A Unified View on PAC-Bayes Bounds for Meta-Learning

Citation for the original published paper (version of record):
Rezazadeh, A. (2022). A Unified View on PAC-Bayes Bounds for Meta-Learning. Proceedings of Machine Learning Research, 162: 18576-18595

(article starts on next page)

A Unified View on PAC-Bayes Bounds for Meta-Learning

Arezou Rezazadeh [1]

Abstract

Meta learning automatically infers an inductive bias, that includes the hyperparameter of the base-learning algorithm, by observing data from a finite number of related tasks. This paper studies PAC-Bayes bounds on meta generalization gap. The meta-generalization gap comprises two sources of generalization gaps: the environment-level and task-level gaps resulting from observation of a finite number of tasks and data samples per task, respectively. In this paper, by upper bounding arbitrary convex functions, which link the expected and empirical losses at the environment and also per-task levels, we obtain new PAC-Bayes bounds. Using these bounds, we develop new PAC-Bayes meta-learning algorithms. Numerical examples demonstrate the merits of the proposed novel bounds and algorithm in comparison to prior PAC-Bayes bounds for meta-learning.

1. Introduction

Based on Mitchell's definition (Mitchell, 1997), a machine learns a task from an experience when its performance improves with training examples of the task. In other words, during the learning process, the learner can produce a *hypothesis* that performs well on future examples of the same task. This learning process is done based on the set of assumptions known as *inductive bias* (Baxter, 2000). In many machine learning problems, finding methods for automatically learning the inductive bias is desirable. *Meta learning* also known as *learning to learn* (Thrun & Pratt, 1998) formalizes this goal by observing data from a number of inherently related tasks. Then, it uses the gained experience and knowledge to learn appropriate bias which can be fine-tuned to perform well on new tasks. Thus, the meta-learner speeds up the learning of a new, previously unseen task (Baxter, 2000). For instance, learning the initialization and the learning rate of a training algorithm (Finn et al., 2017; Li et al., 2017), the model architectures of a neural network (Zoph et al., 2018), or the optimization algorithm of a neural network (Ravi & Larochelle, 2017), all are within the scope of meta-learning.

As mentioned, the goal is extracting knowledge from several observed tasks referred to as *meta-training set*, and using the knowledge to improve performance on a novel task. The meta-learner generalizes well if after observing sufficiently training tasks, it infers a *hyperparameter* which contains good solutions to novel tasks. The good solution means that *meta-generalization loss*, which is defined as the average loss incurred by the hyperparameter when used on a new task, is minimized. However, since both data and task distributions are unknown, the meta-generalization loss can not be optimized. Instead, the meta-learner evaluates the *empirical meta-training loss* for the hyperparameter based on the meta-training set. *Meta-generalization gap* is defined as the difference between the meta-generalization loss and the meta-training loss. If the meta-generalization gap is small, it means that the meta-training loss is a good estimation of the meta-generalization loss.

Thus, bounding the meta-generalization gap is a key technique to understanding how the prior knowledge acquired from previous tasks may improve the performance of learning an unseen task. Here, a key question is 'how to regularize the meta-learner, to avoid overfitting?' The probably approximately correct (PAC)-Bayes generalization bound, is one way to answer this question.

In this paper, we derive a general framework that gives PAC-Bayes bounds on the meta-generalization gap. Under certain setups, different families of PAC-Bayes bounds, namely classic, quadratic and fast-rate families, can be re-obtained by the general framework. We also propose new PAC-Bayes classic bounds which reduce the meta-overfitting problem.

Related Work In statistical meta-learning problems, one line of research is learning of the parameters of the optimization algorithms, and analyzing gradients based on meta-learning methods (Finn et al., 2017; Konobeev et al.,

This work has been funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 893082.
[1]Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden. Correspondence to: Arezou Rezazadeh <arezour@chalmers.se>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

2020). For example, (Balcan et al., 2019; Khodak et al., 2019) worked on an online convex optimization framework with the assumption that tasks are close to a global task parameter. Additionally, (Denevi et al., 2019; 2018) studied algorithms which incrementally update the bias regularization parameter using a sequence of observed tasks. Another line of research is studying the meta-generalization gap, and finding bounds on it on average (Jose & Simeone, 2021; Rezazadeh et al., 2021) or with high probability (Pentina & Lampert, 2014; Amit & Meir, 2018; Rothfuss et al., 2021; Liu et al., 2021; Guan et al., 2022).

We recall that in the ordinary learning problem, the bound for *generalization gap* can be obtained for average generalization error scenario (Russo & Zou, 2016; Xu & Raginsky, 2017; Bu et al., 2019; Negrea et al., 2019) and PAC-Bayes scenario (McAllester, 1999; Seeger, 2002; Maurer, 2004; Catoni, 2007; Alquier, 2008; McAllester, 2013; Guedj & Pujol, 2019; Guedj, 2019; Dziugaite1 et al., 2021; Ohnishi & Honorio, 2021; Rivasplata et al., 2020). In the former case, the bound of generalization error is derived by averaging over the training set and hypothesis. While, the PAC-Bayes bounds hold with high probability.

Following the initial work of McAllester (McAllester, 1999), PAC-Bayes bounds for conventional learning have been widely investigated. Selecting different convex functions, which link the expected and empirical losses, such as KL-divergence (Seeger, 2002), square function (Mcallester, 2003) or linear function (Alquier et al., 2016) implies different PAC-Bayes bounds. The dependency on the sample size, in most of these bounds, is inversely proportional to the square root of the number of samples. In (McAllester, 2013), by choosing the convex function as $\mathrm{D}_\gamma(a||b) = \gamma a - \log(1 - b + be^\gamma)$, a family of PAC-Bayes bounds known as fast-rate bounds were obtained. In these kinds of bounds, the dependence on the sample size can be improved by the inverse of the number of samples. Directly relevant to this paper, in (Rivasplata et al., 2020) by proposing a general approach of finding PAC-Bayes bounds, various known and also new PAC-Bayes bounds were obtained.

In the meta-learning setup, inspired by the PAC-Bayes bounds for conventional learning problem, by using different convex functions, different kinds of bounds were obtained (Pentina & Lampert, 2014; Amit & Meir, 2018; Rothfuss et al., 2021; Liu et al., 2021; Guan et al., 2022). Initially, an extension of generalization error bounds to meta-learning was provided in (Pentina & Lampert, 2014) with a convergence rate $O(1/\sqrt{\mathrm{N}}) + O(1/(\mathrm{N}\sqrt{\mathrm{M}}) + 1/\sqrt{\mathrm{M}})$. To have tighter bounds, the approaches proposed in (McAllester, 1999) and (Alquier et al., 2016) have been extended to the meta-learning problem in (Amit & Meir, 2018) and (Rothfuss et al., 2021), respectively. In (Amit & Meir, 2018) with a rate $O(\sqrt{\log(\mathrm{N})/\mathrm{N}}) + O(\sqrt{\log(\mathrm{NM})/\mathrm{M}}$, by minimizing the obtained PAC-Bayes bound, a gradient-based algorithm was proposed. In (Rothfuss et al., 2021) with a convergence rate $O(1/\sqrt{\mathrm{N}}) + O(1/(\mathrm{N}\sqrt{\mathrm{M}}) + 1/\sqrt{\mathrm{N}})$, by optimizing the obtained bound, a class of PAC-optimal meta-learning algorithms was developed. To achieve meta-learning algorithms with rapid convergence ability, (Liu et al., 2021) and (Guan et al., 2022) have studied fast-rate bounds for the meta-learning setup with improved complexities.

Contributions Here, we summarize the main contributions of the paper.

- Firstly, inspired by (Rivasplata et al., 2020), by upper bounding arbitrary convex functions, which link the expected and empirical losses at environment and also per-task levels, we propose the general PAC-Bayes meta-generalization bounds (Section 3).
- Proper choices of the convex functions recover known PAC-Bayes bounds including classic, quadratic and fast-rate families (Section 4).
- We provide a new fast-rate bound and also a new classic bound with better performance on the meta-test set and with convergence rate $O(\sqrt{(1/\mathrm{N} + 1/\mathrm{M})})$ (Section 5). Following the meta-learning by adjusting the priors (MLAP) algorithm (Amit & Meir, 2018), we develop the MLAP algorithm for our new obtained bounds in the Section 6. We demonstrate the usefulness of the proposed bounds in an example in Section 7. The main merit of our new classic bound is its significant performance to avoid meta overfitting.

2. Notations, Definitions and Methods

In this paper, the sample Z takes on a value in the instance space $\mathcal{Z}$. The hypothesis space (named also as model parameter space) is denoted by $\mathcal{W}$. The non-negative loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$ measures the model parameter $w \in \mathcal{W}$ on a datasample $z \in \mathcal{Z}$, the hyperparameter space is represented by $\mathcal{U}$, and the task environment is defined by a set of tasks $\mathcal{T}$ which can be a discrete or a continuous set. The Kullback-Leibler (KL) divergence between two Bernoulli distributions with respective parameters p and q, is given by $kl(p, q)$. In other cases, the KL divergence between distributions Q and P is denoted by $\mathrm{D}(\mathrm{Q}||\mathrm{P})$.

2.1. Conventional single-task learning

In conventional learning, each task $t \in \mathcal{T}$ is associated with an underlying unknown data distribution $\mathrm{P}_{Z|T=t}$ on $\mathcal{Z}$. For a given task $t_i \in \mathcal{T}$, the *base-learner* observes a data set $\boldsymbol{Z}_i^{\mathrm{M}} = (Z_i^1, \dots, Z_i^{\mathrm{M}})$ of M independently and identically distributed (i.i.d.) samples from $\mathrm{P}_{Z|T=t_i}$. For the conventional single-task learning, the inductive bias comprising of the hyperparameter vector $u \in \mathcal{U}$ of the base-learner is fixed. For the fixed $u \in \mathcal{U}$, the base-learner uses u and the training set $\boldsymbol{Z}_i^{\mathrm{M}}$ to output a distribution over $\mathcal{W}$.

The goal of the base-learner is to infer the model parameter $w \in \mathcal{W}$ that minimizes the *per-task generalization loss* (named also as the per-task expected loss)

$$\mathrm{L}_{\mathrm{P}_{Z|t_i}}(w) = \mathbb{E}_{\mathrm{P}_{Z|t_i}}\left[\ell(w, Z)\right], \tag{1}$$

where the average is taken over a test sample $Z \sim \mathrm{P}_{Z|T=t_i}$ drawn independently from $\boldsymbol{Z}_i^{\mathrm{M}}$. Since $\mathrm{P}_{Z|T=t_i}$ is unknown, the generalization loss $\mathrm{L}_{\mathrm{P}_{Z|T=t_i}}(w)$ cannot be computed. Instead, the base-learner evaluates the *training loss*

$$\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(w) = \frac{1}{\mathrm{M}}\sum_{j=1}^{\mathrm{M}} \ell(w, Z_i^j). \tag{2}$$

The difference between the generalization loss and the training loss is referred to as the generalization gap

$$\Delta\mathrm{L}(w|\boldsymbol{Z}_i^{\mathrm{M}}, u, t_i) = \mathrm{L}_{\mathrm{P}_{Z|t_i}}(w) - \mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(w). \tag{3}$$

Roughly speaking, if the generalization gap is small, then with high probability, the performance of the inferred model parameter w on the training set can be taken as a reliable measure of the per-task generalization loss. Here, the question is that if we want to avoid overfitting and minimize per-task generalization loss with respect to w, what should be optimized on the training data $\boldsymbol{Z}_i^{\mathrm{M}}$? The PAC-Bayes framework studies this problem.

Given hyperparameter vector $u \in \mathcal{U}$, and task $t_i \in \mathcal{T}$, in the conventional single-task PAC-Bayes setting (Alquier, 2021), the base-learner assumes a prior distribution P over $\mathcal{W}$. By observing the training data $\boldsymbol{Z}_i^{\mathrm{M}}$, the base learner updates the prior distribution to a data-dependent distribution referred as posterior distribution Q_i. Having a new instance, the base learner randomly picks a model parameter $w \in \mathcal{W}$ according to Q_i. To have a guarantee that the performance of training loss for the picked w holds with high probability as the performance of per-task generalization loss, we bound generalization gap averaged over the posterior distribution, i.e., $\mathbb{E}_{W\sim\mathrm{Q}_i}[\Delta\mathrm{L}(W|\boldsymbol{Z}_i^{\mathrm{M}}, u, t_i)]$.

Roughly speaking, most PAC-Bayes proofs follow four key steps. (Alquier, 2021) presents a comprehensive tutorial about PAC-Bayes bounds. Here, we review the key steps of finding PAC-Bayes bounds. Let $\mathrm{F}(a, b)$ be a convex function in both a and b. Firstly, a suitable convex function such as $\mathrm{F}(\cdot, \cdot)$ links the expected loss averaged over the posterior distribution with the empirical loss averaged over the posterior distribution. Then, by applying Jensen's inequality, the function over the expectation (posterior distribution) is bounded by the expectation of the function. By using a change of measure inequality (Ohnishi & Honorio, 2021), we find a bound in terms of a divergence (usually KL-divergence between posterior and prior distributions), and the expectation of the function over prior distribution. Then, by applying Markov's inequality, we usually bound the expectation of the function with the logarithm of the confidence parameter. Thus, the convex function linking the expected and empirical losses is bounded by a complexity term, like $\mathrm{F}(a, b) \leq c$. Usually, a further bounding technique, which we refer to as the 'affine transformation', is used to bound the expected loss as an affine transformation of the complexity term. It means that from $\mathrm{F}(a, b) \leq c$, one can conclude that $a \leq k \cdot b + \mathrm{G}(c)$, where $k \in \mathbb{R}$ is a coefficient, and $\mathrm{G} : \mathbb{R}^+ \to \mathbb{R}$.

