii)). In contrast, the task-level complexities in our derived classic bound can fully utilize the whole nm training samples (i.e., without the extra term $O(\frac{1}{\sqrt{n}})$ or $O(\frac{1}{\sqrt{m}})$) and has a convergence rate of $O\big(\sqrt{\frac{\Delta+\ln\sqrt{nm}}{nm}}\big)$. **(3)** Our derived quadratic bound and λ bound can achieve a fast rate of $O\big(\frac{\Delta+\ln\sqrt{nm}}{nm}\big)$ (as long as the empirical multi-task risk $\widehat{er}(\mathcal{Q})$ is close to zero), much sharper than that of $O\big(\frac{\Delta+\ln(n\sqrt{m})}{m}\big)$ in the quadratic bound and λ bound of Liu et al. (2021). To validate the practical effectiveness of the above three improved bounds, we will set them as training objectives to develop three bound-minimization meta-learning algorithms for classification in Section 4.4.

4.3 PAC-Bayes-kl Bound on Novel Task with the Prior Sampled from the Informative Hyperposterior

With PAC-Bayes-kl analysis, we explore the benefits of generating a prior from hyperposterior for a novel task. Pentina & Lampert (2014) have claimed intuitively that sampling a prior from hyperposterior is similar to setting a localized prior for the novel task. In this work, we rigorously validate this intuition. For the novel task, consider the localized prior P and the posterior Q w.r.t. ν over $\mathcal{H}$ as Gibbs Distributions[1]: $\frac{\mathrm{d}P}{\mathrm{d}\nu}(h) = \frac{1}{\beta}\exp\{-\gamma \mathbf{E}_{z\sim D} l(h,z)\}$, $\quad \frac{\mathrm{d}Q}{\mathrm{d}\nu}(h) = \frac{1}{\beta'}\exp\{-\frac{\gamma}{m}\sum_{j=1}^{m} l(h,z_j)\}$, where $\beta = \int_{\mathcal{H}} \exp\{-\gamma \mathbf{E}_{z\sim D} l(h,z)\}\mathrm{d}\nu(h)$, $\beta' = \int_{\mathcal{H}} \exp\{-\frac{\gamma}{m}\sum_{j=1}^{m} l(h,z_j)\}\mathrm{d}\nu(h)$ ($\gamma > 0$) are normalization constants. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, Lever et al. (2013, Theorem 6) gives the tight single-task PAC-Bayes-kl bound with localized prior as follow:

$$\mathrm{kl}(\widehat{er}(Q,S)||er(Q,D)) \leq \frac{1}{m}\Big(\frac{\gamma^2}{2m} + \gamma\sqrt{\frac{2}{m}\ln\frac{2\sqrt{m}}{\delta}} + \ln\frac{2\sqrt{m}}{\delta}\Big). \tag{5}$$

To derive a PAC-Bayes bound on a novel task (randomly sampled from the environment) with the prior generated from hyperposterior $\mathcal{Q}$, we need to choose $\mathcal{Q}$ as a special form that has a high mass around the informative prior, since directly analyzing the generalization ability of arbitrary $\mathcal{Q}$ is hard and even unfeasible. In this work, we set the learned hyperposterior as the Dirac measure that only

[1] Gibbs distribution P has a form of $\frac{\mathrm{d}P}{\mathrm{d}\nu}(h) = \frac{1}{\beta}\exp\{-H_p(h)\}$, where H_p is called energy function.

has mass at an informative prior P. Such informative prior P should contain as much information of the n training tasks as possible for transferring knowledge to the novel task. Thus, analogous to the form of the localized Gibbs prior above Eq. (5) whose energy function is the expected loss on the novel task $\gamma \mathbf{E}_{z\sim D} l(h,z)$, we assume the meta-learned prior P as a Gibbs distribution whose energy function is defined as the expected loss on n training tasks $\frac{\gamma}{n}\sum_{i=1}^{n} \mathbf{E}_{z\sim D_i} l(h,z)$. The formal definition of the prior P sampled from $\mathcal{Q}$ is presented in Eq. (6) and the proof of the existence of the intermediate measure ν in Eq. (6) can be found in Claim 1 in Appendix B.3.1. Then we can derive a tight PAC-Bayes-kl bound on the novel task D that is randomly sampled from the environment τ.

Theorem 3 *In meta-test phase, consider a novel task equipped with training sample* $S = \{z_j\}_{j=1}^{m} \sim D^m$ *where the distribution* D *is sampled randomly from environment* τ*. Consider the prior* P *sampled from the learned hyperposterior, and the posterior* Q *for this novel task as follow:*

$$\frac{\mathrm{d}P}{\mathrm{d}\nu}(h)=\widetilde{P}(h)=\frac{1}{\beta}\exp\{-\frac{\gamma}{n}\sum_{i=1}^{n}\mathbf{E}_{z\sim D_i} l(h,z)\},\ \frac{\mathrm{d}Q}{\mathrm{d}\nu}(h)=\widetilde{Q}(h)=\frac{1}{\beta'}\exp\{-\frac{\gamma}{m}\sum_{j=1}^{m} l(h,z_j)\}, \quad (6)$$

where β, β' *are both normalization constants. Denote* $\underset{D\sim\tau,S\sim D^m}{\mathbf{E}}$ *by* $\underset{D,S}{\mathbf{E}}$ *for brevity. Then for any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, the following inequalities hold simultaneously,*

$$\forall Q,\ \underset{D,S}{\mathbf{E}}\,\mathrm{kl}(\widehat{er}(Q,S)||er(Q,D)) \le \frac{\mathbf{E}_{D,S}\,\mathrm{KL}(Q(S,P)||P)+\ln(2\sqrt{m}/\delta)}{m},$$

$$\mathbf{E}_{D,S}\,\mathrm{KL}(Q(S,P)||P) \le \frac{\gamma^2}{2m}+O(\sqrt{\frac{\ln\sqrt{n}}{n}})+\gamma(\frac{\ln(3\sqrt{m}/\delta)}{2m}+O(\sqrt{\frac{\ln\sqrt{n}+\ln\sqrt{m}}{nm}}))^{\frac{1}{2}}.$$

Denote the RHS of the above second inequality on $\mathbf{E}_{D,S}\,\mathrm{KL}(Q(S,P)||P)$ by $\phi(n,m)$, and the RHS of the inequality in Eq. (5) by $\psi(m) = \frac{1}{m}\big(\frac{\gamma^2}{2m}+\gamma\sqrt{\frac{2}{m}\ln\frac{2\sqrt{m}}{\delta}}+\ln\frac{2\sqrt{m}}{\delta}\big)$. We expect to compare $\phi(n,m)$ and $\psi(m)$ to reveal the benefits of meta-learning over single-task learning. For fair comparisons, in both settings, we set the same Gibbs posterior distribution Q over hypothesis space for the novel task. The only difference is that, in meta-learning setting, the prior P sampled from the hyperposterior has the form of Eq. (6) and only contains the information about the previous n training tasks, and the upper bound $\phi(n,m)$ is calculated on the expectation of the KL-divergence w.r.t. the task environment distribution; while in single-task learning setting, the prior P is set as an informative localized prior, and the bound $\psi(m)$ is calculated between such localized prior and the Gibbs posterior. Therefore, for any fixed m, $\lim_{n\to\infty}\phi(n,m) = \frac{\gamma^2}{2m}+\gamma\sqrt{\frac{\ln(3\sqrt{m}/\delta)}{2m}}$, and $\lim_{n\to\infty}\frac{\phi(n,m)+\ln(2\sqrt{m}/\delta)}{m} \le \psi(m)$. This means that, when the number n of training task becomes large, the upper bound on $\mathbf{E}_{D,S}\,\mathrm{KL}(Q(S,P)||P)$ in meta-learning is tighter than the bound on $\mathrm{KL}(Q||P)$ obtained by setting a localized prior in single-task learning. Such result does theoretically reveal the benefits of sampling the prior from an informative hyperposterior in meta-learning.

4.4 PAC-Bayes Bound-Minimization Algorithms for Meta-Learning with Stochastic Deep Neural Networks

By setting the three bounds in Theorem 2 as minimization objectives, we can develop three PAC-Bayes bound-minimization algorithms for meta-learning with deep neural networks. Concretely, we use stochastic neural networks (e.g. Blundell et al. 2015) and thus define the hypothesis space $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$ as the set of neural networks with certain parameters. We further need to specify the form of hyperprior and hyperposterior. Following previous work (Amit & Meir, 2018), we set both the hyperprior and hyperposterior as the isotropic Gaussian distribution: $\mathcal{P} = \mathcal{N}(0,\kappa_{\mathcal{P}}^2 I_{d\times d})$, $\mathcal{Q}_\theta = \mathcal{N}(\theta,\kappa_{\mathcal{Q}}^2 I_{d\times d})$, where $\kappa_{\mathcal{P}}, \kappa_{\mathcal{Q}} > 0$ are constants, d is the dimension of the parameter w of prior P, and θ is the optimization parameter. Then the KL-divergence between $\mathcal{Q}_\theta$ and $\mathcal{P}$ can be calculated as: $\mathrm{KL}(\mathcal{Q}_\theta||\mathcal{P}) = \frac{||\theta||_2^2}{2\kappa_{\mathcal{P}}^2}$. Next, we consider the form of prior and posterior over the hypothesis space $\mathcal{H}$. Recall that $\mathcal{H}$ is the family of functions parameterized by a weight vector $\{h_w : w \in \mathbb{R}^d\}$, the posterior/prior is thus the distribution over $\mathbb{R}^d$. We choose the prior P_θ and the posteriors Q_{ϕ_i} ($\phi_i \in \mathbb{R}^d$ is the hyperparameter) to be factorized Gaussian: $P_\theta(w) = \prod_{k=1}^{d}\mathcal{N}(w_k;\mu_{P,k},\sigma_{P,k}^2)$, $Q_{\phi_i}(w) = \prod_{k=1}^{d}\mathcal{N}(w_k;u_{i,k},\sigma_{i,k}^2)$, where $\theta = (\mu_P,\rho_P) \in \mathbb{R}^{2d}$ is composed of the means $\mu_{P,k}$ and log-variances of each weight

Table 2: Comparison of different PAC-Bayes bounds on 20 test tasks (the ± shows the 95% confidence interval) in 100/200-swap shuffled pixels environment and permuted labels environment.

	100-Swap Shuffled pixels		**200-Swap Shuffled pixels**		**Permuted labels**	
Method	**Test bound**	**Test error** $(\%)$	**Test bound**	**Test error** $(\%)$	**Test bound**	**Test error** $(\%)$
(Pentina & Lampert, 2014)	0.189 ± 0.023	1.939 ± 0.001	0.240 ± 0.030	2.631 ± 0.002	6.026 ± 0.436	15.660 ± 0.063
MLAP-M (Amit & Meir, 2018)	0.137 ± 0.037	1.607 ± 0.000	0.197 ± 0.019	1.948 ± 0.001	1.799 ± 0.056	4.229 ± 0.003
MLAP-S (Amit & Meir, 2018)	0.133 ± 0.034	1.629 ± 0.001	0.285 ± 0.049	1.972 ± 0.001	4.947 ± 0.339	8.600 ± 0.016
MLAP-VB (Amit & Meir, 2018)	0.138 ± 0.024	1.606 ± 0.001	0.161 ± 0.002	1.962 ± 0.001	4.623 ± 0.308	9.754 ± 0.130
PACOH (Rothfuss et al., 2021)	0.181 ± 0.023	1.919 ± 0.001	0.221 ± 0.029	2.630 ± 0.002	5.434 ± 0.416	12.520 ± 0.061
λ-bound (Liu et al., 2021)	0.067 ± 0.015	1.630 ± 0.001	0.151 ± 0.015	2.097 ± 0.001	3.830 ± 0.181	11.340 ± 0.017
classic bound (ours)	0.138 ± 0.019	1.585 ± 0.000	0.193 ± 0.018	$\mathbf{1.911 \pm 0.001}$	1.790 ± 0.054	4.164 ± 0.003
quadratic bound (ours)	0.081 ± 0.019	1.644 ± 0.000	0.157 ± 0.024	1.929 ± 0.001	1.950 ± 0.051	4.753 ± 0.003
λ bound (ours)	$\mathbf{0.043 \pm 0.008}$	$\mathbf{1.575 \pm 0.001}$	$\mathbf{0.093 \pm 0.012}$	1.932 ± 0.001	$\mathbf{1.698 \pm 0.051}$	$\mathbf{4.064 \pm 0.003}$

$\rho_{P,k} = \ln \sigma_{P,k}^2, k \in [d]$. The posterior vectors $\phi_i = (\mu_i, \rho_i) \in \mathbb{R}^{2d}$ has a similar structure. Thus, the KL-divergence in task-level complexity of meta-learning bounds has a simple analytic form: $\mathbf{E}_{P_\theta \sim \mathcal{Q}_\theta} \sum_{i=1}^n \mathrm{KL}(Q_{\phi_i} || P_\theta) = \mathbf{E}_{P_\theta \sim \mathcal{Q}_\theta} \frac{1}{2} \sum_{i,k=1}^{n,d} \{\ln \frac{\sigma_{P,k}^2}{\sigma_{i,k}^2} + \frac{\sigma_{i,k}^2 + (\mu_{i,k} - \mu_{P,k})^2}{\sigma_{P,k}^2} - 1\}$, where the expectation $P_\theta \sim \mathcal{Q}_\theta$ can be approximated through Monte-Carlo method by adding Gaussian noise to the parameter θ, i.e., $\theta := \theta + \epsilon, \epsilon \sim \mathcal{N}(0, \kappa_{\mathcal{Q}}^2 I_{d \times d})$. Computing the PAC-Bayes bounds in Theorem 2 with the above results and using the gradient descent method to optimize the parameter $\{\theta, \phi_1, ..., \phi_n\}$ can learn the hyperposterior $\mathcal{Q}$. The detailed pseudo-code of our PAC-Bayes meta-learning algorithms (including meta-training and meta-test parts) can be found in Appendix C.2.