To look through the mentioned concepts in detail, we consider the conventional PAC-Bayes bound in (McAllester, 1999). In (McAllester, 1999), by setting $\mathrm{F}^{\text{Task}}(a, b) = 2(\mathrm{M}-1)(a-b)^2$, it is proved that given the prior distribution P, for any confidence parameter $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$2(\mathrm{M}-1)\left(\mathbb{E}_{W\sim\mathrm{Q}_i}\left[\mathrm{L}_{\mathrm{P}_{Z|t_i}}(W) - \mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right]\right)^2 \leq \mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right) + \log\left(\frac{\mathrm{M}}{\delta}\right). \tag{4}$$

The right hand side of (4) is known as the *complexity term*, and contains KL-divergence, as the information gain in specializing from the prior to posterior distributions, and the log-term, as the dependence expression on the confidence parameter, and the number of samples M. A learning algorithm with generalization guarantee selects a posterior distribution Q_i which minimizes (4). Since minimizing (4) is not easy, to find bounds which are convenient to minimize, we apply the affine transformation step. In other words, for the convex function $\mathrm{F}^{\text{Task}}(a, b) = 2(\mathrm{M}-1)(a-b)^2$, since from $\mathrm{F}^{\text{Task}}(a, b) \leq c_{\text{tsk}}$, we have $a \leq 1.b + \sqrt{c_{\text{tsk}}/(2(\mathrm{M}-1))}$, the affine transformation leads to $k_t = 1$ and $\mathrm{G}_{\text{Task}}(c) = \sqrt{c_{\text{tsk}}/(2(\mathrm{M}-1))}$. It means that from (4), the following inequality holds uniformly for all posteriors distributions Q_i

$$\mathbb{E}_{W\sim\mathrm{Q}_i}\left[\mathrm{L}_{\mathrm{P}_{Z|t_i}}(W) - \mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right] \leq \sqrt{\frac{\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right) + \log(\frac{\mathrm{M}}{\delta})}{2(\mathrm{M}-1)}}. \tag{5}$$

2.2. Meta-Learning

The goal of meta-learning is automatically infer the hyperparameter u of the base learner from training data pertaining to a number of related tasks. The tasks are assumed to belonging to a task environment, which is defined by a task distribution P_T on the space of tasks $\mathcal{T}$, and by the per-task data distributions $\{\mathrm{P}_{Z|T=t}\}_{t\in\mathcal{T}}$. The meta-learner observes a meta-training set $\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}} = (\boldsymbol{Z}_1^{\mathrm{M}}, \ldots, \boldsymbol{Z}_{\mathrm{N}}^{\mathrm{M}})$ of N data sets.

Without loss of generality, we assume that number of samples of all tasks equals to M. The obtained results can be easily generalized to the case where per-task data samples are not equal. Each $\boldsymbol{Z}_i^{\mathrm{M}}$ is generated independently by first drawing a task $T_i \sim \mathrm{P}_T$ and then a task-specific dataset $\boldsymbol{Z}_i^{\mathrm{M}} \sim \mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}|T_i}$.

The meta-learner uses the meta-training set $\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}$ to infer the hyperparameter u. In the PAC-Bayes setup for meta learning, the goal of the meta-learner is to infer hyperparameter u from the observed tasks, and then use u as a prior knowledge for learning new (yet unobserved) tasks from task environment $\mathcal{T}$. The quality of u is is measured by the *meta-generalization loss* when using it to learn new tasks. Formally, the objective of the meta-learner is to infer the hyperparameter u that minimizes the meta-generalization loss

$$\mathrm{L}_{\mathrm{P}_{\mathcal{T}\boldsymbol{Z}^{\mathrm{M}}}}(u) = \mathbb{E}_{\mathrm{P}_T \mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}|T}}\left[\mathbb{E}_{W\sim \mathrm{Q}}[\mathrm{L}_{\mathrm{P}_{Z|T}}(W)]\right], \tag{6}$$

where the expectation is taken over an independently generated meta-test task $T \sim \mathrm{P}_T$, over the associated data set $\boldsymbol{Z}^{\mathrm{M}} \sim \mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}|T}$, and over the output of the base-learner. Since P_T and $\{\mathrm{P}_{Z|T=t}\}_{t\in\mathcal{T}}$ are unknown, the meta-generalization loss (6) cannot be computed. Instead, the meta-learner can evaluate the *meta-training loss*, which for a given hyperparameter u, is defined as the average training loss on the meta-training set

$$\mathrm{L}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}(u) = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \mathbb{E}_{W\sim \mathrm{Q}_i}[\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)]. \tag{7}$$

Here, the average is taken over the output of the base-learner. The difference between meta-generalization loss and the meta-training loss is the *meta-generalization gap*

$$\Delta\mathrm{L}(u|\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}) = \mathrm{L}_{\mathrm{P}_{\mathcal{T}\boldsymbol{Z}^{\mathrm{M}}}}(u) - \mathrm{L}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}(u). \tag{8}$$

Small meta-generalization gap means that with high probability, the performance of the inferred hyperparameter u on the meta-training set can be taken as a reliable measure of the meta-generalization loss (6).

In the PAC-Bayes setup for meta learning, the meta- learner assumes a *hyper-prior distribution* $\mathcal{P} \in \mathcal{P}_{\mathcal{U}}$ over hyperparameter space $\mathcal{U}$, observes the meta-training set $\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}$, and updates the hyper-prior distribution to a data-dependent distribution referred as *hyper-posterior distribution* $\mathcal{Q} \in \mathcal{P}_{\mathcal{U}}$. The goal is to use the hyper-posterior distribution for learning new and unseen tasks. In other words, having a new task, the meta learner randomly picks u according to hyper-posterior distribution $\mathcal{Q}$ and then use it for learning of posterior Q_i.

One approach for finding the PAC-Bayes bounds for meta learning, is decomposing the meta-generalization gap into environment-level and within-task generalization gaps. We define the decomposition term as

$$\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \tilde{\mathrm{L}}_{\boldsymbol{Z}_i^{\mathrm{M}}}^{t_i}(U), \tag{9}$$

where $\tilde{\mathrm{L}}_{\boldsymbol{Z}_i^{\mathrm{M}}}^{t_i}(u)$ is the average per-task generalization loss

$$\tilde{\mathrm{L}}_{\boldsymbol{Z}_i^{\mathrm{M}}}^{t_i}(u) = \mathbb{E}_{W\sim \mathrm{Q}_i}\left[\mathrm{L}_{\mathrm{P}_{Z|t_i}}(W)\right]. \tag{10}$$

From (6), we can express the meta-expected loss as $\mathrm{L}_{\mathrm{P}_{\mathcal{T}\boldsymbol{Z}^{\mathrm{M}}}}(U) = \mathbb{E}_{\mathrm{P}_{\mathcal{T}\boldsymbol{Z}^{\mathrm{M}}}}[\tilde{\mathrm{L}}_{\boldsymbol{Z}^{\mathrm{M}}}^{T}(U)]$. Recalling that $\mathrm{F}^{\mathrm{Env}}(a,b)$ is a convex function in both a and b. In the PAC-Bayes setup for meta learning, we can follow the mentioned four steps for both environment-level generalization gap

$$\mathrm{F}^{\mathrm{Env}}\left(\mathbb{E}_{U\sim\mathcal{Q}}\mathbb{E}_{\mathrm{P}_{\mathcal{T}\boldsymbol{Z}^{\mathrm{M}}}}\left[\tilde{\mathrm{L}}_{\boldsymbol{Z}^{\mathrm{M}}}^{T}(U)\right], \mathbb{E}_{U\sim\mathcal{Q}}\left[\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}_{\boldsymbol{Z}_i^{\mathrm{M}}}^{t_i}(U)\right]\right), \tag{11}$$

and within-task generalization gap

$$\mathrm{F}^{\mathrm{Task}}\left(\mathbb{E}_{U\sim\mathcal{Q}}\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathbb{E}_{W\sim\mathrm{Q}_i}\left(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)\right)\right),\right.$$
$$\left.\mathbb{E}_{U\sim\mathcal{Q}}\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathbb{E}_{W\sim\mathrm{Q}_i}[\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)]\right)\right), \tag{12}$$

separately (Pentina & Lampert, 2014; Amit & Meir, 2018; Rothfuss et al., 2021; Liu et al., 2021; Guan et al., 2022).

3. General Meta-Learning PAC-Bayes Bounds

In this section, inspired by (Rivasplata et al., 2020), we find a general approach for finding PAC-Bayes bounds for meta-generalization gap.

Theorem 3.1 (General PAC-Bayes Bounds)**.** *Let* $\mathrm{F}^{\mathrm{Task}}(a,b)$ *and* $\mathrm{F}^{\mathrm{Env}}(a,b)$ *be two functions which are convex in both a and b. Additionally, assuming that the tasks are drawn independently from the task environment* $\mathcal{T}$ *according to distribution* P_T. *For the task and environment level priors* P *and* $\mathcal{P}$, *with a probability at least* $1-\delta$, *under* $\mathrm{P}_{T_{1:\mathrm{N}}}\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}$, *for* $\theta_{\mathrm{tsk}}, \theta_{\mathrm{env}} \geq 0$ *we have*

$$\mathrm{F}^{\mathrm{Env}}\left(\mathbb{E}_{U\sim\mathcal{Q}}\left(\mathrm{L}_{\mathrm{P}_{\mathcal{T}\boldsymbol{Z}^{\mathrm{M}}}}(U)\right), \mathbb{E}_{U\sim\mathcal{Q}}\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}_{\boldsymbol{Z}_i^{\mathrm{M}}}^{T_i}(U)\right)\right)$$
$$+\,\mathrm{F}^{\mathrm{Task}}\left(\mathbb{E}_{U\sim\mathcal{Q}}\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}_{\boldsymbol{Z}_i^{\mathrm{M}}}^{T_i}(U)\right), \mathbb{E}_{U\sim\mathcal{Q}}\left(\mathrm{L}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}(U)\right)\right)$$
$$\leq \left(\frac{1}{\theta_{\mathrm{tsk}}}+\frac{1}{\theta_{\mathrm{env}}}\right)\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) + \frac{1}{\mathrm{N}\cdot\theta_{\mathrm{tsk}}}\mathbb{E}_{\mathcal{Q}}\left(\sum_{i=1}^{\mathrm{N}}\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right)$$
$$+\log\frac{\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}}\left(\Upsilon_{\mathrm{tsk}}^{\frac{1}{\mathrm{N}\cdot\theta_{\mathrm{tsk}}}}\cdot\Upsilon_{\mathrm{env}}^{\frac{1}{\theta_{\mathrm{env}}}}\right)}{\delta}, \tag{13}$$

where

$$\Upsilon_{\text{env}} = \mathbb{E}_{\mathcal{P}} e^{\theta_{\text{env}} \mathrm{F}^{\text{Env}}\left(\mathrm{L}_{\mathrm{P}_{\boldsymbol{TZ}^{\mathrm{M}}}}(U), \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)\right)} \tag{14}$$

$$\Upsilon_{\text{tsk}} = \prod_{i=1}^{\mathrm{N}} \mathbb{E}_{\mathcal{P}\mathrm{P}} e^{\theta_{\text{tsk}} \mathrm{F}^{\text{Task}}\left(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W), \mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right)}. \tag{15}$$

Proof. See Appendix A. □

Generally, to obtain (13), we applied only one Markov's inequality. Thus, on the left hand side of (13), we have the sum of $\mathrm{F}^{\text{Env}}(\cdot)$ and $\mathrm{F}^{\text{Task}}(\cdot)$. A relaxed form of (13), can be obtained by applying the affine transformation and also Markov's inequality two times at the task and environment levels, separately. As discussed in (5), the affine transformation leads to a new function denoted by $\mathrm{G}()$. For example, if the convex function is $\mathrm{F}(a,b) = (a-b)^2$, since from $\mathrm{F}(a,b) = (a-b)^2 \leq c$, we conclude that $a \leq b + \sqrt{c}$, the affine transformation leads to $k=1$ and $\mathrm{G}(c) = \sqrt{c}$. Similarly, $\mathrm{F}(a,b) = kl(a,b) \leq c$ leads to $k = 1/(1-0.5\lambda)$, and $\mathrm{G}(c) = c/(\mathrm{M}.\lambda(1-0.5\lambda))$ (Thiemann et al., 2017) . The following corollary is a relaxation of (13).

Corollary 3.2. *Under the setting of Theorem 3.1, assume that $\mathrm{G}_{\text{Task}}(\cdot)$ and $\mathrm{G}_{\text{Env}}(\cdot)$ are two functions where from $\mathrm{F}^{\text{Task}}(a,b) \leq c_{\text{tsk}}$ (res. $\mathrm{F}^{\text{Env}}(a,b) \leq c_{\text{env}}$), we can conclude $a \leq k_t \cdot b + \mathrm{G}_{\text{Task}}(c_{\text{tsk}})$ (resp. $a \leq k_e \cdot b + \mathrm{G}_{\text{Env}}(c_{\text{env}})$) for $k_t \in \mathbb{R}^+$ (resp. $k_e \in \mathbb{R}^+$). In this case, with probability at least $1-\delta$, under $\mathrm{P}_{T_{1:\mathrm{N}}}\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}$, we have*

$$\mathbb{E}_{U\sim\mathcal{Q}}\left[\mathrm{L}_{\mathrm{P}_{\boldsymbol{TZ}^{\mathrm{M}}}}(U)\right] \leq k_e \cdot k_t \cdot \mathbb{E}_{U\sim\mathcal{Q}}\left[\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)\right] + \mathrm{G}_{\text{Env}}(\mathrm{B}_{\text{Env}}) + \frac{k_e}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \mathrm{G}_{\text{Task}}(\mathrm{B}_{\text{Task}}), \tag{16}$$

where setting $\Lambda_{\text{env}} = e^{\theta_{\text{env}} \mathrm{F}^{\text{Env}}\left(\mathrm{L}_{\mathrm{P}_{\boldsymbol{TZ}^{\mathrm{M}}}}(U), \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)\right)}$

$$\mathrm{B}_{\text{Env}} = \frac{1}{\theta_{\text{env}}}\mathrm{D}(\mathcal{Q}||\mathcal{P}) + \log\left(\frac{2\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}}\mathbb{E}_{\mathcal{P}}\Lambda_{\text{env}}}{\delta}\right)^{\frac{1}{\theta_{\text{env}}}}, \tag{17}$$

and setting $\Lambda_{\text{Task}} = e^{\theta_{\text{tsk}} \mathrm{F}^{\text{Task}}\left(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W), \mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right)}$

$$\mathrm{B}_{\text{Task}} = \frac{1}{\theta_{\text{tsk}}}\mathrm{D}(\mathcal{Q}||\mathcal{P}) + \frac{1}{\theta_{\text{tsk}}}\mathbb{E}_{\mathcal{Q}}\left[\mathrm{D}(\mathrm{Q}_i||\mathrm{P})\right] + \log\left(\frac{2\mathrm{N}\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}}\mathbb{E}_{\mathcal{P}\mathrm{P}}\Lambda_{\text{Task}}}{\delta}\right)^{\frac{1}{\theta_{\text{tsk}}}}. \tag{18}$$

Proof. See Appendix A. □

4. Re-obtaining Existing Results

In this section, by applying different $\mathrm{F}^{\text{Env}}(\cdot,\cdot)$ and $\mathrm{F}^{\text{Task}}(\cdot,\cdot)$ to (16), we re-obtain all the previous main PAC-Bayes bounds for the meta-learning problem. Table 1, summarizes the results. For the derivation see Appendix B.