5 Experiments

In this section, we run our proposed PAC-Bayes bound-minimization meta-learning algorithms on image classification problems, and make detailed comparisons with existing methods. All experimental details are set the same as that in (Amit & Meir, 2018; Liu et al., 2021) for fair comparisons.

5.1 Experimental Setup

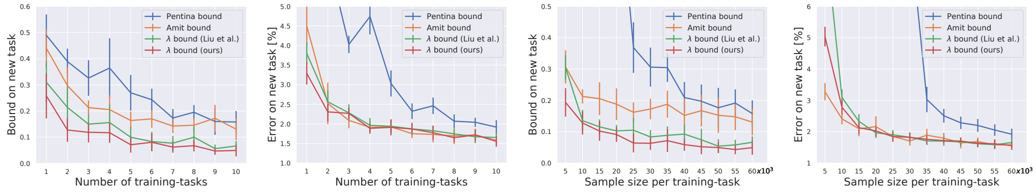
Dataset and Task Environment. We conduct experiments with three different task environments, constructed by augmentations of the MNIST dataset (Lecun et al., 1998). The first/second environment is termed *permuted pixels*, where each task is created by swapping a number of pixel locations (i.e., 100 and 200 locations) for each image. The set of pixels to be swapped is fixed and different for each classification task. The third environment is called *permuted labels*, where each task is created by a random permutation of image labels. During the meta-training phase, each task is constructed with 60,000 training images and 10,000 test images. During the meta-test phase, each task is constructed with much fewer training examples (2,000) and 10,000 test images. For permuted pixels and permuted labels experiments, the number n of training task is set as 10 and 5 respectively.

Neural Network Architecture. We choose a 4-layer fully-connected network for shuffled pixels experiments, and a 4-layer convolutional network for permuted labels experiments. We use the bounded cross-entropy loss (Pérez-Ortiz et al., 2021) for model optimization. The optimizer is Adam (Kingma & Ba, 2015) with a learning rate of 10^{-3} for running different algorithms. More details about the network architecture and the experimental setting can be found in Appendix C.1.

5.2 Experimental Results

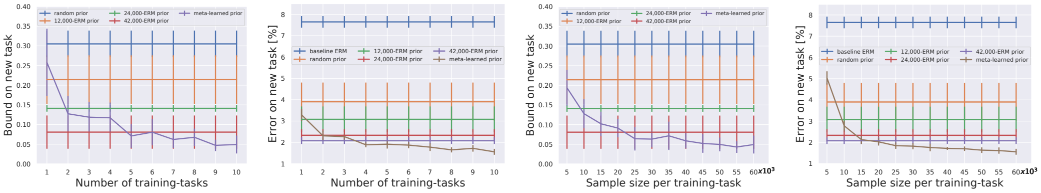
Different PAC-Bayes Methods. We compare average test error and test bound on 20 meta-test tasks among the following methods: (Pentina & Lampert, 2014), MLAP-M (Amit & Meir, 2018, Theorem 2), MLAP-S (Amit & Meir, 2018, Theorem 4), MLAP-VB (Amit & Meir, 2018, Eq.(23)), PACOH (Rothfuss et al., 2021, Theorem 2), λ bound (Liu et al., 2021, Theorem 1). The detailed formulations of the above PAC-Bayes meta-learning bounds can be found in Table 1. We reproduce the experimental results of these methods by directly running the code released online[2] from (Liu et al., 2021), and run our algorithm by replacing others' bounds with our bounds derived from Theorem 2. Concretely, the test bound is calculated as a single-task PAC-Bayes bound on a novel task with the $2,000$ training samples, with a prior sampled from the learned hyperposterior. The test error is the classification error on a novel task. The number of training epochs during the meta-training/meta-test phase is set as 150/200 respectively. More details about the computation of the test bound and test error can be found in Appendix C.2. Experimental results are reported in Table 2.

[2] `https://github.com/tyliu22/Meta-learning-PAC-Bayes-bound-with-data-depedent-prior`



(a) Test Bound (b) Test Error (c) Test Bound (d) Test Error

Figure 1: Comparisons of average test bounds and test errors between other three latest bounds and our λ bound on new tasks (average over 20 meta-test tasks from 100-pixel-shuffled environment).



(a) Test Bound (b) Test Error (c) Test Bound (d) Test Error

Figure 2: Comparisons of average test bounds and test errors between localized priors, and meta-learned priors obtained by minimizing our λ bound across a wide range of number of training tasks and sample size per task (average over 20 meta-test tasks from 100-pixel-shuffled environment).

Classification Performance. We can draw the following conclusions from the comparisons in Table 2: **(i)** In all settings, minimizing our PAC-Bayes-λ bound can always achieve the tightest test bounds the lowest test error over the meta-test tasks for different environments, which is consistent with our theoretical analysis that λ bound is the tightest bound derived in this work. **(ii)** Besides PAC-Bayes-λ bound, our derived PAC-Bayes-classic bound and quadratic bound can also obtain competitive performance in terms of test error and the quantity of test bound w.r.t. other methods.

5.3 More Discussions

Comparisons between Our λ Bound and Others. We provide detailed comparisons of the convergence performance in Figure 1 between our tightest λ bound and other three classical bounds (Pentina & Lampert, 2014; Amit & Meir, 2018; Liu et al., 2021), across a wide range of the number n of the training tasks and the sample size m per task in 100-pixel-shuffled environment. When n changes, m is set as $60,000$; when m changes, n is set as 10. We can find that, minimizing our λ bound always obtains the tightest test bounds and achieves competitive test errors on novel tasks.

Comparisons between Localized Priors and our Meta-Learned Priors. We provide detailed comparisons between our meta-learned priors and different localized priors in Figure 2. Our meta-learned priors are sampled from the hyperposterior obtained through minimizing our λ bound in Theorem 2. The setting of n, m and the environment is the same as that in Figure 1. 'Baseline ERM' represents purely empirical risk minimization (ERM) over 2,000 training samples on each novel task. 'k-ERM prior' represents the localized prior learned from ERM over the $k(\in(0, 58,000))$ training samples (non-overlapped with the 2,000 training samples) from each novel task. 'random prior' can be considered as '0-ERM' prior. The learned k-ERM prior is subsequently used for PAC-Bayes single-task learning (with λ bound) to compute test bound/error. More details of learning k-ERM priors can be found in Appendix C.4. We can see that: with the increase of n or m, the test bounds and test errors of meta-learned prior can decrease to the low values achieved by setting the optimal 42,000-ERM prior, validating the benefits of sampling a prior from an informative hyperposterior.

6 Conclusion

In this work, we generalize the PAC-Bayes-kl bound from the i.i.d. setting to the independent meta-learning setting. Based on the extended PAC-Bayes-kl bound, we provide three improved PAC-Bayes meta-learning bounds. Minimizing the training objectives derived from these bounds leads to three PAC-Bayes meta-learning algorithms, yielding competitive experimental results on novel tasks w.r.t. existing methods. Moreover, we show that the PAC-Bayes-kl bound obtained by sampling a prior from an informative hyperposterior is equivalent to the one obtained by setting localized prior for the novel task, from both theoretical and experimental standpoints. Our ongoing research includes extending our theoretical results to unbounded loss and heavy-tailed data.

REFERENCES

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107 (5):887–902, 2018.

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes bounds. In *NeurIPS*, pp. 9–16, 2006.

Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *ICML*, pp. 205–214, 2018.

Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

Shai Ben-David and Reba Schuller. Exploiting task relatedness for mulitple task learning. In *COLT*, pp. 567–580, 2003.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *ICML*, pp. 1613–1622, 2015.

Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Inst of Mathematical Statistic, 2007.

Jiaxin Chen, Xiao-Ming Wu, Yanke Li, Qimai LI, Li-Ming Zhan, and Fu-Lai Chung. A closer look at the training strategy for modern meta-learning. In *NeurIPS*, 2020.

Alec Farid and Anirudha Majumdar. PAC-BUS: meta-learning bounds via PAC-Bayes and uniform stability. In *NeurIPS*, 2021.

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

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *ICML*, pp. 353–360, 2009.

Pascal Germain, Francis R. Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In *NeurIPS*, pp. 1876–1884, 2016.

Wassily Hoeffding. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, 27(3):713–721, 1956.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. Meta-learning in neural networks: A survey. *IEEE TPAMI*, 2021. doi: 10.1109/TPAMI.2021.3079209.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

John Langford. Tutorial on practical prediction theory for classification. *JMLR*, 6:273–306, 2005.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *NeurIPS*, pp. 423–430, 2002.

Yann Lecun, Leon Bottou, Y. Bengio, and Patrick Haffner. Gradient-Based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278 – 2324, 1998.

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28, 2013.

Hao Liu, Richard Socher, and Caiming Xiong. Taming MAML: efficient unbiased meta-reinforcement learning. In *ICML*, pp. 4061–4071, 2019.

Tianyu Liu, Jie Lu, Zheng Yan, and Guangquan Zhang. PAC-Bayes bounds for meta-learning with data-dependent prior. arXiv:2102.03748, 2021.

Andreas Maurer. A note on the PAC-Bayesian theorem. arXiv:cs/0411099, 2004.

Andreas Maurer. Algorithmic stability and meta-learning. *JMLR*, 6:967–994, 2005.

Andreas Maurer. Transfer bounds for linear feature learning. *Machine Learning*, 75(3):327–350, 2009.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *JMLR*, 17:81:1–81:32, 2016.

David A. McAllester. Some PAC-Bayesian theorems. In *COLT*, pp. 230–234, 1998.

David A. McAllester. PAC-Bayesian model averaging. In *COLT*, pp. 164–170, 1999.

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

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *NeurIPS*, pp. 5947–5956, 2017.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *ICLR*, 2018.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-Bayes bounds with data dependent priors. *JMLR*, 13:3507–3531, 2012.

Anastasia Pentina and Christoph H. Lampert. A PAC-Bayesian bound for lifelong learning. In *ICML*, pp. 991–999, 2014.

Anastasia Pentina and Christoph H. Lampert. Lifelong learning with non-i.i.d. tasks. In *NeurIPS*, pp. 1540–1548, 2015.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *JMLR*, 22(227):1–40, 2021.

Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. *JMLR*, 11:1927–1956, 2010.

David Reeb, Andreas Doerr, Sebastian Gerwinn, and Barbara Rakitsch. Learning Gaussian processes by minimizing PAC-Bayesian generalization bounds. In *NeurIPS*, pp. 3341–3351, 2018.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. In *NeurIPS*, pp. 16833–16845, 2020.

Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, and Andreas Krause. PACOH: Bayes-optimal meta-learning with PAC-guarantees. In *ICML*, pp. 9116–9126, 2021.

Matthias W. Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *JMLR*, 3: 233–269, 2002.

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. *IEEE Trans. Inf. Theory*, 58(12):7086–7093, 2012.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, pp. 4077–4087, 2017.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In *ALT*, pp. 466–492, 2017.

Sebastian Thrun and Lorien Pratt. *Learning to Learn*. Kluwer Academic Publishers, 1998.

Nilesh Tripuraneni, Michael I. Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. In *NeurIPS*, pp. 7852–7862, 2020.