5. New PAC-Bayes Bounds for Meta-Learning

In this section, we insert different $\mathrm{F}^{\text{Env}}(,)$ and $\mathrm{F}^{\text{Task}}(,)$ in (13), and then we bound (14) and (15) by using Lemmas presented in Appendix F. For simplicity, we assume that the loss function is bounded on the interval $[0,1]$, and hence it is sub-Gaussian with parameter 0.5.

Mainly, we present a new fast-rate bound and a new classic bound. To obtain fast-rate bound, like existing bounds, we use (16). It means that we apply Markov's inequality and affine-transformation steps in both the environment and task levels. However, to find the classic bound, we find a lower bound for the left-hand side of (13), and then we apply the affine-transformation step. It means that to obtain new classic bound, we apply both Markov's inequality and affine-transformation step, once. Firstly, we preset the fast-rate bound.

Theorem 5.1. *Under the setting of Theorem 3.1, for $\mathrm{N} \geq 2$, the meta-generalization gap is bounded by*

$$\mathbb{E}_{\mathcal{Q}}\left[\mathrm{L}_{\mathrm{P}_{\boldsymbol{TZ}^{\mathrm{M}}}}(U)\right] \leq \min_{\lambda_e,\lambda_t \geq 0.5} \frac{1}{1-\frac{1}{2\lambda_e}} \cdot \frac{\mathbb{E}_{\mathcal{Q}}\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)}{1-\frac{1}{2\lambda_t}} + \frac{\lambda_e}{1-\frac{1}{2\lambda_e}}\left(\frac{\mathrm{D}(\mathcal{Q}||\mathcal{P}) + \log\frac{2}{\delta}}{\mathrm{N}}\right) + \frac{1}{1-\frac{1}{2\lambda_e}} \cdot \frac{\lambda_t}{1-\frac{1}{2\lambda_t}} \cdot \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \frac{\mathrm{D}(\mathcal{Q}||\mathcal{P}) + \mathbb{E}_{\mathcal{Q}}\left[\mathrm{D}(\mathrm{Q}_i||\mathrm{P})\right] + \log\frac{2\mathrm{N}}{\delta}}{\mathrm{M}}, \tag{19}$$

where (19) is referred as the fast-rate bound for meta- learning.

Proof. See Appendix C □

Now, we obtain a new classic bound. Setting $\mathrm{F}^{\text{Task}}(a,b) = 2(\mathrm{M}-1)(a-b)^2$ and $\mathrm{F}^{\text{Env}}(a,b) = (\mathrm{N}-1)(a-b)^2$ in (13), we will obtain a new bound with a single square, unlike existing bounds. The key step to find bounds with a single square is the following inequality

$$\frac{nm}{n+m}(a-c)^2 \leq n(a-b)^2 + m(b-c)^2, \tag{20}$$

where $n, m \in \mathbb{N}$. To show (20), consider the function $f(a,b,c) \triangleq n(a-b)^2 + m(b-c)^2$. Since $\partial^2 f/\partial b^2 = 2(n+m) > 0$, the function f is convex with respect to b. Hence, by setting the first derivative of f with respect to

Table 1: Existing PAC-Bayes bounds on meta generalization gap can be obtained as a special case of (16)

.

Meta-Learning PAC Bayes Bounds			
Bound	$\mathrm{F}^{\text{Task}}(a,b)$, $\mathrm{F}^{\text{Env}}(a,b)$	Other parameters	Affine transformation
MLAP (Amit & Meir, 2018)	$\mathrm{F}^{\text{Task}}(a,b) = 2(\mathrm{M}-1)(a-b)^2$, $\mathrm{F}^{\text{Env}}(a,b) = 2(\mathrm{N}-1)(a-b)^2$	$\theta_{\text{tsk}} = \theta_{\text{env}} = 1$	$\mathrm{G}_{\text{Env}}(c) = \sqrt{c/(2(\mathrm{N}-1))}$, $k_e = 1$, $\mathrm{G}_{\text{Task}}(c) = \sqrt{c/(2(\mathrm{M}-1))}$, $k_t = 1$
PACOH (Rothfuss et al., 2021)	$\mathrm{F}^{\text{Task}}(a,b) = (a-b)$, $\mathrm{F}^{\text{Env}}(a,b) = (a-b)$	$\theta_{\text{tsk}} = \beta$, $\theta_{\text{env}} = \lambda$	$\mathrm{G}_{\text{Env}}(c) = c$, $k_e = 1$, $\mathrm{G}_{\text{Task}}(c) = c$, $k_t = 1$
λ-Bound (Liu et al., 2021)	$\mathrm{F}^{\text{Task}}(a,b) = \mathrm{M}kl(a,b)$, $\mathrm{F}^{\text{Env}}(a,b) = 2(\mathrm{N}-1)(a-b)^2$	$\theta_{\text{tsk}} = \theta_{\text{env}} = 1$, $\lambda \in (0,2)$	$\mathrm{G}_{\text{Env}}(c) = \sqrt{c/(2(\mathrm{N}-1))}$, $k_e = 1$, $k_t = 1/(1-0.5\lambda)$, $\mathrm{G}_{\text{Task}}(c) = c/(\mathrm{M}\lambda(1-0.5\lambda))$
Classic bound (Guan et al., 2022)	$\mathrm{F}^{\text{Task}}(a,b) = \mathrm{M}kl(a,b)$, $\mathrm{F}^{\text{Env}}(a,b) = \mathrm{N}kl(a,b)$	$\theta_{\text{tsk}} = \theta_{\text{env}} = 1$	$k_t = k_e = 1$, $\mathrm{G}_{\text{Task}}(c) = \sqrt{c/2\mathrm{M}}$, $\mathrm{G}_{\text{Env}}(c) = \sqrt{c/2\mathrm{N}}$
Quadratic bound (Guan et al., 2022)	$\mathrm{F}^{\text{Task}}(a,b) = \mathrm{M}kl(a,b)$, $\mathrm{F}^{\text{Env}}(a,b) = \mathrm{N}kl(a,b)$	$\theta_{\text{tsk}} = \theta_{\text{env}} = 1$	$k_e = 1$, $\mathrm{G}_{\text{Env}}(c) = \sqrt{c/2\mathrm{N}}$, $b \leq (\sqrt{a + (c/2)} + \sqrt{c/2})^2$
λ bound (Guan et al., 2022)	$\mathrm{F}^{\text{Task}}(a,b) = \mathrm{M}kl(a,b)$, $\mathrm{F}^{\text{Env}}(a,b) = \mathrm{N}kl(a,b)$	$\lambda \in (0,2)$ $\theta_{\text{tsk}} = \theta_{\text{env}} = 1$	$k_t 1/(1-0.5\lambda)$, $\mathrm{G}_{\text{Task}}(c) = c/(\lambda(1-0.5\lambda))$ $k_e = 1$, $\mathrm{G}_{\text{Env}}(c) = \sqrt{c/2\mathrm{N}}$

b equal to zero, $b^\star = (na+mc)/(n+m)$ minimizes f. Since $f(a,b^\star,c) \leq f(a,b,c)$, and $f(a,b^\star,c)$ equals to the left-hand side of (20), we conclude (20).

Now, in (20), we set $a = \mathbb{E}_{U\sim\mathcal{Q}}(\mathrm{L}_{\mathrm{P}_{\mathcal{T}\boldsymbol{Z}^{\mathrm{M}}}}(U))$, $b = \mathbb{E}_{U\sim\mathcal{Q}}(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{t_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U))$ and $c = \mathbb{E}_{U\sim\mathcal{Q}}(\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U))$. usually the right hand side of (20) gives us the left hand side of (13). Thus, if the right hand side of (20) is upper bounded by B, from (20) we conclude that

$$\left|\mathbb{E}_{U\sim\mathcal{Q}}\left[\mathrm{L}_{\mathrm{P}_{\mathcal{T}\boldsymbol{Z}^{\mathrm{M}}}}(U) - \mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)\right]\right| \leq \sqrt{\frac{n+m}{nm}\mathrm{B}}. \quad (21)$$

Now, in view of (20), we insert various convex functions to (13). In the following, we present one of them where we set $\mathrm{F}^{\text{Task}}(a,b) = 2(\mathrm{M}-1)(a-b)^2$ and $\mathrm{F}^{\text{Env}}(a,b) = (\mathrm{N}-1)(a-b)^2$ in (13), and $n = (\mathrm{N}-1)$, $m = 2(\mathrm{M}-1)$, in (20).

Theorem 5.2. *Under the setting of Theorem 3.1,* $\mathrm{N} \geq 2$, *the meta-generalization gap is bounded by*

$$\left|\mathbb{E}_{U\sim\mathcal{Q}}\left[\mathrm{L}_{\mathrm{P}_{\mathcal{T}\boldsymbol{Z}^{\mathrm{M}}}}(U) - \mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)\right]\right| \leq \sqrt{\frac{(\mathrm{N}-1)+2(\mathrm{M}-1)}{2(\mathrm{N}-1)(\mathrm{M}-1)}} \cdot \sqrt{2\mathrm{D}(\mathcal{Q}\|\mathcal{P}) + \mathbb{E}_{\mathcal{Q}}\left[\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathrm{D}(\mathrm{Q}_i\|\mathrm{P})\right] + \log\frac{\mathrm{M}\sqrt{\mathrm{N}}}{\delta}}. \quad (22)$$

Proof. See Appendix D. □

We recall that (22) is expressed in terms of a single square. Thus, compared to existing bounds, minimizing (22) is less complicated and one can enjoy the properties of presented bounds in (Rothfuss et al., 2021). In other words, it may reduce the problem of meta overfitting.

In this sections, we only presented a member of fast-rate and a member of classic families. We can apply different $\mathrm{F}^{\text{Env}}(a,b)$ and $\mathrm{F}^{\text{Task}}(a,b)$ functions, and obtain new different bounds. For more new bounds, see Appendix E.

6. Meta-Learning Algorithm

By minimizing the bounds obtained in Section 5, we can develop PAC-Bayes bound-minimization algorithms for meta-learning with deep neural networks. Here, for (22), and in view of (Amit & Meir, 2018) we obtain a meta-learning algorithm. Similar algorithms can be found for all bounds presented in Section 5. Following previous work (Amit & Meir, 2018), we consider $\mathcal{W} = \{h_w, w \in \mathbb{R}^d\}$ as a set of neural networks with certain parameters. For $\mathrm{N}_p, d \in \mathbb{N}$, let $\mathcal{U} \subset \mathbb{R}^{\mathrm{N}_p}$ and $\mathcal{W} \subset \mathbb{R}^d$ for all tasks t_i. Like previous works, (Amit & Meir, 2018; Guan et al., 2022), we set the hyper-prior distribution as

$$\mathcal{P} \triangleq \mathcal{N}(0, \kappa_p^2 \mathrm{I}_{\mathrm{N}_p} \times \mathrm{I}_{\mathrm{N}_p}), \quad (23)$$

where $\kappa_p^2 > 0$ is a predefined constant. We limit the space of hyper-posteriors as a family of isotropic Gaussian distributions defined by

$$\mathcal{Q} \triangleq \mathcal{N}(\theta, \kappa_s^2 \mathrm{I}_{\mathrm{N}_p} \times \mathrm{I}_{\mathrm{N}_p}), \quad (24)$$

Table 2: Comparison of different PAC-Bayes bounds on 20 test tasks. During the meta-training phase, each task is constructed with 8000 images, and during the meta-test phase, each task is constructed with 2000 images. The number of training task is set as $\mathrm{N} = 5$. The number of epochs is 100.

Meta-Learning PAC Bayes Bounds: Test Error (%)		
Bound	100-Swap Shuffled pixels	Permuted labels
MLAP (Amit & Meir, 2018)	31.4 with STD 1.99%	7.95 with STD 0.441%
kl-Bound (Liu et al., 2021)	26.7 with STD 1.14%	9.86 with STD 0.915%
λ-Bound (Liu et al., 2021)	28.1 with STD 1.28%	9.8 with STD 0.859%
Bound 1 (Guan et al., 2022)	31.8 with STD 1.31%	8.026 with STD 0.356%
Bound 2 (Guan et al., 2022)	32 with STD 1.44%	11.06 with STD 0.588%
Bound 3 (Guan et al., 2022)	28.8 with STD 1.6%	10.3 with STD 0.784%
Our classic bound (22)	9.39 with STD 0.454%	3.1 with STD 0.279%
Our fast-rate bound (19)	25.2 with STD 1.23%	9.77 with STD 0.824%

where $\theta \in \mathbb{R}^{\mathrm{N}_p}$ is the optimization parameter, and $\kappa_s^2 > 0$ is a predefined constant.