James Yeh. *Real Analysis: Theory of Measure and Integration*. World Scientific, 2014.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, 2021.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

APPENDIX

Appendix A Explicit Form of Different PAC-Bayes Bounds for Meta-Learning

We provide more details about different PAC-Bayes meta-learning bounds. The explicit form of these bounds can be found in Table 3, which is the detailed version of Table 1 in the main paper.

Table 3: Explicit forms of different PAC-Bayes bounds on $er(\mathcal{Q})$. **Meta-Learning Bound = Empirical Error + Environment-Level Complexity + Task-Level Complexity**. n is the number of observed tasks, m is the number of samples in S_i ($i \in [n]$). $\mathcal{P}, \mathcal{Q} \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))$ are hyperprior and hyperposterior respectively. $P, Q_i = Q(S_i, P) \in \mathcal{M}_1(\mathcal{H})$ are the prior and the posterior for the i-th training task. $\delta \in (0,1)$ is the confidence level. In MLAP-S bound, $\Delta_i = \mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}} \mathrm{KL}(Q_i||P)$. In our quadratic meta-learning bound, $\Delta = \mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}} \sum_{i=1}^n \mathrm{KL}(Q_i||P)$. In λ bounds, $\lambda \in (0,2)$.

Different Bounds	**Empirical Error**	**Environment-Level Complexity**	**Task-Level Complexity**
(Pentina & Lampert, 2014)	$\widehat{er}(\mathcal{Q})$	$\frac{1}{\sqrt{n}}\big(\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P}) + \frac{1}{8} + \ln\frac{\delta}{2}\big)$	$\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\sum_{i=1}^n \mathbf{E}_{P\sim\mathcal{Q}}\mathrm{KL}(Q_i\vert\vert P)}{n\sqrt{m}} + \frac{1}{8\sqrt{m}} + \frac{1}{n\sqrt{m}}\ln\frac{2}{\delta}$
MLAP-M (Amit & Meir, 2018)	$\widehat{er}(\mathcal{Q})$	$\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\ln\frac{2n}{\delta}}{2(n-1)}}$	$\frac{1}{n}\sum_{i=1}^n \sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\mathrm{KL}(Q_i\vert\vert P)+\ln\frac{2nm}{\delta}}{2(m-1)}}$
MLAP-S (Amit & Meir, 2018)	$\widehat{er}(\mathcal{Q})$	$\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\ln\frac{2n}{\delta}}{2(n-1)}}$	$\frac{2}{n}\sum_{i=1}^n [\frac{\Delta_i+\ln\frac{4n\sqrt{m}}{\delta}}{m} + \sqrt{\frac{\Delta_i+\ln\frac{4n\sqrt{m}}{\delta}}{2m}\widehat{er}(Q_i,S_i)}]$
PACOH (Rothfuss et al., 2021)	$\widehat{er}(\mathcal{Q})$	$\frac{1}{\sqrt{n}}\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P}) + \frac{1}{8\sqrt{n}}$	$\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\sum_{i=1}^n \mathbf{E}_{P\sim\mathcal{Q}}\mathrm{KL}(Q_i\vert\vert P)}{n\sqrt{m}} + \big(\frac{1}{8n\sqrt{m}} + \frac{1}{\sqrt{n}}\ln\frac{1}{\delta}\big)$
λ bound (Liu et al., 2021)	$\frac{\widehat{er}(\mathcal{Q})}{1-\lambda/2}$	$\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\ln\frac{2n}{\delta}}{2(n-1)}}$	$\frac{1}{n}\sum_{i=1}^n \frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\mathrm{KL}(Q_i\vert\vert P)+\ln\frac{4n\sqrt{m}}{\delta}}{m\lambda(1-\lambda/2)}$
classic bound (ours)	$\widehat{er}(\mathcal{Q})$	$\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\ln\frac{2\sqrt{n}}{\delta}}{2n}}$	$\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^n \mathrm{KL}(Q_i\vert\vert P)+\ln\frac{2\sqrt{nm}}{\delta}}{2nm}}$
quadratic bound (ours)	$\widehat{er}(\mathcal{Q})$	$\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\ln\frac{2\sqrt{n}}{\delta}}{2n}}$	$\frac{\Delta+\ln\frac{4\sqrt{nm}}{\delta}}{nm} + 2\sqrt{\widehat{er}(\mathcal{Q}) + \frac{\Delta+\ln\frac{4\sqrt{nm}}{\delta}}{2nm}}\sqrt{\frac{\Delta+\ln\frac{4\sqrt{nm}}{\delta}}{2nm}}$
λ bound (ours)	$\frac{\widehat{er}(\mathcal{Q})}{1-\lambda/2}$	$\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\ln\frac{2\sqrt{n}}{\delta}}{2n}}$	$\frac{\mathrm{KL}(\mathcal{Q}\vert\vert\mathcal{P})+\mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^n \mathrm{KL}(Q_i\vert\vert P)+\ln\frac{4\sqrt{nm}}{\delta}}{nm\lambda(1-\lambda/2)}$

Appendix B Proof of Our Theoretical Results

B.1 Proof of Generalized PAC-Bayes-kl Bound from i.i.d. Setting to Independently but Non-identically Distributed Meta-Learning Setting

We first give the proof of Lemma 1 in the main paper. Actually, the proof is proceeded with almost the same sequence of arguments as that of (Lever et al., 2013, Theorem 1) for i.i.d. case, with the only difference being that the focused samples in Lemma 1 are independent but non-identically distributed. We include its proof just for the sake of the completeness.

Proof Lemma 1. Using Fubini's theorem to exchange $\mathbf{E}_{f\sim\pi}$ and $\mathbf{E}_{\mathcal{S}}$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have

$$\begin{aligned}
&\ln \frac{\mathbf{E}_{\mathcal{S}}\mathbf{E}_{f\sim\pi} e^{t\Phi(R(f),r(f))}}{\delta} \\
\geq &\ln \mathbf{E}_{f\sim\pi} e^{t\Phi(R(f),r(f))} \qquad \text{(Markov)} \\
= &\ln \int_{\mathcal{F}} e^{t\Phi(R(f),r(f))} \frac{\mathrm{d}\pi}{\mathrm{d}\rho} \mathrm{d}\rho \qquad \text{(change of measure)} \\
\geq &\int_{\mathcal{F}} \big(\ln e^{t\Phi(R(f),r(f))} + \ln\frac{\mathrm{d}\pi}{\mathrm{d}\rho}\big)\mathrm{d}\rho \qquad \text{(Jensen)} \\
= &t\mathbf{E}_\rho \Phi(R(f),r(f)) - \mathrm{KL}(\rho||\pi) \\
\geq &t\Phi(\mathbf{E}_\rho R(f), \mathbf{E}_\rho r(f)) - \mathrm{KL}(\rho||\pi). \qquad \text{(Jensen)} \qquad \blacksquare
\end{aligned}$$

Next we need the following Lemmas 3-4 to prove our proposed Lemma 2 in the main paper.

Lemma 3 *(Seldin et al., 2012, Lemma 1) Let $X_1, ..., X_K$ be a sequence of independent random variables, such that $X_k \in [0,1]$ almost surely and $\mathbf{E}X_k = \mu_k$, for $k = 1, ..., K$. Let $Y_1, ..., Y_K$ be independent Bernoulli random variables such that $\mathbf{E}Y_k = \mu_k$. Then for any convex function $g : [0,1]^K \to \mathbb{R}$, we have $\mathbf{E}g(X_1, ..., X_K) \leq \mathbf{E}g(Y_1, ..., Y_K)$.*

Lemma 4 *(Hoeffding, 1956, Theorem 3) Let $Y = \sum_{k=1}^{K} Y_k$, where Y_k is a Bernoulli random variable with $\mathbf{E}Y_k = \mu_k$. Let $\bar{Y} \sim Bin(K, \bar{\mu})$ be a Binomial random variable with $\bar{\mu} = \frac{1}{K}\sum_{k=1}^{K} \mu_k$. Then for any strictly convex function $g : [K] \to \mathbb{R}$, $\mathbf{E}g(Y) \leq \mathbf{E}g(\bar{Y})$.*

Based on Lemma 3-4, we can obtain a useful corollary that bound the the sum of independent $[0, 1]$-valued random variables with the sum of i.i.d. Bernoulli random variables as follow.

Corollary 1 *Let $X_1, ..., X_K$ be a sequence of independent random variables with $X_k \in [0, 1]$ almost surely (a.s.) and $\mathbf{E}X_k = \mu_k$, $\forall k = 1, ..., K$. Let $\bar{Y}_1, ..., \bar{Y}_K$ be i.i.d. Bernoulli random variables with $\mathbf{E}\bar{Y}_k = \frac{1}{K}\sum_{k=1}^{K} \mu_k$. Let $X = \sum_{k=1}^{K} X_k, \bar{Y} = \sum_{k=1}^{K} \bar{Y}_k$. Then for any strictly convex function g:*

$$\mathbf{E}g(X) \leq \mathbf{E}g(\bar{Y}).$$

Proof. Let f be an affine function on $[0, 1]^K$, such that $f(X_1, ..., X_K) = \sum_{k=1}^{K} X_k$. Then $g(X) = g \circ f(X_1, ..., X_K)$ can be considered as the composition of the affine function f and the strictly convex function g. Hence $g \circ f$ is a convex function on $[0, 1]^K$. Then according to Lemma 3, there exists independent Bernoulli random variables $\{Y_k\}_{k=1}^{K}$ with $\mathbf{E}Y_k = \mu_k$, such that

$$\mathbf{E}g(X) = \mathbf{E}g \circ f(X_1, ..., X_K) \leq \mathbf{E}g \circ f(Y_1, ..., Y_K) = \mathbf{E}g(Y).$$

Finally, according to Lemma 4, for any strictly convex function g, we can derive the following result,

$$\mathbf{E}g(Y) \leq \mathbf{E}g(\bar{Y}). \quad \blacksquare$$

Corollary 2 *(Lemma 2 in the main paper) In the setting of Corollary 1, we have*

$$\mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{1}{K}\sum_{k=1}^{K} X_k||\bar{\mu})} \leq \mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{1}{K}\sum_{k=1}^{K} \bar{Y}_k||\bar{\mu})}.$$

Proof. Recall that $\mathrm{kl}(p||q)$ is a strictly convex function with respect to p (i.e., the second-order derivative $\frac{\mathrm{d}^2\,\mathrm{kl}(p||q)}{\mathrm{d}p^2} = \frac{1}{p} + \frac{1}{1-p} > 0, \forall p \in (0, 1)$), and exp is a strictly-increasing convex function, hence $\exp\{\lambda\,\mathrm{kl}(p||q)\}$ $(\forall \lambda > 0)$ is a strictly convex function with respect to p. Therefore $\exp\{K\,\mathrm{kl}(\frac{x}{K}||\bar{\mu})\}$ is a strictly convex function w.r.t. x (just show that the second derivative is positive). Combining the above discussion with Corollary 1 finishes the proof. $\blacksquare$

Remark 1 ***(the technical difficulty when obtaining our Corollary 2)***
*A previous result that is similar to our Corollary 2 is the Theorem 1 in (Maurer, 2004), which states that: for **i.i.d.** random variables $X_1, ..., X_K$ with $X_k \in [0, 1]$ a.s. and $\mathbf{E}X_k = \mu$, define the i.i.d. Bernoulli random variables $X'_1, ..., X'_K$ with $\mathbf{E}X'_k = \mu$, then we have $\mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{1}{K}\sum_{k=1}^{K} X_k||\mu)} \leq \mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{1}{K}\sum_{k=1}^{K} X'_k||\mu)}$. It can be seen that our derived Corollary 2 is actually the generalized version of Theorem 1 in (Maurer, 2004) by replacing the '**i.i.d.** $\boldsymbol{X_1, ..., X_K}$' condition with a weaker condition '**independent** $\boldsymbol{X_1, ..., X_K}$'. To core step to obtain such extension is the use of Lemma 4 (i.e. (Hoeffding, 1956, Theorem 3)) to reduce the **independent setting** into the **i.i.d. setting**. Such extension is truly useful when dealing with independent but non-identically distributed setting (e.g. meta-learning setting).*

Now we are ready to employ the above Corollary 2 to derive our generalized PAC-Bayes-kl bound for independent random variables, leading to the following proof of Theorem 1 in the main paper.

Proof of Theorem 1. For any fixed $f \in \mathcal{F}$, let $\{\eta_k\}_{k=1}^{K}$ be i.i.d. Bernoulli random variables with $\mathbf{E}\eta_k = \frac{1}{K}\sum_{k=1}^{K} \mathbf{E}_{\xi_k} g_k(f, \xi_k)$. Then we have, for any fixed $f \in \mathcal{F}$,

$$\begin{aligned}
\mathbf{E}_{\mathcal{S}}\mathrm{e}^{K\,\mathrm{kl}(r(f)||R(f))} &= \mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{1}{K}\sum_{k=1}^{K} g_k(f,\xi_k)||\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\eta_k)} \\
&\leq \mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{1}{K}\sum_{k=1}^{K} \eta_k||\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\eta_k)} \qquad \text{(Corollary 2)} \\
&\leq 2\sqrt{K},
\end{aligned}$$

where the last inequality holds due to the fact that for a binomial random variable $\eta = \sum_{k=1}^{K} \eta_k \sim Bin(K, u)$ with K trials and success μ for each trial, by (Maurer, 2004, Thm 1) we have

$$\mathbf{E}\mathrm{e}^{K\,\mathrm{kl}(\frac{\eta}{K},\mu)} = \sum_{k=0}^{K} \tbinom{K}{k}(\frac{k}{K})^k(\frac{K-k}{K})^{K-k} \qquad \in [\sqrt{K}, 2\sqrt{K}].$$

Notice that $\mathrm{kl}(p||q)$ is also convex with respect to the pair (p, q), thus we can set the function Φ in Lemma 1 as the binary relative entropy kl, and set $t = K$ to finish the proof:

$$\begin{aligned}\mathrm{kl}(\mathbf{E}_{f\sim\rho} r(f)||\mathbf{E}_{f\sim\rho} R(f)) \leq & \frac{1}{K}[\mathrm{KL}(\rho||\pi) + \ln \mathbf{E}_{f\sim\pi}\mathbf{E}_{\mathcal{S}} \mathrm{e}^{K\,\mathrm{kl}(r(f)||R(f))}/\delta] \\ \leq & \frac{\mathrm{KL}(\rho||\pi)}{K} + \frac{\ln(2\sqrt{K}/\delta)}{K}. \quad \blacksquare\end{aligned}$$

B.2 Proof of Three Improved PAC-Bayes Bounds for Meta-Learning Based on Extended PAC-Bayes-kl Bound

Now we first give three relaxations of the generalized PAC-Bayes-kl bound derived in our Theorem 1, based on the Pinsker's inequality. We borrow the relaxation techniques from (McAllester, 1999; Pérez-Ortiz et al., 2021; Thiemann et al., 2017) and hence generalize their PAC-Bayes bounds (i.e. PAC-Bayes-classic/quadratic/λ bound) from i.i.d. setting to independent setting.

Corollary 3 *In the setting of Theorem 1, the following inequality holds with probability at least $1 - \delta$ over the draw of sample $\mathcal{S}$ for any measure ρ over $\mathcal{F}$, we have*
(i) PAC-Bayes-classic bound:

$$|\mathbf{E}_{f\sim\rho} r(f) - \mathbf{E}_{f\sim\rho} R(f)| \leq \sqrt{\frac{\mathrm{KL}(\rho||\pi) + \ln(2\sqrt{K}/\delta)}{2K}}.$$

In particular, if $\mathbf{E}_{f\sim\rho} r(f) < \mathbf{E}_{f\sim\rho} R(f)$, we can obtain two sharper PAC-Bayes inequalities,
(ii) PAC-Bayes-quadratic bound:

$$\mathbf{E}_{f\sim\rho} R(f) \leq \Big(\sqrt{\mathbf{E}_{f\sim\rho} r(f) + \frac{\mathrm{KL}(\rho||\pi) + \ln(2\sqrt{K}/\delta)}{2K}} + \sqrt{\frac{\mathrm{KL}(\rho||\pi) + \ln(2\sqrt{K}/\delta)}{2K}}\Big)^2.$$

(iii) PAC-Bayes-λ bound:

$$\forall\lambda \in (0, 2), \quad \mathbf{E}_{f\sim\rho} R(f) \leq \frac{\mathbf{E}_{f\sim\rho} r(f)}{1 - \lambda/2} + \frac{\mathrm{KL}(\rho||\pi) + \ln(2\sqrt{K}/\delta)}{K\lambda(1 - \lambda/2)}.$$

***Proof*.** (i) For the PAC-Bayes-classic bound, directly using Pinsker's inequality $\mathrm{kl}(p||q) \geq 2(p - q)^2$ finishes the proof. Alternatively, we can also obtain a one-sided inequality for $\mathbf{E}_{f\sim\rho} r(f) - \mathbf{E}_{f\sim\rho} R(f)$ (or $\mathbf{E}_{f\sim\rho} R(f) - \mathbf{E}_{f\sim\rho} r(f)$) by replacing $\frac{2}{\delta}$ in the RHS of two-sided bound with $\frac{1}{\delta}$. $\blacksquare$
(ii) For the PAC-Bayes-quadratic bound, note that if $q > p$, then $\mathrm{kl}(p||q) \geq \frac{(p-q)^2}{2q}$. Therefore, if there exists a constant $\theta > 0$, such that $\mathrm{kl}(p||q) \leq \theta$, then we have $\frac{(p-q)^2}{2q} \leq \theta$, which leads to the below inequality:

$$q - p \leq \sqrt{2q\theta}. \tag{7}$$

Note that Eq. (7) can be considered as the quadratic inequality on $\sqrt{q}$. Solving this inequality results in the following inequality,

$$q \leq (\sqrt{\frac{\theta}{2} + p} + \sqrt{\frac{\theta}{2}})^2,$$

which finally gives the PAC-Bayes-quadratic bound if we set $q = \mathbf{E}_{f\sim\rho} R(f)$, $p = \mathbf{E}_{f\sim\rho} r(f)$, and $\theta = \frac{\mathrm{KL}(\rho||\pi) + \ln(2\sqrt{K}/\delta)}{K}$. $\blacksquare$
(iii) For the PAC-Bayes-λ bound, note that if $q > p$, then $\mathrm{kl}(p||q) \geq \frac{(p-q)^2}{2q}$. Therefore, if there exists a constant $\theta > 0$, such that $\mathrm{kl}(p||q) \leq \theta$, then we have $\frac{(p-q)^2}{2q} \leq \theta$. We then have for any positive $\lambda > 0$, $|p - q| \leq \sqrt{2q\theta} \leq \frac{1}{2}(\lambda q + \frac{2\theta}{\lambda})$. Thus we have

$$-\frac{\lambda q}{2} - \frac{\theta}{\lambda} \leq p - q \leq \frac{\lambda q}{2} + \frac{\theta}{\lambda}. \tag{8}$$

Using just the above left-hand-side inequality, with simple rearrangement we finally obtain that $q(1 - \frac{\lambda}{2}) \leq p + \frac{\theta}{\lambda}$, thus for any $\lambda \in (0, 2)$:

$$q \leq \frac{p}{1 - \lambda/2} + \frac{\theta}{\lambda(1 - \lambda/2)}.$$

Combining the above results with the first assertion, we thus have with probability at least $1-\delta$ over the draw of $\mathcal{S}$,

$$\mathbf{E}_{f\sim\rho}R(f) \leq \frac{\mathbf{E}_{f\sim\rho}r(f)}{1-\lambda/2} + \frac{\mathrm{KL}(\rho||\pi) + \ln(2\sqrt{K}/\delta)}{K\lambda(1-\lambda/2)}. \qquad \blacksquare$$

Next we focus on giving the explicit upper bounds on the deviations $er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})$ and $\widetilde{er}(\mathcal{Q}) - \widehat{er}(\mathcal{Q})$. First, we use Corollary 3(i) to give an explicit bound on $|er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})|$ as follow.

Proposition 1 *For any $\delta \in (0,1)$, any pre-defined hyperprior $\mathcal{P}$, with probability at least $1-\delta$ over the draw of n distributions $\{D_i\}_{i=1}^n$, we have for any hyperposterior $\mathcal{Q}$,*

$$|er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})| \leq \sqrt{\frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \ln\frac{2\sqrt{n}}{\delta}}{2n}}.$$

Proof. Notice that we can rewrite $er(\mathcal{Q}), \widetilde{er}(\mathcal{Q})$ as the following form:

$$er(\mathcal{Q}) = \mathbf{E}_{P\sim\mathcal{Q}}\mathbf{E}_{(D,S)\sim\tau\times D^m} er(Q(S,P),D), \qquad \widetilde{er}(\mathcal{Q}) = \mathbf{E}_{P\sim\mathcal{Q}}\frac{1}{n}\sum_{i=1}^{n} er(Q(S_i,P),D_i).$$

Recall Theorem 1, we thus set $K = n$, $f = P$, set the reference measure $\pi = \mathcal{P}$, the posterior measure $\rho = \mathcal{Q}$, the observation $\xi_k = (D_i, S_i)$, and set the function $g_k(f,\xi_k) = \mathbf{E}_{h\sim Q(S_i,P)}\mathbf{E}_{z\sim D_i} l(h,z) \in [0,1]$. Then we can give an upper bound on the relative entropy $\mathrm{kl}(\widetilde{er}(\mathcal{Q})||er(\mathcal{Q}))$,

$$\mathrm{kl}(\widetilde{er}(\mathcal{Q})||er(\mathcal{Q})) \leq \frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \ln\frac{2\sqrt{n}}{\delta}}{n}.$$

With Pinsker's inequality $\mathrm{kl}(p,q) \geq 2(p-q)^2$,

$$|er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})| \leq \sqrt{\frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \ln\frac{2\sqrt{n}}{\delta}}{2n}}. \qquad \blacksquare$$

Further, we can use Corollary 3(i)-(iii) to give an explicit bound on $\widetilde{er}(\mathcal{Q})$ in Propositions 2-4.

Proposition 2 ***(PAC-Bayes-classic bound for*** $\widetilde{er}(\mathcal{Q})$***)*** *For any predefined hyperprior $\mathcal{P}$, any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draw of the training sample $S = \{S_i\}_{i=1}^n$, for any hyperposterior $\mathcal{Q}$ we have,*

$$|\widetilde{er}(\mathcal{Q}) - \widehat{er}(\mathcal{Q})| \leq \sqrt{\frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^n \mathrm{KL}(Q_i||P) + \ln\frac{2\sqrt{nm}}{\delta}}{2nm}}.$$

Proof. Notice that we can rewrite $\widetilde{er}(\mathcal{Q}), \widehat{er}(\mathcal{Q})$ as the following form:

$$\widetilde{er}(\mathcal{Q}) = \mathbf{E}_{P\sim\mathcal{Q}}\mathbf{E}_{(h_1,\cdots,h_n)\sim Q_1\times\cdots\times Q_n}\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_{z\sim D_i} l(h_i,z),$$

$$\widehat{er}(\mathcal{Q}) = \mathbf{E}_{P\sim\mathcal{Q}}\mathbf{E}_{(h_1,\cdots,h_n)\sim Q_1\times\cdots\times Q_n}\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} l(h_i,z_{ij}).$$

Recall Theorem 1, we can set $f = (P, h_1, ..., h_n)$, the reference measure $\pi = \mathcal{P}\times P^n$, the posterior measure $\rho = \mathcal{Q}\times\prod_{i=1}^n Q_i$, where $Q_i = Q(S_i,P)$, and set the observations $\xi_k = z_{ij}$, $g_k(f,\xi_k) = l(h_i, z_{ij})$. Then using Corollary 3(i), with probability $\geq 1-\delta$ we have,

$$|\widetilde{er}(\mathcal{Q}) - \widehat{er}(\mathcal{Q})| \leq \sqrt{\frac{\mathrm{KL}(\mathcal{Q}\times\prod_i^n Q_i||\mathcal{P}\times P^n) + \ln\frac{2\sqrt{nm}}{\delta}}{2nm}}.$$

Furthermore, notice that $\mathrm{KL}(\mathcal{Q}\times\prod_i^n Q_i||\mathcal{P}\times P^n) = \mathbf{E}_{\mathcal{Q}\times\prod_{i=1}^n Q_i}\ln\frac{\mathrm{d}(\mathcal{Q}\times\prod_{i=1}^n Q_i)}{\mathrm{d}(\mathcal{P}\times P^n)} = \mathbf{E}_{\mathcal{Q}\times\prod_{i=1}^n Q_i}\big(\ln\frac{\mathrm{d}\mathcal{Q}}{\mathrm{d}\mathcal{P}} + \sum_{i=1}^n \ln\frac{\mathrm{d}Q_i}{\mathrm{d}P}\big) = \mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}}\sum_{i=1}^n \mathrm{KL}(Q_i||P)$, which completes the whole proof. $\blacksquare$

Proposition 3 ***(PAC-Bayes-quadratic bound for $\widetilde{er}(\mathcal{Q})$)*** *For any predefined hyperprior $\mathcal{P}$, any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draw of the training sample $S = \{S_i\}_{i=1}^n$, for any hyperposterior $\mathcal{Q}$ we have,*

$$\widetilde{er}(\mathcal{Q}) \le \Big(\sqrt{\widehat{er}(\mathcal{Q}) + \frac{\Delta}{2nm}} + \sqrt{\frac{\Delta}{2nm}}\Big)^2$$

where $\Delta = \mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}} \sum_{i=1}^n \mathrm{KL}(Q_i||P) + \ln \frac{2\sqrt{nm}}{\delta}$.