Next, we consider the posterior and prior distributions over $\mathcal{W}$. For all tasks $t_i \in \mathcal{T}$, $\mathcal{W}$ can be seen as a family of functions parameterized by a weight vector $\boldsymbol{a}^d = [a_1, \ldots, a_d]$. For a given hyperparameter u, let the weight vector is denoted by $\boldsymbol{a}^d$. We define the prior distribution as factorized Gaussian distributions

$$\mathrm{P}(\boldsymbol{a}^d|u) = \prod_{k=1}^{d} \mathrm{P}(a_k|u) = \prod_{k=1}^{d} \mathcal{N}(a_k; \mu_u(k), \sigma_u^2(k)), \tag{25}$$

meaning that, before observing data, the k-th weight denoted by a_k, takes values according to Gaussian distribution with mean $\mu_u(k)$ and variance $\sigma_u^2(k)$. After observing data, for task $t_i \in \mathcal{T}$, a_k takes values according to Gaussian distribution with mean $\mu_i(k)$ and variance $\sigma_i^2(k)$. Thus, the posterior distribution of task t_i is

$$\mathrm{Q}_i(\boldsymbol{a}^d|\boldsymbol{z}_i^{\mathrm{M}_i}, u) = \prod_{k=1}^{d} \mathcal{N}(a_k; \mu_i(k), \sigma_i^2(k)). \tag{26}$$

From (23) and (24), it can be verified that

$$\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) = \frac{||\theta||_2^2 + \kappa_s^2}{2\kappa_p^2} + \log\frac{\kappa_p^2}{\kappa_s^2} - \frac{1}{2}. \tag{27}$$

Similarly, from (25) and (26), for task t_i, we find that

$$\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right) = \frac{1}{2}\sum_{k=1}^{d}\left(\log\frac{\sigma_u^2(k)}{\sigma_i^2(k)} + \log\frac{\sigma_i^2(k) + (\mu_i(k) - \mu_u(k))^2}{\sigma_u^2(k)}\right). \tag{28}$$

Inserting (27) and (28) into (22), it remains to select the parameters of posterior distribution Q_i minimizing (22). Since the square root function is strictly increasing, an equivalent optimization problem is the minimization of the objective function inside the square of (22).

Following the optimization technique described in (Amit & Meir, 2018), approximating the expectation $\mathbb{E}_{U\sim\mathcal{N}(\theta,\kappa_s^2\mathrm{I}_{\mathrm{N}_p}\times\mathrm{I}_{\mathrm{N}_p})}$ by averaging several Monte-Carlo samples of U, the optimal posterior distribution can be obtained by evaluating the gradient of (22) with respect to $(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$ as described in Section 4.4 of (Amit & Meir, 2018).

We recall that, like (Rothfuss et al., 2021), the minimization problem of (22) is equivalent to the simpler problem than the optimization of existing classic bounds. In fact, it suffices to minimize an objective function, which is linear with respect to KL-divergences. This leads to Gibbs posteriors, and might be the reason why the obtained algorithm reduces the meta-overfitting problem.

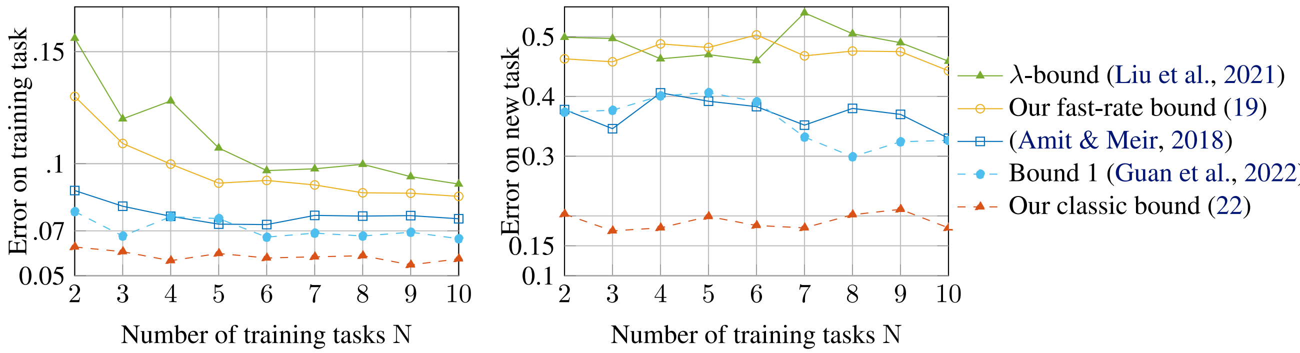
7. Numerical Results

Using the same experiment given by Section 5 of (Amit & Meir, 2018) and also (Liu et al., 2021; Guan et al., 2022), we compare our bounds with previous works. We reproduce the experimental results of our method by directly running the online code[1] from (Amit & Meir, 2018), and run our algorithm by replacing others' bounds with our bounds.

In image classification, the data samples $z = (x, y)$, consist of a an image, x and a label y. We consider an experiment based on augmentations of the MNIST dataset. We study two experiments, namely permuted labels and permuted pixels. For permuted labels, each task is created by a random permutation of image labels. For permuted pixels each task is created by a permutation of image pixels. The pixel permutations are achieved by 100 location swaps to ensure the task relatedness.

The network architecture used for the permuted-label ex-

[1]https://github.com/ron-amit/meta-learning-adjusting-priors2



(a) Meta-training error

(b) Meta-test error

Figure 1: The average training and test errors versus the number of training-tasks. The number of training examples for each task is 600 images, and during the meta-test phase, each task is constructed with 100 images. The number of epochs is 100.

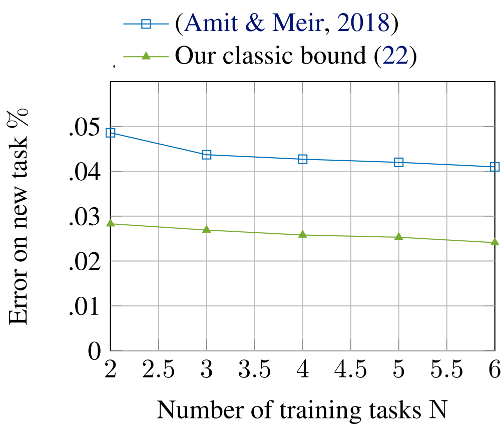


Figure 2: The average test error of learning a new task versus N. The number of training example of each task 8000. The number of epochs is 400.

periment is a small CNN with two convolutions layers, a linear hidden layer and a linear output layer (Amit & Meir, 2018). With a learning rate of 10^{-3}, we use the hyper-prior, prior, hyper-posterior and posterior distributions given by (23), (25), (24) and (26), respectively. We set $\kappa_p^2 = 100$, $\kappa_s^2 = 0.001$, and $\delta = 0.1$. For each task τ_i, and $k = 1, \ldots, d$, the posterior parameter $\log(\sigma_i^2(k))$ initialized by $\mathcal{N}(-10, 0.01)$, $\mu_i(k)$ is initialized randomly with the Glorot method (Glorot & Bengio, 2010). Then, for different bounds, by using backpropagation, we evaluate the gradient of the bound with respect to $\boldsymbol{\mu}_i = (\mu_i(1), \ldots, \mu_i(d))$. Then, we set $\mu_i(k)$ and $\sigma_i(k)$ as the means and variance of k-th weight. The parameters $\mu_u(k)$ and $\sigma_u(k)$ are similar in structure, and the parameter θ is the vector containing the weights of N tasks (Amit & Meir, 2018).

Table 2 shows the comparison of different PAC-Bayes bounds for both permuted pixels and labels experiments. The performance of our classic bounds is significantly better than the existing bounds. Our fast-rate bound achieves competitive performance on novel tasks. For permuted labels, Figure 1a compares the average training error, and Figure 1b shows the test error of learning a new task for different bounds. As shown in Figure 1a, the training error of our classic bound (22) is comparable with other bounds. However, in Figure 1b, for new tasks, the performance of our bound is much better than other bounds. Figure 2 compares the test error when the larger number of training examples is available. Again, our classic bound has better performance.

7.1. Conclusion

In this paper, for meta-learning setup, we have derived a general PAC-Bayes bound which can recover existing known bounds and proposes new bounds. Based on our extended PAC-Bayes bound, we have obtained a bound from the fast-rate family and also a bound from the classic family. The fast-rate bound yields to competitive experimental results on novel tasks with respect to existing methods. Unlike existing bound, to obtain the classic bound, we used only one Markov's inequality and by lower bounding the sum of environment-level and task-level convex functions, we end up with a new classic bound. Practical examples show that the new obtained classic bound reduces the meta overfitting problem. The main property of the new classic bound is that it is expressed in terms of one square. Thus, minimizing the new PAC-Bayes bound leads to a simpler optimization problem, i.e., minimizing an objective function which is linear with respect to KL- divergences of posterior and prior distributions. We guess that due to this property, the new proposed bound has better performance on the meta-test set.

Potentially, our general PAC-Bayes bound holds for both bounded and unbounded loss functions, as well as data-dependent or data-free prior distributions. Here, we only focused on data-free priors and bounded loss functions. Generalizing to other scenarios is left to future work.

Acknowledgements

The author is indebted to Prof. Giuseppe Durisi for his intellectual leadership, many helpful suggestions and invaluable criticism.

This work has been funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 893082.

References

Alquier, P. PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 3: 279–304, Sept. 2008.

Alquier, P. User-friendly introduction to PAC-Bayes bounds. arXiv:2110.11216, 2021. URL https://arxiv.org/abs/2110.11216.

Alquier, P., Ridgway, J., Chopin, N., and Teh, Y. W. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17: 1–41, Dec. 2016.

Amit, R. and Meir, R. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *Proc. Int. Conf. on Machine Learning (ICML)*, Stockholm, Sweden, July 2018.

Balcan, M.-F., Khodak, M., and Talwalkar, A. Provable guarantees for gradient-based meta-learning. In *Proc. Int. Conf. Machine Learning (ICML)*, Long Beach, CA, USA, June 2019.

Baxter, J. A model of inductive bias learning. *Journal of Artif. Intell. Research*, 12:149–198, Mar. 2000.

Bu, Y., Zou, S., and Veeravalli, V. V. Tightening mutual information based bounds on generalization error. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, July 2019.

Catoni, O. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56. IMS Lecture Notes Monogr.Ser., 2007.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. Learning to learn around a common mean. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, Canada, Dec. 2018.

Denevi, G., Ciliberto, C., Grazzi, R., and Pontil, M. Learning-to-learn stochastic gradient descent with biased regularization. In *Proc. Int. Conf. on Machine Learning (ICML)*, Long Beach, CA, USA, June 2019.

Dziugaite1, G., Hsu, K., Gharbieh, W., and Roy, D. M. On the role of data in PAC-Bayes bounds. In *Proc. Artif. Intell. Statist. (AISTATS)*, Apr. 2021.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. Int. Conf. Machine Learning (ICML)*, Sydney, Australia, Aug. 2017.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proc. Artif. Intell. Statist. (AISTATS)*, Chia Laguna Resort, Sardinia, Italy, May 2010.

Guan, J., Lu, Z., and Liu, Y. Improved generalization risk bounds for meta-learning with PAC-Bayes-kl analysis. In *The Int. Conf. on Learning Representations (ICLR)*, Apr. 2022.

Guedj, B. A Primer on PAC-Bayesian Learning. arXiv: 1901.05353, 2019. URL http://arxiv.org/abs/1901.05353.

Guedj, B. and Pujol, L. Still no free lunches: the price to pay for tighter PAC-Bayes bounds. arXiv:1910.04460, 2019. URL http://arxiv.org/abs/1910.04460.

Jose, S. T. and Simeone, O. Information-theoretic generalization bounds for meta-learning and applications. *Entropy*, 23(1), Jan. 2021.

Khodak, M., Balcan, M.-F. F., and Talwalkar, A. S. Adaptive gradient-based meta-learning methods. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2019.

Konobeev, M., Kuzborskij, I., and Szepesvári, C. On Optimality of Meta-Learning in Fixed-Design Regression with Weighted Biased Regularization. arXiv:2011.00344, 2020. URL https://arxiv.org/abs/2011.00344.

Li, Z., Zhou, F., Chen, F., and Li, H. Meta-SGD: Learning to learn quickly for few-shot learning. arXiv:1707.09835, 2017. URL http://arxiv.org/abs/1707.09835.

Liu, T., Lu, J., Yan, Z., and Zhang, G. PAC-Bayes bounds for meta-learning with data-dependent prior. arXiv:2102.03748, 2021. URL https://arxiv.org/abs/2102.03748.

Maurer, A. A note on the PAC-Bayesian theorem. arXiv:0411099, 2004. URL https://arxiv.org/abs/0411099.

McAllester, D. PAC-Bayesian tutorial with a dropout bound. arXiv:1307.2118, 2013. URL https://arxiv.org/abs/1307.2118.pdf.

McAllester, D. A. PAC-Bayesian model averaging. In *Proc. Conf. on Learning Theory (COLT)*, New York, NY, USA, July 1999.

Mcallester, D. A. PAC-Bayesian stochastic model selection. In *Machine Learning*, volume 51, pp. 5–21, Apr. 2003.

Mitchell, T. M. *Machine Learning*. McGraw-Hill, New York, 1997. ISBN 978-0-07-042807-2.

Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A., and Roy, D. M. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2019.

Ohnishi, Y. and Honorio, J. Novel change of measure inequalities with applications to PAC-Bayesian bounds and monte carlo estimation. In *Proc. Artif. Intell. Statist. (AISTATS)*, Apr. 2021.

Pentina, A. and Lampert, C. A PAC-Bayesian bound for lifelong learning. In *Proc. Int. Conf. on Machine Learning (ICML)*, Beijing, China, June 2014.

Ravi, S. and Larochelle, H. Optimization as a model for few-shot learning. In *Proc. Int. Conf. Learning Representations, (ICLR)*, Toulon, France, Apr. 2017.

Rezazadeh, A., Sharu, S. T., Durisi, G., and Simeone, O. Conditional mutual information-based generalization bound for meta learning. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Melbourne, Australia, July 2021.

Rivasplata, O., Kuzborskij, I., and Szepesvári, C. PAC-Bayes analysis beyond the usual bounds. arXiv:2006.13057, 2020. URL https://arxiv.org/abs/2006.13057.

Rothfuss, J., Fortuin, V., and Krause, A. PACOH: Bayes-optimal meta-learning with PAC-guarantees. In *Proc. Int. Conf. on Machine Learning (ICML)*, Vienna, Austria, Feb. 2021.