Proof. The main step is to use the PAC-Bayes-quadratic bound in Corollary 3(ii) to bound $\widetilde{er}(\mathcal{Q})$. The rest proof is similar to that of Proposition 2 and is left to readers. ■

Proposition 4 ***(PAC-Bayes-λ bound for $\widetilde{er}(\mathcal{Q})$)*** *For any predefined hyperprior $\mathcal{P}$, any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draw of the training sample $S = \{S_i\}_{i=1}^n$, for any hyperposterior $\mathcal{Q}$ and any $\lambda \in (0,2)$, we have,*

$$\widetilde{er}(\mathcal{Q}) \le \frac{\widehat{er}(\mathcal{Q})}{1-\lambda/2} + \frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \mathbf{E}_{P\sim\mathcal{Q}} \sum_{i=1}^n \mathrm{KL}(Q_i||P) + \ln \frac{2\sqrt{nm}}{\delta}}{nm\lambda(1-\lambda/2)}.$$

Proof. The main step is to use the PAC-Bayes-λ bound in Corollary 3 (iii) to bound $\widetilde{er}(\mathcal{Q})$. The rest proof is similar to that of Proposition 2. ■

Proof of Theorem 2. The main idea is to give bounds on $\mathrm{kl}(\widetilde{er}(\mathcal{Q})||er(\mathcal{Q}))$ and $\mathrm{kl}(\widehat{er}(\mathcal{Q})||\widetilde{er}(\mathcal{Q}))$ respectively, and then combine them with union bound to give the explicit upper bound on $er(\mathcal{Q})$. (1) To bound $\mathrm{kl}(\widetilde{er}(\mathcal{Q})||er(\mathcal{Q}))$, since we have no idea whether $\widetilde{er}(\mathcal{Q}) < er(\mathcal{Q})$ or not, we can not use the quadratic bound or λ bound. What we can use is just the classic bound given in Proposition 1. Then we can derive a one-sided bound on $er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})$ with $\sqrt{\frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \ln\frac{\sqrt{n}}{\delta}}{2n}}$. (2) To bound $\mathrm{kl}(\widehat{er}(\mathcal{Q})||\widetilde{er}(\mathcal{Q}))$, since the empirical multi-task risk $\widehat{er}(\mathcal{Q})$ is strictly smaller than the expected multi-task risk, we hence can use the PAC-Bayes-quadratic bound in Proposition 3 and the PAC-Bayes-λ bound in Proposition 4 to bound $\widetilde{er}(\mathcal{Q})$. Then combining the upper bound on $\widetilde{er}(\mathcal{Q})$ with the one-sided bound on $er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})$ in discussion (1), we can obtain the PAC-Bayes-quadratic bound or λ bound for meta-learning. For the PAC-Bayes-classic bound, note that we can obtain a one-sided bound for $\widetilde{er}(\mathcal{Q}) - \widehat{er}(\mathcal{Q})$ by replacing $\frac{2}{\delta}$ with $\frac{1}{\delta}$ in the bound of Proposition 2. Thus, combining the one-sided bound on $\widetilde{er}(\mathcal{Q}) - \widehat{er}(\mathcal{Q})$ with the one-sided bound on $er(\mathcal{Q}) - \widetilde{er}(\mathcal{Q})$ in discussion (1) finishes the proof. ■

B.3 Proof of PAC-Bayes-kl Bound on Novel Task with the Prior from the Learned Hyperposterior

B.3.1 Proof of the Existence of the Intermediate Measure ν in Eq. (6)

In this section, we first give a proof of the existence of the intermediate measure ν defined in Eq. (6). Therefore, we can also guarantee the existence of the Gibbs meta-learned prior and the Gibbs posterior defined in Eq. (6). We first give a basic lemma in measure theory as follow.

Lemma 5 *(Yeh, 2014, Page 255, Prob.11.8) Given a measurable space $(X, \mathcal{B})$. Let μ be a σ-finite positive measure on $(X, \mathcal{B})$ and let f be a μ-integrable nonnegative extended real-valued $\mathcal{B}$-measurable function on X such that $f \neq 0$, μ-a.e. on X. Define a set function ν on $\mathcal{B}$ by setting*

$$\nu(E) = \int_E f \mathrm{d}\mu, \forall E \in \mathcal{B}.$$

(a) Show that ν is a σ-finite positive measure on $(X, \mathcal{B})$.
(b) Show that $\nu << \mu$ and $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ exists, and $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} = f$, μ-a.e. on X.
(c) Show that $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ exists and $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} = \frac{1}{f}$ μ- and ν-a.e. on X.

Now we can prove how to construct the measure $\nu \in \mathcal{M}_1(\mathcal{H})$ defined in Eq. (6) of main paper as follow. The core step is to show that the normalization constant $\beta = \int_{\mathcal{H}} \exp\{\frac{\gamma}{n}\sum_{i=1}^{n} \mathbf{E}_{z\sim D_i} l(h,z)\}\mathrm{d}P$ is well-defined (i.e., finite), such that the measure ν is well-defined. The finiteness of β holds due to the fact that the loss function l is a bounded function. The existence of the measure $\nu \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{H}))$ defined above Eq. (5) can be proved in a similar way.

Claim 1 *There exists a probability measure $\nu \in \mathcal{M}_1(\mathcal{H})$ that satisfies the conditions in Eq. (6).*

Proof. Define a nonnegative function $f : \mathcal{H} \to \mathbb{R}^+$ as $f(h) = \frac{1}{\beta}\exp\{\frac{\gamma}{n}\sum_{i=1}^{n} \mathbf{E}_{z\sim D_i} l(h,z)\}$, where $\beta = \int_{\mathcal{H}} \exp\{\frac{\gamma}{n}\sum_{i=1}^{n} \mathbf{E}_{z\sim D_i} l(h,z)\}\mathrm{d}P$ is a normalization constant. Note that the loss function l has range $[0,1]$, hence $\beta \le \int_{\mathcal{H}} \exp\{\gamma\}\mathrm{d}P = \exp\{\gamma\} < \infty$ and hence the density function is well-defined. We then construct a set function over $\mathcal{B}(\mathcal{H})$, as defined by: $\nu(E) = \int_E f\mathrm{d}P, \forall E \in \mathcal{B}(\mathcal{H})$, where $\mathcal{B}(\mathcal{H})$ is a σ-algebra over the space $\mathcal{H}$. Then according to the assertion (a) in Lemma 5, the set function ν is a σ-finite positive (probability) measure on the space $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$. According to the assertion (c) in Lemma 5, the measure P is absolutely continuous with respect to the measure ν (i.e., $P << \nu$), so the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}\nu} = \frac{1}{f} = \exp\{-\frac{\gamma}{n}\sum_{i=1}^{n} \mathbf{E}_{z\sim D_i} l(h,z)\}\beta$, and $\mathrm{d}P = \exp\{-\frac{\gamma}{n}\sum_{i=1}^{n} \mathbf{E}_{z\sim D_i} l(h,z)\}\beta\mathrm{d}\nu$. Therefore we have $1 = \int_{\mathcal{H}} 1\mathrm{d}P = \int_{\mathcal{H}} \beta\exp\{-\frac{\gamma}{n}\sum_{i=1}^{n} \mathbf{E}_{z\sim D_i} l(h,z)\}\mathrm{d}\nu$, then we obtain $\beta = 1/\int_{\mathcal{H}} \exp\{-\frac{\gamma}{n}\sum_{i=1}^{n} \mathbf{E}_{z\sim D_i} l(h,z)\}\mathrm{d}\nu$. Similarly, we can construct the (posterior) probability measure Q_i by defining the set function: $Q_i(E) = \int_E \exp\{-\frac{\gamma}{m_i}\sum_{j=1}^{m_i} l(h,z_{ij})\}\mathrm{d}\nu/\beta_i', \forall E \in \mathcal{B}(\mathcal{H})$, where $\beta_i' = \int_H \exp\{-\frac{\gamma}{m_i}\sum_{j=1}^{m_i} l(h,z_{ij})\}\mathrm{d}\nu$ is the normalization constant. ■

B.3.2 PROOF OF PAC-BAYES-KL BOUND ON NOVEL TASK WITH THE LOCALIZED GIBBS PRIOR FROM THE DIRAC HYPERPOSTERIOR

Next, we give a fundamental lemma which guarantees that the equality $\mathbf{E}_S \int f\mathrm{d}Q_S = \int f\mathrm{d}(\mathbf{E}_S Q_S)$ holds. Such result will be exactly applied to the proof of **Bounding II** for Theorem 3 in the next section. The proof of $\mathbf{E}_S \int f\mathrm{d}Q_S = \int f\mathrm{d}(\mathbf{E}_S Q_S)$ follows the steps of 'standard method' in measure theory (Yeh, 2014), but we have not found this result in the literature, hence we include the details.

Lemma 6 *Denote the posterior distribution $Q(S,P)$ by Q_S for brevity. Define the set function over the σ-algebra $\mathcal{B}(\mathcal{H})$ of the hypothesis space $\mathcal{H}$ as: $\forall A \in \mathcal{B}(\mathcal{H})$, $\big(\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S\big)(A) = \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S(A)$. Then we have following three assertions:*
(a) $\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S$ is a probability measure over $\mathcal{H}$.
(b) $\forall$ integrable function $f : \mathcal{H} \to \mathbb{R}$, $\mathop{\mathbf{E}}_{D\sim\tau}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{h\sim Q(S,P)} f(h) = \mathbf{E}_{h\sim\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m}Q_S} f(h)$.
(c) $\frac{\mathrm{d}\mathbf{E}_{D,S}Q_S}{\mathrm{d}P} = \mathbf{E}_{D,S}\frac{\mathrm{d}Q_S}{\mathrm{d}P}$.

Proof.
(a) For the first assertion, we have:
(i) $\big(\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S\big)(\mathcal{H}) = \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S(\mathcal{H}) = \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} 1 = 1.$
(ii)$\big(\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S\big)(\emptyset) = \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S(\emptyset) = 0.$
(iii) For any pairwise disjoint set $\{A_k\}_{k=1}^{\infty}$, where $A_k \in \mathcal{B}(\mathcal{H})$, we have

$$\begin{aligned}&\big(\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S\big)(\cup_k A_k) = \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S(\cup_k A_k)\\ =&\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m}\sum_{k=1}^{\infty} Q_S(A_k) = \sum_{k=1}^{\infty}\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S(A_k) = \sum_{k=1}^{\infty}\big(\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S\big)(A_k),\end{aligned}$$

where the exchange between $\sum$ and $\mathbf{E}$ holds due to Fubini's theorem. Therefore,$\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S$ is a probability measure over the hypothesis space $\mathcal{H}$. ■

(b) For assertion (b), we take standard method to demonstrate that the equality holds.
(b)(i) First consider the case where $f(h)$ is a simple function.
Let $f = \sum_{k=1}^{n} a_k \mathbf{1}_{A_k}, (a_k \ge 0)$, where $A_k \in \mathcal{B}(\mathcal{H})$, $\mathbf{1}_{A_k}$ is the indicator function defined over A_k.

Then the right hand side of the equality has

$$
\begin{aligned}
RHS =& \int_{\mathcal{H}} \sum_{k=1}^{n} a_k \mathbf{1}_{A_k} \mathrm{d}\big(\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S\big) \\
=& \sum_{k=1}^{n} a_k \big(\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S\big)(A_k) \quad \text{(assertion (a))} \\
=& \sum_{k=1}^{n} a_k \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S(A_k) \\
=& \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} \sum_{k=1}^{n} a_k Q_S(A_k) \\
=& \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m}\mathbf{E}_{h\sim Q_S} f(h) = LHS.
\end{aligned}
$$

(b)(ii) Now consider the case where $f(h)$ is a nonnegative function.
Then we can choose a series of nonnegative non-decreasing simple functions $\{f_k\}_{k=1}^{\infty} \uparrow f$ (i.e., $\lim_k f_k(h) = f(h)$, almost everywhere on $\mathcal{H}$). Using Levy's monotone convergence theorem,

$$
\begin{aligned}
RHS =& \int_{\mathcal{H}} \lim_k f_k \mathrm{d}\big(\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S\big) \\
=& \lim_k \int_{\mathcal{H}} f_k \mathrm{d}\big(\mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S\big) \\
=& \lim_k \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m}\mathbf{E}_{h\sim Q_S} f_k(h) \quad \text{((b)(i))} \\
=& \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m}\mathbf{E}_{h\sim Q_S} \lim_k f_k(h) \quad \text{(Levy)} \\
=& \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m}\mathbf{E}_{h\sim Q_S} f(h) = LHS,
\end{aligned}
$$

(b)(iii) Finally consider the case where $f(h)$ is an integrable function.
Decompose f as $f = f^+ - f^-$, where $f^+ = \max(f, 0), f^- = \max(-f, 0)$. From (b)(ii) we have,

$$
\mathop{\mathbf{E}}_{D\sim\tau} \mathop{\mathbf{E}}_{h\sim Q(S,P)} f^+(h) = \mathbf{E}_{h\sim \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S} f^+(h),
$$

$$
\mathop{\mathbf{E}}_{D\sim\tau} \mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim Q(S,P)} f^-(h) = \mathbf{E}_{h\sim \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S} f^-(h).
$$

Therefore, using the linearity property of the integral we can obtain,

$$
\begin{aligned}
LHS =& \mathop{\mathbf{E}}_{D\sim\tau} \mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim Q(S,P)} (f^+ - f^-)(h) \\
=& \mathop{\mathbf{E}}_{D\sim\tau} \mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim Q(S,P)} f^+(h) - \mathop{\mathbf{E}}_{D\sim\tau} \mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim Q(S,P)} f^-(h) \\
=& \mathbf{E}_{h\sim \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S} f^+(h) - \mathbf{E}_{h\sim \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S} f^-(h) \\
=& \mathbf{E}_{h\sim \mathbf{E}_{D\sim\tau}\mathbf{E}_{S\sim D^m} Q_S} (f^+ - f^-)(h) = RHS,
\end{aligned}
$$

which finishes the proof for assertion (b). ■

(c) For the last assertion, note that the Radon-Nikodym derivative $\frac{\mathrm{d}Q_S}{\mathrm{d}P} \geq 0$, then for any subset $E \subseteq \mathcal{H}$, we have

$$
\begin{aligned}
& \int_E (\mathbf{E}_{D,S} \frac{\mathrm{d}Q_S}{\mathrm{d}P}) \mathrm{d}P \\
=& \int_E \Big(\int_{\mathcal{M}_1(\mathcal{Z})} \int_{\mathcal{Z}^m} \frac{\mathrm{d}Q_S}{\mathrm{d}P} \mathrm{d}D^m \mathrm{d}\tau \Big) \mathrm{d}P \\
=& \int_{\mathcal{M}_1(\mathcal{Z})} \int_{\mathcal{Z}^m} \int_E \frac{\mathrm{d}Q_S}{\mathrm{d}P} \mathrm{d}P \mathrm{d}D^m \mathrm{d}\tau \quad \text{(Fubini)} \\
=& \int_{\mathcal{M}_1(\mathcal{Z})} \int_{\mathcal{Z}^m} \int_E \mathrm{d}Q_S \mathrm{d}D^m \mathrm{d}\tau \quad \text{(change of measure)} \\
=& (\mathbf{E}_{D,S} Q_S)(E).
\end{aligned}
$$

Therefore, we have that the Radon-Nikodym derivative of $\mathbf{E}_{D,S}Q_S$ with respect to P is $\frac{\mathrm{d}\mathbf{E}_{D,S}Q_S}{\mathrm{d}P} = \mathbf{E}_{D,S}\frac{\mathrm{d}Q_S}{\mathrm{d}P}$. Hence we finish the whole proof for assertion (c). ■

Proof of Theorem 3. Using the PAC-Bayes kl-bound in Theorem 1 for the i.i.d. setting, with probability at least $1-\delta$ we have,

$$\mathbf{E}_{D,S}\,\mathrm{kl}(\widehat{er}(Q,S)||er(Q,D)) \leq \frac{\mathbf{E}_{D,S}\,\mathrm{KL}(Q||P) + \ln{(2\sqrt{m}/\delta)}}{m}$$

Hence, we aim to bound the expectation of KL-divergence in the RHS of the above inequality. Actually, we can decompose this term into three parts and bound these parts separately, that is,

$$\begin{aligned}
&\mathbf{E}_{D\sim\tau,S\sim D^m}\,\mathrm{KL}(Q(S,P)||P) = \mathbf{E}_{D\sim\tau,S\sim D^m}\mathbf{E}_{h\sim Q(S,P)}\ln\frac{\mathrm{d}Q(S,P)}{\mathrm{d}P}\\
=&\mathbf{E}_{D,S}\mathbf{E}_{h\sim Q(S,P)}\ln\frac{\beta\exp\{-\frac{\gamma}{m}\sum_{j=1}^{m}l(h,z_j)\}}{\beta'\exp\{-\frac{\gamma}{n}\sum_{i=1}^{n}\underset{z\sim D_i}{\mathbf{E}}l(h,z)\}}\\
=&\mathbf{E}_{D,S}\mathbf{E}_{h\sim Q(S,P)}\ln\frac{\beta}{\beta'} + \mathbf{E}_{D,S}\mathbf{E}_{h\sim Q(S,P)}\big[\frac{\gamma}{n}\sum_{i=1}^{n}\mathbf{E}_{z\sim D_i}l(h,z) - \frac{\gamma}{m}\sum_{j=1}^{m}l(h,z_j)\big]\\
=&\mathbf{E}_{D,S}\mathbf{E}_{h\sim Q(S,P)}\ln\frac{\beta}{\beta'} + \mathbf{E}_{D,S}\big[\frac{\gamma}{n}\sum_{i=1}^{n}er(Q(S,P),D_i) - \gamma\widehat{er}(Q(S,P),S)\big]\\
=&\underbrace{\mathbf{E}_{D,S}\mathbf{E}_{h\sim Q(S,P)}\ln\frac{\beta}{\beta'}}_{\mathbf{I}} + \underbrace{\mathbf{E}_{D,S}\big[\frac{\gamma}{n}\sum_{i=1}^{n}er(Q(S,P),D_i) - \gamma er(Q(S,P),D)\big]}_{\mathbf{II}}\\
&+\underbrace{\mathbf{E}_{D,S}\gamma\big[er(Q(S,P),D) - \widehat{er}(Q(S,P),S)\big]}_{\mathbf{III}}
\end{aligned}$$

Bounding I.

$$\begin{aligned}
&\mathbf{E}_{D,S}\ln\frac{\beta}{\beta'} = -\mathbf{E}_{D,S}\ln\frac{\beta'}{\beta}\\
=&-\mathbf{E}_{D,S}\ln\int_{\mathcal{H}}\frac{\exp\{-\frac{\gamma}{m}\sum_{j=1}^{m}l(h,z_j)\}}{\exp\{-\frac{\gamma}{n}\sum_{i=1}^{n}\mathbf{E}_{z\sim D_i}l(h,z)\}/\widetilde{P}(h)}\mathrm{d}\nu\\
=&-\mathbf{E}_{D,S}\ln\int_{\mathcal{H}}\widetilde{P}(h)\exp\{-\frac{\gamma}{m}\sum_{j=1}^{m}l(h,z_j) + \frac{\gamma}{n}\sum_{i=1}^{n}\mathbf{E}_{z\sim D_i}l(h,z)\}\mathrm{d}\nu\\
\leq&\mathbf{E}_{D,S}\mathbf{E}_{h\sim P}\{\frac{\gamma}{m}\sum_{j=1}^{m}l(h,z_j) - \frac{\gamma}{n}\sum_{i=1}^{n}\mathbf{E}_{z\sim D_i}l(h,z)\} \quad \text{(Jensen Inequality of } -\ln\text{)}\\
=&\mathbf{E}_{D,S}\big[\gamma\widehat{er}(P,S) - \frac{\gamma}{n}\sum_{i=1}^{n}er(P,D_i)\big]\\
=&\mathbf{E}_{D}\big[\gamma er(P,D) - \frac{\gamma}{n}\sum_{i=1}^{n}er(P,D_i)\big]\\
=&\big[\gamma\mathbf{E}_{D\sim\tau}er(P,D) - \frac{\gamma}{n}\sum_{i=1}^{n}er(P,D_i)\big]\\
\leq&\gamma\sqrt{\frac{\ln{(\sqrt{n}/\delta)}}{2n}}, \quad \text{(set } \pi=\rho=\text{P in Corollary 3(i))}
\end{aligned}$$

where the last inequality holds with probability at least $1-\delta$.

Bounding II.

$$
\begin{aligned}
&\mathbf{E}_{D,S}\big[\frac{\gamma}{n}\sum_{i=1}^{n} er(Q(S,P),D_i)-\gamma er(Q(S,P),D)\big]\\
=&\frac{\gamma}{n}\sum_{i=1}^{n}\big[\mathbf{E}_{D,S}\mathbf{E}_{h\sim Q(S,P)}l(h,D_i)-\mathbf{E}_{D,S}\mathbf{E}_{h\sim Q(S,P)}l(h,D)\big]\\
=&\frac{\gamma}{n}\sum_{i=1}^{n}\big[\mathbf{E}_{h\sim \mathbf{E}_{D,S}Q(S,P)}l(h,D_i)-\mathbf{E}_{D}\mathbf{E}_{h\sim \mathbf{E}_{D,S}Q(S,P)}l(h,D)\big] \quad \text{(Lemma 6 (b))}\\
=&\frac{\gamma}{n}\sum_{i=1}^{n}\big[er(\mathbf{E}_{D,S}Q(S,P),D_i)-\mathbf{E}_{D}er(\mathbf{E}_{D,S}Q(S,P),D)\big]\\
\leq&\gamma\sqrt{\frac{\ln{(\sqrt{n}/\delta)}+\mathrm{KL}(\mathbf{E}_{D,S}Q(S,P)||P)}{2n}} \quad \text{(Corollary 3)}\\
\overset{(*)}{=}&\gamma\sqrt{\frac{\ln{(\sqrt{n}/\delta)}+\mathrm{KL}(\mathbf{E}_{D,S}\widetilde{Q}(S,P)||\widetilde{P})}{2n}} \quad \text{(Lemma 6 (c))}\\
\leq&\gamma\sqrt{\frac{\ln{(\sqrt{n}/\delta)}+\mathbf{E}_{D,S}\,\mathrm{KL}(\widetilde{Q}(S,P)||\widetilde{P})}{2n}} \quad (\text{Convexity of KL}-\text{divergence})\\
=&\gamma\sqrt{\frac{\ln{(\sqrt{n}/\delta)}+\mathbf{E}_{D,S}\,\mathrm{KL}(Q(S,P)||P)}{2n}}.
\end{aligned}
$$

For reader's benefit, we provide more explanations for equality $(*)$, where $\widetilde{Q}$, $\widetilde{P}$ are density functions of measures Q and P respectively, with respect to measure ν. Actually from Lemma 6 (c) we have,

$$
\frac{\mathrm{d}\mathbf{E}_{D,S}Q(S,P)}{\mathrm{d}\nu}=\mathbf{E}_{D,S}\frac{\mathrm{d}Q(S,P)}{\mathrm{d}\nu}=\mathbf{E}_{D,S}\widetilde{Q}(S,P).
$$

Then by reusing Lemma 6 (c), we have,

$$
\begin{aligned}
&\mathrm{KL}(\mathbf{E}_{D,S}Q(S,P)||P)\\
=&\mathbf{E}_{h\sim \mathbf{E}_{D,S}Q(S,P)}\ln\frac{\mathrm{d}\mathbf{E}_{D,S}Q(S,P)}{\mathrm{d}P}\\
=&\mathbf{E}_{h\sim \mathbf{E}_{D,S}Q(S,P)}\ln\mathbf{E}_{D,S}\frac{\mathrm{d}Q(S,P)}{\mathrm{d}P}\\
=&\mathbf{E}_{h\sim \mathbf{E}_{D,S}Q(S,P)}\ln\mathbf{E}_{D,S}\frac{\mathrm{d}Q(S,P)}{\mathrm{d}\nu}\big(\frac{\mathrm{d}P}{\mathrm{d}\nu}\big)^{-1}\\
=&\mathbf{E}_{h\sim \mathbf{E}_{D,S}\big[\widetilde{Q}(S,P)(h)\big]}\ln\mathbf{E}_{D,S}\frac{\widetilde{Q}(S,P)}{\widetilde{P}}\\
=&\,\mathrm{KL}(\mathbf{E}_{D,S}\widetilde{Q}(S,P)||\widetilde{P}).
\end{aligned}
$$