Russo, D. and Zou, J. Controlling bias in adaptive data analysis using information theory. In *Proc. Artif. Intell. Statist. (AISTATS)*, Cadiz, Spain, May 2016.

Seeger, M. PAC-Bayesian generalisation error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, Oct. 2002.

Thiemann, N., Igel, C., Wintenberger, O., and Seldin, Y. A strongly quasiconvex pac-bayesian bound. In Hanneke, S. and Reyzin, L. (eds.), *Proc. of the Int. Conf. on Algorithmic Learning Theory (ALT)*, pp. 466–492, Oct. 2017.

Thrun, S. and Pratt, L. *Learning to Learn: Introduction and Overview*, pp. 3–17. Springer, Boston, MA, 1998. ISBN 978-1-4615-5529-2.

Wainwright, M. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Univ. Press, Cambridge, U.K, 2019. ISBN 9781108498029.

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. Learning transferable architectures for scalable image recognition. In *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, USA, June 2018. doi: 10.1109/CVPR.2018.00907.

A. Proof of Theorem 3.1

To bound the meta-generalization gap, we bound the generalization gap at task and environment levels, separately. At the task level, for the task t_i, the base-learner uses a prior and the samples $\boldsymbol{Z}_i^{\mathrm{M}}$ to output a distribution over hypotheses. Here, we consider the prior over hypothesis $(\mathcal{P}, \mathrm{P})$ as a joint distribution of one hyper-prior $\mathcal{P}$ and the prior P depends on the hyper-prior. Note that the posterior over the hypothesis can be any distribution, particularly a tuple $(\mathcal{Q}, \mathrm{Q}_i)$ where firstly the hyperparameter U is sampled from the hyper-posterior $\mathcal{Q}$, and then the model parameter W is sampled from Q_i. Considering this approach, for any $\theta_{\text{tsk}} \geq 0$, we have

$$\begin{aligned}
&\theta_{\text{tsk}} \mathrm{F}^{\text{Task}}\left(\mathbb{E}_{\mathcal{Q}}\mathbb{E}_{\mathrm{Q}_i}\left(\mathrm{L}_{\mathrm{P}_{Z|t_i}}(W)\right), \mathbb{E}_{\mathcal{Q}}\mathbb{E}_{\mathrm{Q}_i}\left(\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right)\right)\\
&\quad\leq \theta_{\text{tsk}}\mathbb{E}_{\mathcal{Q}}\mathbb{E}_{\mathrm{Q}_i}\left[\mathrm{F}^{\text{Task}}\left(\mathrm{L}_{\mathrm{P}_{Z|t_i}}(W), \mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right)\right] &(29)\\
&\quad\leq \mathrm{D}\left(\mathcal{Q}\mathrm{Q}_i||\mathcal{P}\mathrm{P}\right) + \log\left(\mathbb{E}_{\mathcal{P}\mathrm{P}} e^{\theta_{\text{tsk}}\mathrm{F}^{\text{Task}}\left(\mathrm{L}_{\mathrm{P}_{Z|t_i}}(W), \mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right)}\right) &(30)\\
&\quad= \mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) + \mathbb{E}_{\mathcal{Q}}\left[\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right] + \log\left(\mathbb{E}_{\mathcal{P}\mathrm{P}} e^{\theta_{\text{tsk}}\mathrm{F}^{\text{Task}}\left(\mathrm{L}_{\mathrm{P}_{Z|t_i}}(W), \mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right)}\right), &(31)
\end{aligned}$$

where $\tilde{\mathrm{L}}^{t_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(u)$ and $\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(w)$ are defined in (10) (2), respectively. Since $\mathrm{F}^{\text{Task}}(\cdot)$ is convex, in (29) we applied Jensen's inequality, and (30) follows from the Donsker-Varadhan theorem (99). Finally, (31) follows from the definition of the KL-divergence.

Next, we average both sides of (31) over N tasks. Recalling that $\mathrm{F}^{\text{Task}}(a, b)$ is convex in both a and b, we have $\mathrm{F}^{\text{Task}}\left(\frac{1}{\mathrm{N}}\sum_i^{\mathrm{N}} a_i, \frac{1}{\mathrm{N}}\sum_i^{\mathrm{N}} b_i\right) \leq \frac{1}{\mathrm{N}}\sum_i^{\mathrm{N}} \mathrm{F}^{\text{Task}}(a_i, b_i)$. By applying this fact, in view of (9) and (7), using $\log\prod_i a_i = \sum_i \log a_i$, we find that

$$\begin{aligned}
\theta_{\text{tsk}}\mathrm{F}^{\text{Task}}\left(\mathbb{E}_{U\sim\mathcal{Q}}\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{t_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)\right), \mathbb{E}_{U\sim\mathcal{Q}}\mathrm{L}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}(U)\right) &\leq \mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) + \frac{1}{\mathrm{N}}\mathbb{E}_{\mathcal{Q}}\left(\sum_{i=1}^{\mathrm{N}}\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right)\\
&\quad+\frac{1}{\mathrm{N}}\log\underbrace{\left(\prod_{i=1}^{\mathrm{N}}\mathbb{E}_{\mathcal{P}\mathrm{P}} e^{\theta_{\text{tsk}}\mathrm{F}^{\text{Task}}\left(\mathrm{L}_{\mathrm{P}_{Z|t_i}}(W), \mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right)}\right)}_{\Upsilon_{\text{tsk}}}. \qquad (32)
\end{aligned}$$

Similarly, at the environment level, by setting hyper-prior and hyper-posterior as $\mathcal{P}$ and $\mathcal{Q}$, respectively, using Jensen's inequality, and applying the Donsker-Varadhan theorem (99), for $\theta_{\text{env}} \geq 0$ we have

$$\begin{aligned}
&\theta_{\text{env}}\mathrm{F}^{\text{Env}}\left(\mathbb{E}_{U\sim\mathcal{Q}}\left(\mathrm{L}_{\mathrm{P}_{\boldsymbol{TZ}^{\mathrm{M}}}}(U)\right), \mathbb{E}_{U\sim\mathcal{Q}}\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{t_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)\right)\right)\\
&\qquad\leq \mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) + \log\underbrace{\left(\mathbb{E}_{\mathcal{P}} e^{\theta_{\text{env}}\mathrm{F}^{\text{Env}}\left(\mathrm{L}_{\mathrm{P}_{\boldsymbol{TZ}^{\mathrm{M}}}}(U), \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{t_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)\right)}\right)}_{\Upsilon_{\text{env}}}. \qquad (33)
\end{aligned}$$

Now, dividing both sides of (32) (resp. (33)) by θ_{tsk} (resp. θ_{env}), summing up both sides of the obtained inequalities, and using the fact that $\log(a) + \log(b) = \log(a.b)$, we finally obtain

$$\begin{aligned}
&\mathrm{F}^{\text{Env}}\left(\mathbb{E}_{U\sim\mathcal{Q}}\left(\mathrm{L}_{\mathrm{P}_{\boldsymbol{TZ}^{\mathrm{M}}}}(U)\right), \mathbb{E}_{U\sim\mathcal{Q}}\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{t_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)\right)\right) + \mathrm{F}^{\text{Task}}\left(\mathbb{E}_{U\sim\mathcal{Q}}\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{t_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)\right), \mathbb{E}_{U\sim\mathcal{Q}}\mathrm{L}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}(U)\right)\\
&\qquad\leq \frac{1}{\mathrm{N}\cdot\theta_{\text{tsk}}}\mathbb{E}_{\mathcal{Q}}\left(\sum_{i=1}^{\mathrm{N}}\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right) + \left(\frac{1}{\theta_{\text{tsk}}} + \frac{1}{\theta_{\text{env}}}\right)\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) + \log\left(\Upsilon_{\text{tsk}}^{\frac{1}{\mathrm{N}\cdot\theta_{\text{tsk}}}}\cdot\Upsilon_{\text{env}}^{\frac{1}{\theta_{\text{env}}}}\right). \qquad (34)
\end{aligned}$$

Finally, by applying the Markov's inequality, i.e., $\mathbb{P}[\Upsilon \geq \mathbb{E}[\Upsilon]/\delta] \leq \delta$ to the $\Upsilon_{\text{tsk}}^{\frac{1}{\theta_{\text{N}\cdot\text{tsk}}}} \cdot \Upsilon_{\text{env}}^{\frac{1}{\theta_{\text{env}}}}$ term, from (34), we conclude that with probability at least $1-\delta$

$$\mathrm{F}^{\text{Env}}\left(\mathbb{E}_{U\sim\mathcal{Q}}\left(\mathrm{L}_{\mathrm{P}_{\boldsymbol{T}\boldsymbol{Z}^{\mathrm{M}}}}(U)\right), \mathbb{E}_{U\sim\mathcal{Q}}\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}_{\boldsymbol{Z}_i^{\mathrm{M}}}^{T_i}(U)\right)\right) + \mathrm{F}^{\text{Task}}\left(\mathbb{E}_{U\sim\mathcal{Q}}\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}_{\boldsymbol{Z}_i^{\mathrm{M}}}^{T_i}(U)\right), \mathbb{E}_{U\sim\mathcal{Q}}\mathrm{L}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}(U)\right)$$
$$\leq \frac{1}{\mathrm{N}\cdot\theta_{\text{tsk}}}\mathbb{E}_{\mathcal{Q}}\left(\sum_{i=1}^{\mathrm{N}}\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right) + \left(\frac{1}{\theta_{\text{tsk}}}+\frac{1}{\theta_{\text{env}}}\right)\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) + \log\frac{\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}}\left(\Upsilon_{\text{tsk}}^{\frac{1}{\mathrm{N}\cdot\theta_{\text{tsk}}}}\cdot\Upsilon_{\text{env}}^{\frac{1}{\theta_{\text{env}}}}\right)}{\delta}, \tag{35}$$

which proves (13).

Next, to prove (16), we apply Markov's inequality to both (31) and (33). From (31), we find that with probability at least $1-\delta_i$ under distribution $\mathrm{P}_{T_{1:\mathrm{N}}}\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}$, we have

$$\theta_{\text{tsk}}\mathrm{F}^{\text{Task}}\left(\mathbb{E}_{\mathcal{Q}}\mathbb{E}_{\mathrm{Q}_i}\left(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)\right), \mathbb{E}_{\mathcal{Q}}\mathbb{E}_{\mathrm{Q}_i}\left(\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right)\right)$$
$$\leq \mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) + \mathbb{E}_{\mathcal{Q}}\left[\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right] + \log\left(\frac{\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}}\mathbb{E}_{\mathcal{P}\mathrm{P}}e^{\theta_{\text{tsk}}\mathrm{F}^{\text{Task}}\left(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W),\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right)}}{\delta_i}\right). \tag{36}$$

Recalling that from $\mathrm{F}^{\text{Task}}(a,b) \leq c_{\text{tsk}}$, we can conclude $a \leq k_t \cdot b + \mathrm{G}_{\text{Task}}(c_{\text{tsk}})$, from (36), by dividing both sides of $a \leq k_t \cdot b + \mathrm{G}_{\text{Task}}(c_{\text{tsk}})$ by N, with probability at least $1-\delta_i$, we have

$$\frac{1}{\mathrm{N}}\mathbb{E}_{\mathcal{Q}}\mathbb{E}_{\mathrm{Q}_i}\left(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)\right) \leq \frac{k_t}{\mathrm{N}}\mathbb{E}_{\mathcal{Q}}\mathbb{E}_{\mathrm{Q}_i}\left(\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right) + \frac{1}{\mathrm{N}}\mathrm{G}_{\text{Task}}\left(\mathrm{B}_t\right), \tag{37}$$

where

$$\mathrm{B}_t = \frac{1}{\theta_{\text{tsk}}}\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) + \frac{1}{\theta_{\text{tsk}}}\mathbb{E}_{\mathcal{Q}}\left[\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right] + \frac{1}{\theta_{\text{tsk}}}\log\left(\frac{\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}}\mathbb{E}_{\mathcal{P}\mathrm{P}}e^{\theta_{\text{tsk}}\mathrm{F}^{\text{Task}}\left(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W),\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right)}}{\delta_i}\right). \tag{38}$$

Here, in Lemma F.1, we set f_i as the left hand side of (37) and a_i as the right hand side of (37). Thus, from Lemma F.1, we conclude that with probability at least $1-\sum_i \delta_i$,

$$\mathbb{E}_{U\sim\mathcal{Q}}\left[\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathbb{E}_{W\sim\mathrm{Q}_i}\left(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)\right)\right] \leq \frac{k_t}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathbb{E}_{\mathcal{Q}}\mathbb{E}_{\mathrm{Q}_i}\left(\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right) + \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathrm{G}_{\text{Task}}\left(\mathrm{B}_t\right). \tag{39}$$

Finally, in view of (9) and (7), (39) can be written as

$$\mathbb{E}_{\mathcal{Q}}\left[\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}_{\boldsymbol{Z}_i^{\mathrm{M}}}^{T_i}(U)\right] \leq k_t\cdot\mathbb{E}_{\mathcal{Q}}\left[\mathrm{L}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}(U)\right] + \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathrm{G}_{\text{Task}}\left(\mathrm{B}_t\right). \tag{40}$$

Similarly, at the environment level from $\mathrm{F}^{\text{Env}}(a,b) \leq c_{\text{env}}$, we can conclude $a \leq k_e \cdot b + \mathrm{G}_{\text{Env}}(c_{\text{env}})$. Considering this fact, by applying the Markov's inequality to (33), with probability at least $1-\delta_0$, we have

$$\mathbb{E}_{\mathcal{Q}}\big(\mathrm{L}_{\mathrm{P}_{\boldsymbol{T}\boldsymbol{Z}^{\mathrm{M}}}}(U)\big) \leq k_e\cdot\mathbb{E}_{\mathcal{Q}}\Big(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}_{\boldsymbol{Z}_i^{\mathrm{M}}}^{T_i}(U)\Big) + \mathrm{G}_{\text{Env}}\left(\mathrm{B}_e\right), \tag{41}$$

where

$$\mathrm{B}_e = \frac{1}{\theta_{\text{env}}}\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) + \frac{1}{\theta_{\text{env}}}\log\left(\frac{\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}}\mathbb{E}_{\mathcal{P}}e^{\theta_{\text{env}}\mathrm{F}^{\text{Env}}\left(\mathrm{L}_{\mathrm{P}_{\boldsymbol{T}\boldsymbol{Z}^{\mathrm{M}}}}(U),\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}_{\boldsymbol{Z}_i^{\mathrm{M}}}^{T_i}(U)\right)}}{\delta_0}\right). \tag{42}$$

Here, again we use Lemma F.1. In Lemma F.1, we set $\mathrm{N}=2$, f_1 and a_1 as the $k_e \geq 0$ times of the left and right hands side of (40), respectively, and also f_2 and a_2 as the left and right hands side of (41), respectively. Thus, with probability at least $1-\sum_i \delta_i - \delta_0$,

$$\mathbb{E}_{\mathcal{Q}}\left[\mathrm{L}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}(U)\right] \leq k_e \cdot k_t \cdot \mathbb{E}_{\mathcal{Q}}\left[\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)\right] + \mathrm{G}_{\text{Env}}\left(\mathrm{B}_e\right) + \frac{k_e}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \mathrm{G}_{\text{Task}}\left(\mathrm{B}_t\right). \tag{43}$$

Finally, setting $\delta_0 = \frac{\delta}{2}$, $\delta_i = \frac{\delta}{2\mathrm{N}}$ in (43), we conclude (16).

B. Re-obtaining the known PAC-Bayes Bounds

In this section, we present the derivation of bounds, summarized in Table 1. Firstly, to obtain Theorem 2 of (Amit & Meir, 2018), in (16), we set $\theta_{\text{tsk}} = \theta_{\text{env}} = 1$, and $\mathrm{F}^{\text{Task}}(a,b) = 2(\mathrm{M}-1)(a-b)^2$, $\mathrm{F}^{\text{Env}}(a,b) = 2(\mathrm{N}-1)(a-b)^2$. These choices lead to $k_e = k_t = 1$, $\mathrm{G}_{\text{Task}}(c) = \sqrt{\frac{c}{2(\mathrm{M}-1)}}$ and also $\mathrm{G}_{\text{Env}}(c) = \sqrt{\frac{c}{2(\mathrm{N}-1)}}$. To simplify B_{Task} and B_{Env} given by (18) and (17), we use Lemma F.2. Since the prior in independent of the data, by interchanging the order of expectations over $\mathrm{P}_{T_{1:\mathrm{N}}}\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}$ and priors, in view of (18) and (17), we find that

$$\mathbb{E}_{\mathcal{P}\mathrm{P}}\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}}\left[e^{2(\mathrm{M}-1)\left(\mathrm{L}_{\mathrm{P}_{Z|t_i}}(W) - \mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_i}(W)\right)^2}\right] \leq \mathrm{M}, \tag{44}$$

$$\mathbb{E}_{\mathcal{P}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}}\left[e^{2(\mathrm{N}-1)\left(\mathrm{L}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}(U) - \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \tilde{\mathrm{L}}^{t_i}_{\boldsymbol{Z}^{\mathrm{M}}_i}(U)\right)^2}\right] \leq \mathrm{N}, \tag{45}$$

where for (44) (res. (45)), we used Lemma F.2 by setting $\lambda = \frac{2(\mathrm{M}-1)}{\mathrm{M}}$ (resp. $\lambda = \frac{2(\mathrm{N}-1)}{\mathrm{N}}$) and $\sigma = 0.5$. We recall that since in (Amit & Meir, 2018) the loss function is bounded on $[0,1]$, we set $\sigma = 0.5$ in Lemma F.2. Now, inserting (44) and (45) into (18) and (17), from (16) we conclude that

$$\mathbb{E}_{U\sim\mathcal{Q}}\left(\mathrm{L}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}(U)\right) - \mathbb{E}_{U\sim\mathcal{Q}}\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U) \leq \sqrt{\frac{\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) + \log\left(\frac{2\mathrm{N}}{\delta}\right)}{2(\mathrm{N}-1)}} + \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\sqrt{\frac{\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) + \mathbb{E}_{\mathcal{Q}}\left[\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right] + \log\frac{2\mathrm{NM}}{\delta}}{2(\mathrm{M}-1)}}, \tag{46}$$

which is the same as the bound presented in Theorem 2 of (Amit & Meir, 2018).

Next, to obtain Theorem 2 of (Rothfuss et al., 2021), in (13) we set $\mathrm{F}^{\text{Env}}(a,b) = \mathrm{F}^{\text{Task}}(a,b) = (a-b)$, $\theta_{\text{tsk}} = \beta$, and $\theta_{\text{env}} = \lambda$, and these choices leads to $k_t = k_e = 1$ and $\mathrm{G}_{\text{Task}}(c) = \mathrm{G}_{\text{Env}}(c) = c$. To simplify the log-term, since the prior in independent of the data, by interchanging the order of expectations over $\mathrm{P}_{T_{1:\mathrm{N}}}\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}$ and priors, and recalling that the loss function is bounded on the interval $[0,1]$, and hence it is sub-Gaussian with parameter $\sigma = 0.5$, we can conclude

$$\mathbb{E}_{U\sim\mathcal{Q}}\left(\mathrm{L}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}(U)\right) - \mathbb{E}_{U\sim\mathcal{Q}}\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U) \leq \min_{\beta,\lambda\geq 0}\left(\frac{\lambda}{8\mathrm{N}} + \frac{\lambda}{8\mathrm{M}}\right) - \frac{1}{\sqrt{\mathrm{N}}}\log\delta + \frac{1}{\beta}\mathbb{E}_{\mathcal{Q}}\left(\sum_{i=1}^{\mathrm{N}}\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right) + \left(\frac{1}{\beta} + \frac{1}{\lambda}\right)\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right), \tag{47}$$

which is the same as Theorem 2 of (Rothfuss et al., 2021).

Next, we re-obtain Theorem 1 of (Liu et al., 2021). Firstly, in (16), we set $\theta_{\text{tsk}} = \theta_{\text{env}} = 1$, and $\mathrm{F}^{\text{Task}}(a,b) = \mathrm{M}kl(a,b)$ and $\mathrm{F}^{\text{Env}}(a,b) = 2(\mathrm{N}-1)(a-b)^2$. These choices lead to $k_e = 1$ $\mathrm{G}_{\text{Env}}(c) = \sqrt{\frac{c}{2(\mathrm{N}-1)}}$, and using further relaxation, from $\mathrm{M}kl(a,b) \leq c_{\text{tsk}}$, it can be proved that $a \leq b/(1-0.5\lambda) + c_{\text{tsk}}/(\mathrm{M}.\lambda(1-0.5\lambda))$ for $\lambda \in (0,2)$ (Thiemann et al., 2017). Thus, we have k_t and $\mathrm{G}_{\text{Task}}(c) = c/(\mathrm{M}\lambda(1-0.5\lambda))$. It remains to obtain the log-terms of B_{Task} and B_{Env} given by (18) and (17). Again, assuming the prior in independent of the data, by interchanging the order of expectations over $\mathrm{P}_{T_{1:\mathrm{N}}}\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}$

and priors, we find (45), and

$$\mathbb{E}_{\mathcal{P}\mathrm{P}}\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}} e^{\mathrm{M}kl\left(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W),\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)\right)} \leq 2\sqrt{\mathrm{M}}, \tag{48}$$

where in (48), we used Lemma F.5. Applying all these facts to (16), we obtain

$$\begin{aligned}\mathbb{E}_{U\sim\mathcal{Q}}\left[\mathrm{L}_{\mathrm{P}_{\boldsymbol{T}\boldsymbol{Z}^{\mathrm{M}}}}(U)\right] \leq \frac{1}{(1-0.5\lambda)}\mathbb{E}_{U\sim\mathcal{Q}}\left[\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)\right] + \sqrt{\frac{\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right)+\log\frac{2\mathrm{N}}{\delta}}{2(\mathrm{N}-1)}} \\ +\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\frac{\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right)+\mathbb{E}_{\mathcal{Q}}\left[\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right]+\log\frac{4\mathrm{N}\sqrt{\mathrm{M}}}{\delta}}{(\mathrm{M}\lambda(1-0.5\lambda))},\end{aligned} \tag{49}$$

which is the same as Theorem 1 of (Liu et al., 2021).

Similar approach can be applied to the bounds presented in (Guan et al., 2022). For the three bounds considered in (Guan et al., 2022), KL-divergence is chosen for both task level and environment level. To bound the log-terms of (18) and (17), we need to use Lemma 2 of (Guan et al., 2022). For the affine transformation steps, at the environment level, we use Pinsker's inequality.

C. Proof Theorem 5.1

To prove Theorem 5.1, in (16), we set $\theta_{\text{tsk}} = \theta_{\text{env}} = 1$, $\mathrm{F}^{\text{Env}}(a,b) = \mathrm{ND}_\gamma(b,a)$ and $\mathrm{F}^{\text{Task}}(a,b) = \mathrm{MD}_\gamma(b,a)$. Using Lemma F.7, from $\mathrm{ND}_\gamma(b,a) \leq c_e$, we conclude that for $\gamma \in (-2,0)$, $a \leq b/(1+0.5\gamma) - c_e/(\mathrm{N}.\gamma(1+0.5\gamma))$. In other words, $k_e = 1/(1+0.5\gamma)$ and $\mathrm{G}_{\text{Env}}(c) = \frac{-c}{\mathrm{N}\gamma(1+0.5\gamma)}$ (similarly for the task level). It remains to determine the log-terms appeared in B_{Task} and B_{Env} given by (18) and (17), respectively. Since the prior is independent of the data, by interchanging the order of expectations over $\mathrm{P}_{T_{1:\mathrm{N}}}\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}$ and priors, using Lemma F.6, in view of (18) and (17), we find that

$$\mathbb{E}_{\mathcal{P}}\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}} e^{\mathrm{ND}_\gamma\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U),\mathrm{L}_{\mathrm{P}_{\boldsymbol{T}\boldsymbol{Z}^{\mathrm{M}}}}(U)\right)} \leq 1, \tag{50}$$

$$\mathbb{E}_{\mathcal{P}\mathrm{P}}\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}} e^{\mathrm{MD}_\gamma\left(\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W),\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)\right)} \leq 1. \tag{51}$$

Applying all these facts to (16), we find that

$$\begin{aligned}\mathbb{E}_{U\sim\mathcal{Q}}\left[\mathrm{L}_{\mathrm{P}_{\boldsymbol{T}\boldsymbol{Z}^{\mathrm{M}}}}(U)\right] \leq \frac{1}{(1+0.5\gamma_e)}\cdot\frac{1}{(1+0.5\gamma_t)}\cdot\mathbb{E}_{U\sim\mathcal{Q}}\left[\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)\right] - \frac{\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right)+\log\frac{2}{\delta}}{\mathrm{N}\gamma_e(1+0.5\gamma_e)} \\ -\frac{1}{\mathrm{N}(1+0.5\gamma_e)}\sum_{i=1}^{\mathrm{N}}\frac{\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right)+\mathbb{E}_{\mathcal{Q}}\left[\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right]+\log\frac{2\mathrm{N}}{\delta}}{\mathrm{M}\gamma_t(1+0.5\gamma_t)}.\end{aligned} \tag{52}$$

Setting $\lambda_e = -1/\gamma_e$ and $\lambda_t = -1/\gamma_t$, we conclude the proof.

D. Proof of Theorem 5.2

Setting $\theta_{\text{tsk}} = \theta_{\text{env}} = 1$, $\mathrm{F}^{\text{Task}}(a,b) = 2(\mathrm{M}-1)(a-b)^2$ and $\mathrm{F}^{\text{Env}}(a,b) = (\mathrm{N}-1)(a-b)^2$ (13), leads to

$$\begin{aligned}(\mathrm{N}-1)\left(\mathbb{E}_{U\sim\mathcal{Q}}\left(\mathrm{L}_{\mathrm{P}_{\boldsymbol{T}\boldsymbol{Z}^{\mathrm{M}}}}(U)\right)-\mathbb{E}_{U\sim\mathcal{Q}}\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)\right)\right)^2 \\ +2(\mathrm{M}-1)\left(\mathbb{E}_{U\sim\mathcal{Q}}\big(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)\big)-\mathbb{E}_{U\sim\mathcal{Q}}\big(\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)\big)\right)^2 \\ \leq \mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right)+\frac{1}{\mathrm{N}}\mathbb{E}_{\mathcal{Q}}\left(\sum_{i=1}^{\mathrm{N}}\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right)+\log\frac{\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}}\left(\Upsilon_{\text{tsk}}^{\frac{1}{\mathrm{N}}}\cdot\Upsilon_{\text{env}}\right)}{\delta}.\end{aligned} \tag{53}$$

Then, following the same steps to obtain (20), we can show that

$$\frac{2(\mathrm{M}-1)(\mathrm{N}-1)}{2(\mathrm{M}-1)+(\mathrm{N}-1)}\left(\mathbb{E}_{U\sim\mathcal{Q}}\left(\mathrm{L}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}(U)\right)-\mathbb{E}_{U\sim\mathcal{Q}}\left(\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)\right)\right)^2$$
$$\leq(\mathrm{N}-1)\left(\mathbb{E}_{U\sim\mathcal{Q}}\left(\mathrm{L}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}(U)\right)-\mathbb{E}_{U\sim\mathcal{Q}}\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}^{\mathrm{M}}_i}(U)\right)\right)^2$$
$$+2(\mathrm{M}-1)\left(\mathbb{E}_{\mathcal{Q}}\left(\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}^{\mathrm{M}}_i}(U)\right)-\mathbb{E}_{\mathcal{Q}}\left(\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)\right)\right)^2, \tag{54}$$

and hence

$$\frac{2(\mathrm{M}-1)(\mathrm{N}-1)}{2(\mathrm{M}-1)+(\mathrm{N}-1)}\left(\mathbb{E}_{U\sim\mathcal{Q}}\left(\mathrm{L}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}(U)\right)-\mathbb{E}_{U\sim\mathcal{Q}}\left(\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)\right)\right)^2$$
$$\leq \mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right)+\frac{1}{\mathrm{N}}\mathbb{E}_{\mathcal{Q}}\left(\sum_{i=1}^{\mathrm{N}}\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right)+\log\frac{\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}}\left(\Upsilon_{\text{tsk}}^{\frac{1}{\mathrm{N}}}\cdot\Upsilon_{\text{env}}\right)}{\delta}. \tag{55}$$