Bounding III.
Reusing the one-sided inequality in Corollary 3(i) and the Jensen's inequality of the concave square root function (i.e., $f(t)=\sqrt{t}$), we have with probability at least $1-\delta$,

$$
\begin{aligned}
\mathbf{E}_{D,S}\gamma\big[er(Q,D)-\widehat{er}(Q,S)\big]\leq&\gamma\mathbf{E}_{D,S}\sqrt{\frac{\ln{(\sqrt{m}/\delta)}+\mathrm{KL}(Q||P)}{2m}}\\
\leq&\gamma\sqrt{\frac{\ln{(\sqrt{m}/\delta)}+\mathbf{E}_{D,S}\,\mathrm{KL}(Q||P)}{2m}}.
\end{aligned}
$$

Bounding I + II + III.
Combining the above upper bounds I-III, and denoting $\mathbf{E}_{D,S}(\mathrm{KL}(Q(S,P)||P))$ by $\boldsymbol{\theta}$ for brevity, we have the following inequality,

$$
\boldsymbol{\theta}\leq\gamma\sqrt{\frac{\ln{(3\sqrt{n}/\delta)}}{2n}}+\gamma\sqrt{\frac{\ln{(3\sqrt{n}/\delta)}+\boldsymbol{\theta}}{2n}}+\gamma\sqrt{\frac{\ln{(3\sqrt{m}/\delta)}+\boldsymbol{\theta}}{2m}}
$$

If $\boldsymbol{\theta} \leq \gamma\sqrt{\frac{\ln{(3\sqrt{n}/\delta)}}{2n}}$, then we are done. Otherwise, using simple calculation and rearrangement, as well as the basic inequalities $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $\sqrt{a} \leq \frac{a+1}{2}$, we have

$$\begin{aligned}
&\boldsymbol{\theta}^2 + \gamma^2 \frac{\ln{(3\sqrt{n}/\delta)}}{2n} - 2\boldsymbol{\theta}\gamma\sqrt{\frac{\ln{(3\sqrt{n}/\delta)}}{2n}} \\
\leq&\gamma^2 \big(\sqrt{\frac{\ln{(3\sqrt{n}/\delta)} + \boldsymbol{\theta}}{2n}} + \sqrt{\frac{\ln{(3\sqrt{m}/\delta)} + \boldsymbol{\theta}}{2m}}\big)^2 \\
=&\gamma^2 \big(\frac{\ln{(3\sqrt{n}/\delta)} + \boldsymbol{\theta}}{2n} + \frac{\ln{(3\sqrt{m}/\delta)} + \boldsymbol{\theta}}{2m} + \sqrt{\frac{\ln\frac{3\sqrt{n}}{\delta}\ln\frac{3\sqrt{m}}{\delta} + \boldsymbol{\theta}(\ln\frac{3\sqrt{n}}{\delta} + \ln\frac{3\sqrt{m}}{\delta}) + \boldsymbol{\theta}^2}{nm}}\big) \\
\leq&\gamma^2 \big(\frac{\ln{(3\sqrt{n}/\delta)} + \boldsymbol{\theta}}{2n} + \frac{\ln{(3\sqrt{m}/\delta)} + \boldsymbol{\theta}}{2m} + \sqrt{\frac{\ln\frac{3\sqrt{n}}{\delta}\ln\frac{3\sqrt{m}}{\delta}}{nm}} + \sqrt{\frac{\boldsymbol{\theta}(\ln\frac{3\sqrt{n}}{\delta} + \ln\frac{3\sqrt{m}}{\delta})}{nm}} + \frac{\boldsymbol{\theta}}{\sqrt{nm}}\big) \\
\leq&\gamma^2 \big(\frac{\ln{(3\sqrt{n}/\delta)} + \boldsymbol{\theta}}{2n} + \frac{\ln{(3\sqrt{m}/\delta)} + \boldsymbol{\theta}}{2m} + \sqrt{\frac{\ln\frac{3\sqrt{n}}{\delta}\ln\frac{3\sqrt{m}}{\delta}}{nm}} + \sqrt{\frac{(\ln\frac{3\sqrt{n}}{\delta} + \ln\frac{3\sqrt{m}}{\delta})}{nm}}\frac{\boldsymbol{\theta}+1}{2} + \frac{\boldsymbol{\theta}}{\sqrt{nm}}\big).
\end{aligned}$$

Thus we have

$$\begin{aligned}
&\boldsymbol{\theta}^2 - \boldsymbol{\theta}\big(2\gamma\sqrt{\frac{\ln{(3\sqrt{n}/\delta)}}{2n}} + \frac{\gamma^2}{2m} + \frac{\gamma^2}{2n} + \frac{\gamma^2\sqrt{(\ln\frac{3\sqrt{n}}{\delta} + \ln\frac{3\sqrt{m}}{\delta})}}{2\sqrt{nm}} + \frac{\gamma^2}{\sqrt{nm}}\big) \\
\leq&\gamma^2\big(\frac{\ln{(3\sqrt{m}/\delta)}}{2m} + \sqrt{\frac{\ln\frac{3\sqrt{n}}{\delta}\ln\frac{3\sqrt{m}}{\delta}}{nm}} + \frac{\sqrt{\ln\frac{3\sqrt{n}}{\delta} + \ln\frac{3\sqrt{m}}{\delta}}}{2\sqrt{nm}}\big).
\end{aligned}$$

Note that if a quadratic function $x^2 - ax - b \leq 0$, we have $x \leq \frac{a}{2} + \sqrt{b + \frac{a^2}{4}} \leq a + \sqrt{b}$. Thus,

$$\begin{aligned}
\boldsymbol{\theta} \leq& 2\gamma\sqrt{\frac{\ln{(3\sqrt{n}/\delta)}}{2n}} + \frac{\gamma^2}{2m} + \frac{\gamma^2}{2n} + \frac{\gamma^2\sqrt{(\ln\frac{3\sqrt{n}}{\delta} + \ln\frac{3\sqrt{m}}{\delta})}}{2\sqrt{nm}} + \frac{\gamma^2}{\sqrt{nm}} \\
&+ \gamma\big(\frac{\ln{(3\sqrt{m}/\delta)}}{2m} + \sqrt{\frac{\ln\frac{3\sqrt{n}}{\delta}\ln\frac{3\sqrt{m}}{\delta}}{nm}} + \frac{\sqrt{(\ln\frac{3\sqrt{n}}{\delta} + \ln\frac{3\sqrt{m}}{\delta})}}{2\sqrt{nm}}\big)^{\frac{1}{2}}. \quad \blacksquare
\end{aligned}$$

APPENDIX C MORE DETAILS OF EXPERIMENTS

C.1 ARCHITECTURE OF DEEP NEURAL NETWORKS AND OPTIMIZATION SETTINGS

For shuffled pixels experiment, the network structure is designed as a 4-layer (3 hidden layers and a linear output layer) fully-connected neural network, with 400 unites per layer. The total number of parameters to be learned is $28^2 \times 400 + 400^2 + 400^2 + 400 \times 10 = 637,600$. For permuted labels experiment, the network structure is chose as a 4-layer convolutional neural network, with 2 convolutional layers of 10 and 20 filters (5 × 5 kernels), a linear hidden layer with 50 units and a linear output layer. The total number of parameters is about $5^2 \times 10 + 10 \times 5^2 \times 20 + 20^2 \times 50 + 50 \times 10 = 25,750$. We set hyperprior and hyperposterior parameters as $\kappa_{\mathcal{P}} = 2,000$ and $\kappa_{\mathcal{Q}} = 0.01$ respectively. The confidence level in PAC framework is $\delta = 0.1$. We run all methods with 150 training epochs during the meta-training phase and 200 training epochs during the meta-test phase.

C.2 THREE BOUND-MINIMIZATION META-LEARNING ALGORITHMS FOR CLASSIFICATION PROBLEM

In this section, we detail the procedure of our PAC-Bayes bound-minimization algorithms for meta-learning. The algorithm is composed of two parts: meta-training part (Algorithm 1) and meta-test part (Algorithm 2). Note that in Section C.1 we have set the deep neural network as Bayesian network (Blundell et al., 2015), therefore we can obtain the analytic form of different KL-divergences and approximate the expectation with Monte-Carlo methods. Thus we can implement our algorithm as follow. We only show the procedure of PAC-Bayes-λ-bound-minimization algorithm. The

Algorithm 1 PAC-Bayes-λ bound-minimization algorithm, meta-training phase

```
Input: Datasets from $n$ training tasks: $S_1, ..., S_n$, hyperprior $\mathcal{P}$, hyperparameter $\lambda = 1$.
Output: Parameter $\theta$ of hyperposterior $\mathcal{Q}_\theta$.
Initialize:
$\theta = (\mu_P, \rho_P) \in \mathbb{R}^{2d}$, $\phi_i = (\mu_i, \rho_i) \in \mathbb{R}^{2d}, i = 1, ..., n$.
while not converged do
    for $i \in \{1, ..n\}$ do
        Sample a mini-batch $S'_i$ from datasets $S_i$.
        Compute $\mathbf{E}_{P_\theta \sim \mathcal{Q}_\theta} \widehat{er}(Q_i, S_i)$ with the mini-batch $S'_i$ by averaging Monte-Carlo draws.
        Compute $\mathrm{KL}(\mathcal{Q}_\theta || \mathcal{P})$ with Eq. (9).
        Compute $\mathbf{E}_{P_\theta \sim \mathcal{Q}_\theta} \mathrm{KL}(Q_{\phi_i} || P_\theta)$ with Eq. (10) by averaging Monte-Carlo draws.
    end for
    Compute the PAC-Bayes-$\lambda$ bound in Theorem 2 with $\mathbf{E}_{P_\theta \sim \mathcal{Q}_\theta} \widehat{er}(Q_i, S_i)$, $\mathrm{KL}(\mathcal{Q}_\theta || \mathcal{P})$ and
    $\mathbf{E}_{P_\theta \sim \mathcal{Q}_\theta} \mathrm{KL}(Q_{\phi_i} || P_\theta)$, $i = 1, ..., n$.
    Compute the gradient of PAC-Bayes-$\lambda$ bound w.r.t $\{\theta, \phi_1, ..., \phi_n\}$ using backpropagation.
    Take an optimization step.
end while
return $\theta$
```

Algorithm 2 PAC-Bayes-λ bound-minimization algorithm, meta-test phase

```
Input: Learned hyperposterior $\mathcal{Q}_\theta$, dataset $S_{n+1}$ from task $D_{n+1}$, test data $S^*$, hyperparameter
$\lambda = 1$.
Output: Posterior $Q_{\phi_{n+1}}$ for $D_{n+1}$, and the single-task PAC-Bayes-$\lambda$ bound $B(\phi_{n+1})$ (i.e.,
test bound), classification error (i.e., test error).
Sample an informative prior $P$ from $\mathcal{Q}_\theta$
Initialize: posterior $Q_{\phi_{n+1}}$ as $P$
while not converged do
    Sample a mini-batch $S'_{n+1}$ from datasets $S_{n+1}$.
    Compute the empirical loss $\mathbf{E}_{h \sim Q_{n+1}} \widehat{er}(h, S_{n+1})$ with the mini-batch $S'_{n+1}$ by averaging
    Monte-Carlo draws.
    Compute $\mathrm{KL}(Q_{\phi_{n+1}} || P)$ with Eq. (10).
    Compute the single-task PAC-Bayes-$\lambda$ bound $B(\phi_{n+1})$ in Corollary 3 (iii) with
    $\mathbf{E}_{h \sim Q_{n+1}} \widehat{er}(h, S_{n+1})$ and $\mathrm{KL}(Q_{\phi_{n+1}} || P)$.
    Compute the gradient of the bound $B(\phi_{n+1})$ w.r.t $\phi_{n+1}$ using backpropagation.
    Take an optimization step.
end while
Use the random classifier $h \sim Q_{\phi_{n+1}}$ to classify $S^*$ to output test error.
return $Q_{\phi_{n+1}}$, test bound $B(\phi_{n+1})$ and test error.
```

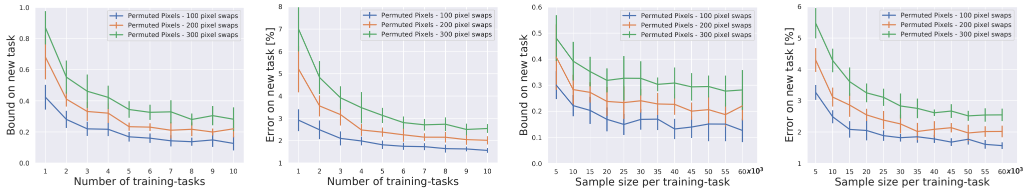
other two algorithms based on our classic bound and quadratic bound can be inferred in a similar way. Concretely, as shown in Section 4.4, by setting both the hyperprior and hyperposterior as the isotropic Gaussian distribution: $\mathcal{P} = \mathcal{N}(0, \kappa_{\mathcal{P}}^2 I_{d\times d})$, $\quad \mathcal{Q}_\theta = \mathcal{N}(\theta, \kappa_{\mathcal{Q}}^2 I_{d\times d})$, the KL-divergence between hyperposterior $\mathcal{Q}_\theta$ and hyperprior $\mathcal{P}$ can be calculated as:

$$\mathrm{KL}(\mathcal{Q}_\theta || \mathcal{P}) = \frac{||\theta||_2^2}{2\kappa_{\mathcal{P}}^2}. \tag{9}$$