Now, the log-term appeared in (55) can be bounded as

$$\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}|\mathrm{P}_{T_1:T_\mathrm{N}}}\left(\Upsilon_{\text{tsk}}^{\frac{1}{\mathrm{N}}}\cdot\Upsilon_{\text{env}}\right)\leq\sqrt{\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}|\mathrm{P}_{T_1:T_\mathrm{N}}}\left(\Upsilon_{\text{tsk}}^{\frac{1}{\mathrm{N}}}\right)^2\cdot\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}|\mathrm{P}_{T_1:T_\mathrm{N}}}\left(\Upsilon_{\text{env}}\right)^2}, \tag{56}$$

where in (56), we applied Cauchy-Schwartz inequality (or Hölder's inequality). Next, by setting $\kappa=2$, $\theta_{\text{tsk}}=\theta_{\text{env}}=1$, $\mathrm{F}^{\text{Task}}\left(a,b\right)=2(\mathrm{M}-1)(a-b)^2$ and $\mathrm{F}^{\text{Env}}\left(a,b\right)=(\mathrm{N}-1)(a-b)^2$, in (14) and (15), we have

$$\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}|\mathrm{P}_{T_1:T_\mathrm{N}}}\left(\Upsilon_{\text{tsk}}^{\frac{1}{\mathrm{N}}}\right)^2=\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}|\mathrm{P}_{T_1:T_\mathrm{N}}}\left(\prod_{i=1}^{\mathrm{N}}\mathbb{E}_{\mathcal{P}\mathrm{P}}e^{\theta_{\text{tsk}}\mathrm{F}^{\text{Task}}\left(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W),\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_i}(W)\right)}\right)^{\frac{2}{\mathrm{N}}} \tag{57}$$
$$\leq\left(\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}|\mathrm{P}_{T_1:T_\mathrm{N}}}\prod_{i=1}^{\mathrm{N}}\mathbb{E}_{\mathcal{P}\mathrm{P}}e^{2(\mathrm{M}-1)\left(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)-\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_i}(W)\right)^2}\right)^{\frac{2}{\mathrm{N}}} \tag{58}$$
$$=\left(\prod_{i=1}^{\mathrm{N}}\mathbb{E}_{\mathcal{P}\mathrm{P}}\mathbb{E}_{\mathrm{P}_{T_i\boldsymbol{Z}^{\mathrm{M}}_i}}e^{2(\mathrm{M}-1)\left(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)-\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_i}(W)\right)^2}\right)^{\frac{2}{\mathrm{N}}} \tag{59}$$
$$\leq\left(\prod_{i=1}^{\mathrm{N}}\mathrm{M}\right)^{\frac{2}{\mathrm{N}}}=\mathrm{M}^2 \tag{60}$$

where in (57) we applied (15). In (58), since $a^{2/\mathrm{N}}$ is a concave function for $\mathrm{N}\geq 2$, we used Jensen's inequality. Since, tasks are assumed to be independent, and the prior is independent of the data, by interchanging the order of expectations over $\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}_i}$ and $\mathcal{P}\mathrm{P}$, we obtained (59). Finally, in (60), we used Lemma F.4. Recalling that the loss function is bounded on $[0,1]$, we face with sub-Gaussian variables with parameter $\sigma=0.5$. By setting $\lambda_{\text{Tsk}}=2(\mathrm{M}-1)/\mathrm{M}$ (where $\lambda_{\text{Tsk}}\leq 1/2\sigma^2$) in Lemma F.4, and recalling that $\sigma=0.5$, from (80), we found (60).

Similarly, inserting (14) into (85) , we find that

$$\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}|\mathrm{P}_{T_1:T_\mathrm{N}}}\left(\Upsilon_{\text{env}}\right)^2 = \mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}|\mathrm{P}_{T_1:T_\mathrm{N}}}\left(\mathbb{E}_{\mathcal{P}}e^{\theta_{\text{env}}\mathrm{F}^{\text{Env}}\left(\mathrm{L}_{\mathrm{P}_{\boldsymbol{TZ}^{\mathrm{M}}}}(U),\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}^{\mathrm{M}}_i}(U)\right)}\right)^2 \tag{61}$$

$$\leq \mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}|\mathrm{P}_{T_1:T_\mathrm{N}}}\mathbb{E}_{\mathcal{P}}\left(e^{(\mathrm{N}-1)\left(\mathrm{L}_{\mathrm{P}_{\boldsymbol{TZ}^{\mathrm{M}}}}(U)-\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}^{\mathrm{M}}_i}(U)\right)}\right)^2 \tag{62}$$

$$= \mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}|\mathrm{P}_{T_1:T_\mathrm{N}}}\mathbb{E}_{\mathcal{P}}e^{2(\mathrm{N}-1)\left(\mathrm{L}_{\mathrm{P}_{\boldsymbol{TZ}^{\mathrm{M}}}}(U)-\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}^{\mathrm{M}}_i}(U)\right)} \tag{63}$$

$$= \mathbb{E}_{\mathcal{P}}\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}|\mathrm{P}_{T_1:T_\mathrm{N}}}e^{2(\mathrm{N}-1)\left(\mathrm{L}_{\mathrm{P}_{\boldsymbol{TZ}^{\mathrm{M}}}}(U)-\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}^{\mathrm{M}}_i}(U)\right)} \tag{64}$$

$$\leq \mathrm{N} \tag{65}$$

where in (62), since a^2 is a convex function, we applied Jensen's inequality. In (63), we used the fact that $\exp(a)^b = \exp(a.b)$. Since the prior is independent of the data, by interchanging the order of expectations over $\mathrm{P}_{T_{1:\mathrm{N}}}\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}$ and prior, we obtained (64). Finally, in (65) we used Lemma F.4. By setting $\lambda_{\text{Env}} = 2(\mathrm{N}-1)/\mathrm{N}$ (where $\lambda_{\text{Env}} \leq 1/2\sigma^2$) in Lemma F.4, and recalling that $\sigma = 0.5$, from (81), we found (65).

Inserting (60) and (65) into (56), we have

$$\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}|\mathrm{P}_{T_1:T_\mathrm{N}}}\left(\Upsilon_{\text{tsk}}^{\frac{1}{\mathrm{N}}}\cdot\Upsilon_{\text{env}}\right) \leq \sqrt{\mathrm{M}^2\mathrm{N}} = \mathrm{M}\sqrt{\mathrm{N}}. \tag{66}$$

Next, we focus on the affine transformation. Since from $2(\mathrm{M}-1)(a-b)^2 \leq c_{\text{tsk}}$ (resp. $(\mathrm{N}-1)(a-b)^2 \leq c_{\text{env}}$), we can conclude that $a \leq b + \sqrt{c_{\text{tsk}}/(2(\mathrm{M}-1))}$ (resp. $a \leq b + \sqrt{c_{\text{env}}/(\mathrm{N}-1)}$).

Inserting (66) into (55), and applying affine transformation, we conclude the proof.

E. Presenting New PAC-Bayes Bounds

Theorem E.1. *Under the setting of Theorem 3.1, for $k \in \mathbb{N} = \{1, 2, ...\}$ and $\mathrm{N} \geq 2$, the meta-generalization gap is bounded by*

$$\left|\mathbb{E}_{U\sim\mathcal{Q}}\left[\mathrm{L}_{\mathrm{P}_{\boldsymbol{TZ}^{\mathrm{M}}}}(U) - \mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)\right]\right| \leq \sqrt{\frac{(\mathrm{N}-\mathrm{N}^{\frac{1}{2k}}) + 2(\mathrm{M}-\mathrm{M}^{\frac{1}{2k}})}{2(\mathrm{N}-\mathrm{N}^{\frac{1}{2k}})(\mathrm{M}-\mathrm{M}^{\frac{1}{2k}})}}$$
$$\sqrt{2\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) + \mathbb{E}_{\mathcal{Q}}\left[\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right] + \log\frac{(\sqrt{\mathrm{N}}\cdot\mathrm{M})^{\frac{1}{2}-\frac{1}{4k}}}{\delta}}. \tag{67}$$

Proof. We set $\theta_{\text{env}} = \theta_{\text{tsk}} = 1$, $\mathrm{F}^{\text{Env}}(a,b) = (\mathrm{N}-\mathrm{N}^{\frac{1}{2k}})(b-a)^2$ and $\mathrm{F}^{\text{Task}}(a,b) = 2(\mathrm{M}-\mathrm{M}^{\frac{1}{2k}})(b-a)^2$. To bound the log-term, we use Lemma F.4, and in (82) and (83), we set $\sigma = 0.5$, $\lambda_{\text{Tsk}} = 2 - 2\mathrm{M}^{-1+1/(2k)}$ and $\lambda_{\text{Env}} = 2 - 2\mathrm{N}^{-1+1/(2k)}$. By following exactly the same steps presented in the proof of Theorem 5.2, we conclude the proof. □

Theorem E.2. *Under the setting of Theorem 3.1, for $\mathrm{N} \geq 2$, the meta-generalization gap is bounded by*

$$\left|\mathbb{E}_{U\sim\mathcal{Q}}\left[\mathrm{L}_{\mathrm{P}_{\boldsymbol{TZ}^{\mathrm{M}}}}(U) - \mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)\right]\right| \leq \sqrt{\frac{0.5\mathrm{N}+\mathrm{M}}{0.5\mathrm{N}\cdot\mathrm{M}}}$$
$$\sqrt{\mathrm{D}\left(\mathcal{Q}||\mathcal{P}\right) + \mathbb{E}_{\mathcal{Q}}\left[\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathrm{D}\left(\mathrm{Q}_i||\mathrm{P}\right)\right] + \log\frac{2\sqrt{2}}{\delta}}. \tag{68}$$

Proof. We set $\theta_{\text{env}} = \theta_{\text{tsk}} = 1$, $\mathrm{F}^{\text{Env}}(a,b) = 0.5\mathrm{N}(b-a)^2$ and $\mathrm{F}^{\text{Task}}(a,b) = \mathrm{M}(b-a)^2$. To bound the log-term, we use Lemma F.4, and in (82) and (83), we set $\sigma = 0.5$, $\lambda_{\text{Tsk}} = 1$ and $\lambda_{\text{Env}} = 1$. By following exactly the same steps presented in the proof of Theorem 5.2, we conclude the proof. □

We can apply different $\mathrm{F}^{\mathrm{Env}}(a,b)$ and $\mathrm{F}^{\mathrm{Task}}(a,b)$ functions, and obtain different bound.

F. General Lemmas

In this appendix, we provide a number of general lemmas that will be used throughout the paper.

Lemma F.1. *Let X_i for $i=1,...,\mathrm{N}$ be independent random variables. Suppose that for given $a_i \in \mathbb{R}^+$, and measurable function f_i*

$$\mathbb{P}_{X_i}[f_i(X_i) \geq a_i] \leq \delta_i, \tag{69}$$

where $\delta_i \in [0,1]$. Then,

$$\mathbb{P}_{X_{1:\mathrm{N}}}\left[\sum_i f_i(X_i) \leq \sum_i a_i\right] \geq 1 - \sum_i \delta_i. \tag{70}$$

Proof. Firstly, we show that

$$\underbrace{\left\{(x_1,...,x_\mathrm{N}):\ \sum_i f_i(x_i) \geq \sum_i a_i\right\}}_{\mathcal{A}} \subseteq \underbrace{\bigcup_{i=1}\{x_i:\ f_i(x_i) \geq a_i\}}_{\mathcal{B}}. \tag{71}$$

Let $(x_1,...,x_\mathrm{N}) \notin \mathcal{B}$. It means that for all $i=1,...,\mathrm{N}$, $f_i(x_i) < a_i$ which leads that $\sum_i f_i(x_i) < \sum_i a_i$, or $(x_1,...,x_\mathrm{N}) \notin \mathcal{A}$. Thus, $\mathcal{B}^c \subset \mathcal{A}^c$, or equivalently $\mathcal{A} \subset \mathcal{B}$.

Next, from (71), one can conclude

$$\mathbb{P}_{X_{1:\mathrm{N}}}\left[\sum_i f_i(X_i) \geqslant \sum_i a_i\right] \leq \sum_i \mathbb{P}_{X_{1:\mathrm{N}}}\left[f_i(X_i) \geqslant a_i\right] = \sum_i \mathbb{P}_{X_i}\left[f_i(X_i) \geqslant a_i\right] \leq \sum_i \delta_i \tag{72}$$

where the last inequality follows from (69). The proof can be concluded from (72).

□

Lemma F.2. *Let $X_1,...,X_m$ be independent random variables, and $g:\mathcal{X} \to \mathbb{R}$ be a sub-Gaussian function with parameter σ. Assume $\Delta \triangleq \mathbb{E}[g(X)] - \frac{1}{m}\sum_{k=1}^m g(X_i)$, where for $\epsilon > 0$, we have $\mathbb{P}[\Delta \geq \epsilon] \leq \exp(\frac{-m\epsilon^2}{2\sigma^2})$. Then*

$$\mathbb{E}\left[e^{\lambda m \Delta^2}\right] \leq \frac{1}{1-2\lambda\sigma^2}, \tag{73}$$

for $\lambda \leq \frac{1}{2\sigma^2}$.

Proof. The proof is similar to Lemma 3 of (McAllester, 1999). For completeness, we repeat it again. Let $f^\star$ denotes the density function maximizing $\mathbb{E}[e^{\lambda m \Delta^2}]$ subject to the constraint that $\mathbb{P}[\Delta \geq \epsilon] \leq \exp(\frac{-m\epsilon^2}{2\sigma^2})$. The maximum occurs when $\mathbb{P}_{f^\star}[\Delta \geq \epsilon] = \exp(\frac{-m\epsilon^2}{2\sigma^2})$ leading to

$$f^\star(\Delta) = \frac{m\Delta}{\sigma^2}\exp\left(-\frac{m\Delta^2}{2\sigma^2}\right)\mathbb{1}\{\Delta \geq 0\}. \tag{74}$$

Thus, we have

$$\mathbb{E}\left[e^{\lambda m \Delta^2}\right] \leq \int_0^\infty \exp\left(\lambda m \Delta^2\right)\frac{m\Delta}{\sigma^2}\exp\left(-\frac{m\Delta^2}{2\sigma^2}\right)d\Delta = \frac{1}{1-2\lambda\sigma^2}, \qquad \lambda < \frac{1}{2\sigma^2}. \tag{75}$$

□

Lemma F.3. *Let $X_1, ..., X_m$ be independent random variables. Assume $\Delta = \mathbb{E}[g(X)] - \frac{1}{m}\sum_{k=1}^m g(X_i)$, where $g(\cdot)$ is sub-Gaussian function with parameter σ. Then ,*

$$\mathbb{E}\left[e^{\lambda m \Delta^2}\right] \leq \frac{1}{\sqrt{1-2\lambda\sigma^2}}, \tag{76}$$

for $\lambda \leq \frac{1}{2\sigma^2}$.

Proof. The proof is similar to Theorem 2.6 (Wainwright, 2019). For completeness, we repeat it again. Since $g(\cdot)$ is sub-Gaussian, we have

$$\mathbb{E}\left[e^{\lambda\Delta}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2m}\right). \tag{77}$$

Multiplying both sides of (77) by $\exp(-\frac{\lambda^2\sigma^2}{2sm})$ for $s \in (0,1)$, we find that

$$\mathbb{E}\left[e^{\lambda\Delta - \frac{\lambda^2\sigma^2}{2sm}}\right] \leq \exp\left(\frac{-\lambda^2\sigma^2}{2ms}(-s+1)\right). \tag{78}$$

Next, we take integration with respect to λ. Since (78) is valid for any $\lambda \in \mathbb{R}$, by using Fubini's theorem, we exchange the order of expectation and integration, leading to

$$\mathbb{E}\left[\exp\left(\frac{sm\Delta^2}{2\sigma^2}\right)\right] \leq \frac{1}{\sqrt{1-s}}, \qquad \text{for } 0 < s < 1. \tag{79}$$

By defining $\lambda = \frac{s}{2\sigma^2}$, we conclude the proof. $\square$

Lemma F.4. *Consider $\mathrm{L}_{\mathrm{P}_{Z|t_i}}(w_i)$, $\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(w_i)$, $\mathrm{L}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}(u)$, $\mathrm{L}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}(u)$ and $\tilde{\mathrm{L}}^t_{\boldsymbol{Z}_i^{\mathrm{M}}}(u)$ defined by (1), (2), (6), (7) and (10), respectively. Assume that the loss function $\ell(\cdot,\cdot)$ is bounded on the interval $[0,1]$, and hence it is sub-Gaussian with parameter $\sigma = (b-a)/2$. For $\lambda_{\text{Env}}, \lambda_{\text{Tsk}} \leq 1/2\sigma^2$, and data-free priors we have*

$$\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}}\mathbb{E}_{\mathcal{P}\mathrm{P}}e^{\lambda_{\text{Tsk}}\mathrm{M}(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)-\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W))^2} \leq \frac{1}{1-2\lambda_{\text{Tsk}}\sigma^2}, \tag{80}$$

$$\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}}\mathbb{E}_{\mathcal{P}}e^{\lambda_{\text{Env}}\mathrm{N}\left(\mathrm{L}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}(U)-\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)\right)^2} \leq \frac{1}{1-2\lambda_{\text{Env}}\sigma^2}, \tag{81}$$

and also

$$\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}}\mathbb{E}_{\mathcal{P}\mathrm{P}}e^{\lambda_{\text{Tsk}}\mathrm{M}(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)-\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W))^2} \leq \frac{1}{\sqrt{1-2\lambda_{\text{Tsk}}\sigma^2}}, \tag{82}$$

$$\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}}\mathbb{E}_{\mathcal{P}}e^{\lambda_{\text{Env}}\mathrm{N}\left(\mathrm{L}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}(U)-\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)\right)^2} \leq \frac{1}{\sqrt{1-2\lambda_{\text{Env}}\sigma^2}}. \tag{83}$$

Proof. We recall that since the prior is independent of the data, by interchanging the order of expectations over $\mathrm{P}_{T_{1:\mathrm{N}}}\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}$ and priors To show (80), and (82), we note that $\mathrm{P}_{T_i\boldsymbol{Z}_i^{\mathrm{M}}}$ is the marginal distribution of $\mathrm{P}_{T_{1:\mathrm{N}}}\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}$. Since the priors are data-free, we have

$$\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}|T_{1:\mathrm{N}}}}\mathbb{E}_{\mathcal{P}\mathrm{P}}e^{\lambda_{\text{Tsk}}\mathrm{M}(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)-\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W))^2} = \mathbb{E}_{\mathrm{P}_{T_i\boldsymbol{Z}_i^{\mathrm{M}}}}\mathbb{E}_{\mathcal{P}\mathrm{P}}e^{\lambda_{\text{Tsk}}\mathrm{M}(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)-\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W))^2} \tag{84}$$

$$= \mathbb{E}_{\mathcal{P}\mathrm{P}}\mathbb{E}_{\mathrm{P}_{T_i\boldsymbol{Z}_i^{\mathrm{M}}}}e^{\lambda_{\text{Tsk}}\mathrm{M}(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)-\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W))^2}. \tag{85}$$

Next, we set $\Delta = \mathrm{L}_{\mathrm{P}_{Z|t_i}}(W) - \mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)$, and $m = \mathrm{M}$ in Lemma F.2, and also Lemma F.3, for $\lambda_{\text{Tsk}} \leq \frac{1}{2\sigma^2}$, we respectively conclude that

$$\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}_i^{\mathrm{M}}|T_i}}e^{\lambda_{\text{Tsk}}\mathrm{M}(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)-\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W))^2} \leq \frac{1}{1-2\lambda_{\text{Tsk}}\sigma^2}, \tag{86}$$

$$\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}_i^{\mathrm{M}}|T_i}}e^{\lambda_{\text{Tsk}}\mathrm{M}(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)-\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W))^2} \leq \frac{1}{\sqrt{1-2\lambda_{\text{Tsk}}\sigma^2}}, \tag{87}$$

where by averaging both sides of (86) and (87) over T_i, from (86) and (87), we find that

$$\mathbb{E}_{\mathcal{P}\mathrm{P}}\mathbb{E}_{\mathrm{P}_{T_i \boldsymbol{Z}_i^{\mathrm{M}}}} e^{\lambda_{\text{Tsk}}\mathrm{M}(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)-\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W))^2} \leq \frac{1}{1-2\lambda_{\text{Tsk}}\sigma^2}, , \tag{88}$$

$$\mathbb{E}_{\mathcal{P}\mathrm{P}}\mathbb{E}_{\mathrm{P}_{T_i \boldsymbol{Z}_i^{\mathrm{M}}}} e^{\lambda_{\text{Tsk}}\mathrm{M}(\mathrm{L}_{\mathrm{P}_{Z|T_i}}(W)-\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W))^2} \leq \frac{1}{\sqrt{1-2\lambda_{\text{Tsk}}\sigma^2}}. \tag{89}$$

Inserting the right hand-sides of (88) and (89) into (85), we conclude (80), and (82).

Similarly, to show To show (81), and (83), we use the fact that $\mathrm{L}_{\mathrm{P}_{\boldsymbol{T}\boldsymbol{Z}^{\mathrm{M}}}}(U) = \mathbb{E}_{\mathrm{P}_{\boldsymbol{T}\boldsymbol{Z}^{\mathrm{M}}}}[\tilde{\mathrm{L}}^{T}_{\boldsymbol{Z}^{\mathrm{M}}}(U)]$. Using the fact that prior is data-free, by setting $\Delta = \mathbb{E}_{\mathrm{P}_{\boldsymbol{T}\boldsymbol{Z}^{\mathrm{M}}}}[\tilde{\mathrm{L}}^{T}_{\boldsymbol{Z}^{\mathrm{M}}}(U)] - \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)$, and $m=\mathrm{N}$ in Lemma F.2, and also Lemma F.3, for $\lambda_{\text{Env}} \leq \frac{1}{2\sigma^2}$, we respectively conclude that

$$\mathbb{E}_{\mathcal{P}}\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}} e^{\lambda_{\text{Env}}\mathrm{N}\left(\mathbb{E}_{\mathrm{P}_{\boldsymbol{T}\boldsymbol{Z}^{\mathrm{M}}}}[\tilde{\mathrm{L}}^{T}_{\boldsymbol{Z}^{\mathrm{M}}}(U)]-\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)\right)^2} \leq \frac{1}{1-2\lambda_{\text{Env}}\sigma^2}, \tag{90}$$

$$\mathbb{E}_{\mathcal{P}}\mathbb{E}_{\mathrm{P}_{T_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}|T_{1:\mathrm{N}}}} e^{\lambda_{\text{Env}}\mathrm{N}\left(\mathbb{E}_{\mathrm{P}_{\boldsymbol{T}\boldsymbol{Z}^{\mathrm{M}}}}[\tilde{\mathrm{L}}^{T}_{\boldsymbol{Z}^{\mathrm{M}}}(U)]-\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\tilde{\mathrm{L}}^{T_i}_{\boldsymbol{Z}_i^{\mathrm{M}}}(U)\right)^2} \leq \frac{1}{\sqrt{1-2\lambda_{\text{Env}}\sigma^2}}, \tag{91}$$

concluding (81), and (83). □

Lemma F.5. *Let $X_1, ..., X_n$ be i.i.d random variables, and $f : \mathcal{X} \to [0, 1]$ be a bounded function. For all $n > 8$, we have*

$$\mathbb{E}\left[e^{n\mathrm{D}\left(\frac{1}{n}\sum_i f(X_i)||\mathbb{E}[f(X)]\right)}\right] \leq 2\sqrt{n}. \tag{92}$$

Proof. See Theorem 1 of (Maurer, 2004), then for $n \geq 8$, the right hand side of Eq. 5 of (Maurer, 2004) is smaller than $\sqrt{n}$. □

Lemma F.6. *Let $X_1, ..., X_n$ be i.i.d random variables. For the given function $f : \mathcal{X} \to [0, 1]$, we have*

$$\mathbb{E}\left[e^{n\mathrm{D}_\gamma\left(\frac{1}{n}\sum_i f(X_i)||\mathbb{E}[f(X)]\right)}\right] \leq 1, \tag{93}$$

where $\mathrm{D}_\gamma(a||b) = \gamma a - \log(1 - b + be^\gamma)$.

Proof. See Equation (18) of (McAllester, 2013). For completeness, we repeat it again. Since for $a \in [0, 1]$ and $\gamma \in \mathbb{R}$, we have $e^{\gamma a} \leq 1 - a + a \cdot e^\gamma$, we conclude $e^{\frac{\gamma}{n}\sum_i f(X_i)} \leq 1 - \frac{1}{n}\sum_i f(X_i) + e^\gamma \frac{1}{n}\sum_i f(X_i)$, by taking expectation from both sides, we find that $\mathbb{E}\left[e^{\frac{\gamma}{n}\sum_i f(X_i)}\right] \leq 1 - \mathbb{E}[f(X)] + e^\gamma\mathbb{E}[f(X)]$. Taking logarithm from both sides leads to

$$\mathbb{E}\left[e^{\frac{\gamma}{n}\sum_i f(X_i)-\log(1-\mathbb{E}[f(X)]+e^\gamma\mathbb{E}[f(X)])}\right] \leq 1 \tag{94}$$

Now, since X_is are independent

$$\mathbb{E}_{\mathrm{P}_{\boldsymbol{X}_{1:n}}}\left[e^{n\mathrm{D}_\gamma\left(\frac{1}{n}\sum_i f(X_i)||\mathbb{E}[f(X)]\right)}\right] = \mathbb{E}_{\mathrm{P}_{\boldsymbol{X}_{1:n}}}\left[\prod_{i=1}^{n} e^{\mathrm{D}_\gamma\left(\frac{1}{n}\sum_i f(X_i)||\mathbb{E}[f(X)]\right)}\right] \tag{95}$$

$$= \prod_{i=1}^{n}\mathbb{E}_{\mathrm{P}_{X_i}}\left[e^{\mathrm{D}_\gamma\left(\frac{1}{n}\sum_i f(X_i)||\mathbb{E}[f(X)]\right)}\right] \tag{96}$$

$$= \prod_{i=1}^{n}\mathbb{E}_{\mathrm{P}_{X_i}}\left[e^{\frac{\gamma}{n}\sum_i f(X_i)-\log(1-\mathbb{E}[f(X)]+e^\gamma\mathbb{E}[f(X)])}\right] \leq 1, \tag{97}$$

where the equality in (97) and the last inequality follow from the definition of D_γ and (94), respectively. □

Lemma F.7. *Let $\mathrm{D}_\gamma(a||b) = \gamma a - \log(1 - b + be^\gamma)$. For $\lambda > 0.5$ and $\mathrm{C} \in \mathbb{R}$, if $\mathrm{D}_{-\frac{1}{\lambda}}(a||b) < \mathrm{C}$, then*

$$b \leq \frac{a + \lambda\mathrm{C}}{1 - \frac{1}{2\lambda}}. \tag{98}$$

Proof. See Lemma 2 of (McAllester, 2013). □

As mentioned before, to find PAC-Bayes bounds, usually we have four steps, namely choosing a suitable convex function, applying Jensen's, change of measure and Markov's inequalities. For the most PAC-Bayesian proofs, Donsker-Varadhan's inequality is used as the change of measure inequality:

Lemma F.8. *For any measurable function $\phi(\cdot)$, and two distributions P and Q, we have*

$$\mathbb{E}_{\mathrm{Q}}[\phi(X)] \leq \mathrm{D}(\mathrm{Q}||\mathrm{P}) + \log\left(\mathbb{E}_{\mathrm{P}}\left[e^{\phi(X)}\right]\right). \tag{99}$$