Next, by choosing the prior P_θ and the posteriors Q_{ϕ_i} ($\phi_i \in \mathbb{R}^d$) as factorized Gaussian: $P_\theta(w) = \prod_{k=1}^d \mathcal{N}(w_k; \mu_{P,k}, \sigma_{P,k}^2)$, $\quad Q_{\phi_i}(w) = \prod_{k=1}^d \mathcal{N}(w_k; u_{i,k}, \sigma_{i,k}^2)$, the KL-divergence term $\mathrm{KL}(Q_{\phi_i} || P_\theta)$ in task-level complexity has a simple analytic form:

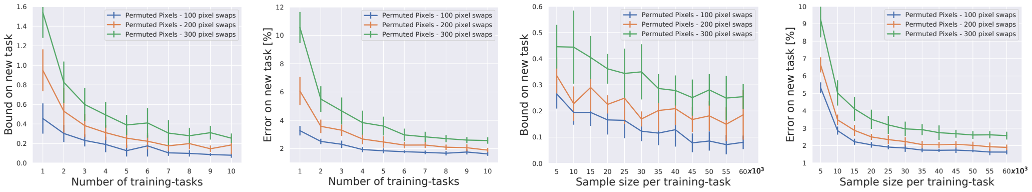
$$\mathbf{E}_{P_\theta \sim \mathcal{Q}_\theta} \sum_{i=1}^{n} \mathrm{KL}(Q_{\phi_i} || P_\theta) = \mathbf{E}_{P_\theta \sim \mathcal{Q}_\theta} \frac{1}{2} \sum_{i,k=1}^{n,d} \{\ln \frac{\sigma_{P,k}^2}{\sigma_{i,k}^2} + \frac{\sigma_{i,k}^2 + (\mu_{i,k} - \mu_{P,k})^2}{\sigma_{P,k}^2} - 1\}, \tag{10}$$

where the expectation $P_\theta \sim \mathcal{Q}_\theta$ can be approximated through Monte-Carlo method by adding Gaussian noise to the parameter θ, as defined by $\theta := \theta + \epsilon, \epsilon \sim \mathcal{N}(0, \kappa_{\mathcal{Q}}^2 I_{d\times d})$.



(a) Classic Bound (b) Classic Bound's error (c) Classic Bound (d) Classic Bound's error

Figure 3: Average test bounds and test errors of our PAC-Bayes-classic bound on 20 meta-test tasks for different pixel-shuffled environments. (a)-(b): Test bounds and test errors for different number of training tasks. (c)-(d): Test bounds and test errors for different sample size per training task.



(a) Quadratic Bound (b) Quad Bound's Error (c) Quadratic Bound (d) Quad Bound's Error

Figure 4: Average test bounds and test errors of our quadratic bound on 20 meta-test tasks for different pixel-shuffled environments. (a)-(b): Test bounds and test errors for different number of training tasks. (c)-(d): Test bounds and test errors for different sample size per training task.

C.3 Convergence Analysis of Our PAC-Bayes Classic Bound and Quadratic Bound for Meta-Learning

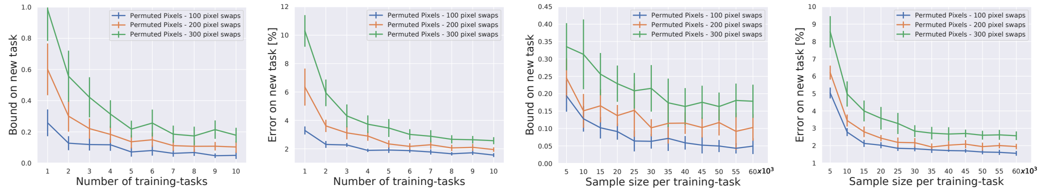
We provide detailed convergence analysis of our derived three improved PAC-Bayes bounds for meta-learning (i.e., classic/quadratic/λ bound) in Figure 3-5. We also show performance comparisons among our derived three bounds on a novel task in Figure 6. The experiments are conducted across a wide range of the number n of the training tasks and the sample size m per training task. When n changes, m is set as $60,000$ consistently; when m changes, n is set as 10. We plot the average test error and average test bound on 20 meta-test tasks from three environments with 100/200/300 pixel swaps. We can observe that: **(i)** With the increase of the number of training tasks or the sample size per training task, the test error and the test bound over the new task can both decrease to a lower level, validating the asymptotic behaviour of our derived bounds. **(ii)** Our bound-minimization method achieves better performance when the number of pixel swaps gets smaller, which reveals the importance of task relatedness of the environment.

C.4 How to Learn Data-Dependent Localized Prior via Empirical Risk Minimization (ERM)

In this section, we provide more details of how to learn data-dependent localized prior via empirical risk minimization (ERM). Such content can be considered as the supplementary explanations of Figure 2 in the main paper. Note that the localized prior defined in Eq. (5) is distribution-dependent. **However, it is difficult to set a distribution-dependent prior in practical applications, since the data distribution is always unknown.** Instead, we choose to set the data-dependent prior (that can be obtained through ERM over a number of samples) as the localized prior on the novel task. Actually, our strategy of learning a localized data-dependent prior is originated from (Parrado-Hernández et al., 2012; Liu et al., 2021). Concretely, there are three main methods in Figure 2: baseline ERM, random prior, and k-ERM prior. Given a novel task associated with 2,000 i.i.d. training samples S, the details of running these three methods are listed as follow.

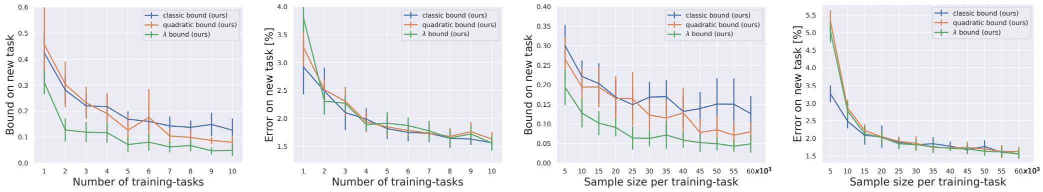
Baseline ERM Method. The baseline ERM method aims to minimize the following empirical risk over the $2,000$ training samples from the novel task with respect to a posterior Q:

$$\widehat{er}(Q, S) = \mathbf{E}_{h\sim Q} \frac{1}{2000} \sum_{i=1}^{2000} l(h, z_i),$$



(a) λ Bound (b) λ Bound's Error (c) λ Bound (d) λ Bound's Error

Figure 5: Average test bounds and test errors of our PAC-Bayes-λ bound on 20 meta-test tasks for different pixel-shuffled environments. (a)-(b): Test bounds and test errors for different number of training tasks. (c)-(d): Test bounds and test errors for different sample size per training task.



(a) Test Bound (b) Test Error (c) Test Bound (d) Test Error

Figure 6: Comparisons of average test bounds and test errors between our three PAC-Bayes bounds (i.e., classic bound, quadratic bound and λ bound) on a new task from 100-pixel-shuffled environment (average over 20 meta-test tasks). (a)-(b): Test bounds and test errors for different number of training tasks. (c)-(d): Test bounds and test errors for different sample size per training-task.

and returns a learned posterior Q for final classification, where the distribution Q is initialized as a random prior (i.e., Gaussian prior).

Random Prior Method. For the random prior method, the minimization objective is the following PAC-Bayes-λ bound for single-task learning ($\lambda \in (0, 2)$, in practice we set $\lambda = 1$):

$$\frac{\widehat{er}(Q, S)}{1 - \lambda/2} + \frac{\mathrm{KL}(Q||P) + \ln{(2\sqrt{2000}/\delta)}}{2000\lambda(1 - \lambda/2)},$$

where the prior P is chose as the random Gaussian prior. The returned posterior Q is applied for final classification over unseen data from the novel task.

k-ERM Prior Method. For the k-ERM prior method, we first use k training samples S' on this novel task (note that in MNIST dataset the $k \in (0, 58000)$ training samples are non-overlapped with the predefined $2,000$ training samples) to run the following ERM procedure:

$$\widehat{er}(Q^0, S') = \mathbf{E}_{h \sim Q^0} \frac{1}{k} \sum_{i=1}^{k} l(h, z_i).$$

The returned posterior Q^0 is called data-dependent localized prior, and is applied as the initialized prior to the following PAC-Bayes-λ bound minimization procedure over $2,000$ training samples:

$$\frac{\widehat{er}(Q, S)}{1 - \lambda/2} + \frac{\mathrm{KL}(Q||Q^0) + \ln{(2\sqrt{2000}/\delta)}}{2000\lambda(1 - \lambda/2)}.$$

The returned posterior Q will be applied for final classification on test data over the novel task. Details of the k-ERM localized prior algorithm can be found in Algorithm 3. Note that in Figure 2, we set $k = 12,000$, $24,000$, and $42,000$ respectively, and the 'random prior' can be considered as the 0-ERM prior. We can see that with the increase of k, the k-ERM prior method obtains better test performance. This indicates that with the increase of k, we can derive a localized prior with higher quality for faster adaptation to novel task. However, in practice, we find that when we increase k to a higher level (i.e., $k > 42,000$), the test performance of the k-ERM method can not be better than that of $42,000$-ERM method. Therefore, we just plot the test bounds and test errors of k-ERM methods (i.e., $k \leq 42,000$) in Figure 2. Further, since all localized priors are set over the novel

Algorithm 3 k-ERM localized prior algorithm with PAC-Bayes-λ bound, single-task learning setting

```
Input: training dataset $S_{n+1}, S$ from task $D_{n+1}$ ($S_{n+1} \bigcap S = \emptyset$, $|S| = k$), test data $S^*$, hyperparameter $\lambda = 1$.
Output: Posterior $Q_{\phi_{n+1}}$ for $D_{n+1}$, and the single-task PAC-Bayes-$\lambda$ bound $B(\phi_{n+1})$ (i.e., test bound), classification error (i.e., test error).
Initialize: localized prior $Q^0$ as a random Gaussian prior.
while not converged do        ▷ ERM over $k$ samples to learn localized prior $Q^0$
    Sample a mini-batch $S'$ from datasets $S$.
    Compute the empirical loss $\mathbf{E}_{h \sim Q^0} \widehat{er}(h, S)$ with the mini-batch $S'$ by averaging Monte-Carlo draws.
    Compute the gradient of $\widehat{er}(Q^0, S)$ w.r.t the parameter of $Q^0$ using backpropagation.
    Take an optimization step.
end while
Initialize: posterior $Q_{\phi_{n+1}}$ as the localized prior $Q^0$
while not converged do        ▷ PAC-Bayes learning with localized prior $Q^0$
    Sample a mini-batch $S'_{n+1}$ from datasets $S_{n+1}$.
    Compute the empirical loss $\mathbf{E}_{h \sim Q_{n+1}} \widehat{er}(h, S_{n+1})$ with the mini-batch $S'_{n+1}$ by averaging Monte-Carlo draws.
    Compute $\mathrm{KL}(Q_{\phi_{n+1}} || Q^0)$ with Eq. (10).
    Compute the single-task PAC-Bayes-$\lambda$ bound $B(\phi_{n+1})$ in Corollary 3 (iii) with $\mathbf{E}_{h \sim Q_{n+1}} \widehat{er}(h, S_{n+1})$ and $\mathrm{KL}(Q_{\phi_{n+1}} || Q^0)$.
    Compute the gradient of the bound $B(\phi_{n+1})$ w.r.t $\phi_{n+1}$ using backpropagation.
    Take an optimization step.
end while
Use the random classifier $h \sim Q_{\phi_{n+1}}$ to classify $S^*$ to output test error.
return $Q_{\phi_{n+1}}$, test bound $B(\phi_{n+1})$ and test error.
```

tasks, their test performance (i.e., test bounds and test errors) on novel tasks is irrelevant to the setting of training tasks (i.e., the number n of training tasks and the sample size m per training task). **Therefore, in Figure 2 of the main paper, the plots of the test bound/error of localized priors are all straight lines (with standard deviation) and do not change with the increase of n or m.**