Journal of Machine Learning Research 26 (2025) 1-60 Submitted 3/23; Revised 4/24; Published 1/25

Bayes Meets Bernstein at the Meta Level: an Analysis of Fast Rates in Meta-Learning with PAC-Bayes

Charles Riou CHARLES@MS.K.U-TOKYO.AC.JP
The University of Tokyo & RIKEN Center for AIP
University of Tokyo, Department of Computer Science,
Graduate School of Information Science and Technology
7-3-1 Hongo, Bunkyo-ku
Tokyo 113-0033, Japan.

Pierre Alquier ALQUIER@ESSEC.EDU
ESSEC Business School
Asia-Pacific campus
5 Nepal Park
575749 Singapore.

Badr-Eddine Chérief-Abdellatif BADR-EDDINE.CHERIEF-ABDELLATIF@CNRS.FR
CNRS & LPSM, Sorbonne Université & Université Paris Cité
LPSM, Campus Pierre et Marie Curie
4, place Jussieu
75252 Paris Cedex 05, France.

Editor: Daniel Roy

Abstract

Bernstein's condition is a key assumption that guarantees fast rates in machine learning. For example, under this condition, the Gibbs posterior with prior π has an excess risk in $O(d_\pi/n)$, as opposed to $O(\sqrt{d_\pi/n})$ in the general case, where n denotes the number of observations and d_π is a complexity parameter which depends on the prior π. In this paper, we examine the Gibbs posterior in the context of meta-learning, i.e., when learning the prior π from T previous tasks. Our main result is that Bernstein's condition always holds at the meta level, regardless of its validity at the observation level. This implies that the additional cost to learn the Gibbs prior π, which will reduce the term d_π across tasks, is in $O(1/T)$, instead of the expected $O(1/\sqrt{T})$. We further illustrate how this result improves on the standard rates in three different settings: discrete priors, Gaussian priors and mixture of Gaussian priors.

Keywords: Bernstein's condition, meta-learning, fast rates, PAC-Bayes bounds, information bounds, the Gibbs algorithm, variational approximations.

1. Introduction

One of the greatest promises of artificial intelligence is the ability to design autonomous systems that can learn from different situations throughout their lives and adapt quickly to new environments, as humans, animals and other living things naturally do. Based on the intuition that a new problem often has significant similarities to previously encountered tasks, the use of past experience is particularly important in areas such as computer vision (Quattoni et al., 2008; Kulis et al., 2011; Li et al., 2018; Achille et al., 2019), natural language processing (Huang et al., 2018; Gu

et al., 2018; Dou et al., 2019; Qian and Yu, 2019) and reinforcement learning (Finn et al., 2017; Mishra et al., 2018; Wang et al., 2016; Yu et al., 2020) where the learner has access to only a few training examples for the task of interest, but for which a vast amount of data sets from a variety of related tasks is available. In the area of digit recognition for example, it is possible to leverage the experience gained from millions of similar open source image classification data sets, as the key features needed to classify cats from dogs or pants from shirts can be used to classify handwritten digits. This idea is at the heart of meta-learning (Thrun and Pratt, 1998; Baxter, 2000; Vanschoren, 2019), a field that has recently attracted a lot of attention due to its huge success in real-world applications, and which aims to improve performance on a particular task by transferring the knowledge contained in different but related tasks.

Meta-learning has been widely studied in recent literature. It must be noted that *meta-learning* was used to refer to a wide range of situations. Providing a precise definition of meta-learning is a challenge. In particular, the terms *transfer learning* and *multi-task learning*, although distinct, are often used interchangeably instead of meta-learning. *Transfer learning* is a very general concept that involves two tasks that share similarities - a source and a target - and consists in transferring the knowledge acquired on the source data set to better process the target data (Pan and Yang, 2010; Zhuang et al., 2020). In practice, this can be formulated in many different ways, but the most popular approach is to pre-train a model on the source data, e.g., images of cats and dogs, and then to fine-tune it on the target training data set, e.g., images of handwritten digits. In particular, a challenging problem in transfer learning is to find a measure that quantifies the similarity between the source and target tasks. *Multi-task learning* adopts a different framework, where multiple learning tasks are considered and the goal is to learn a model that can handle all tasks simultaneously (Caruana, 1997; Zhang and Yang, 2021). The model usually has a common representation, e.g., the first layers of a deep neural network, and a task-specific component, e.g., the last layer of the network. *Meta-learning* also considers a collection of data sets from a variety of tasks, but unlike multi-task learning, we are not interested in learning the fixed number of tasks, but rather in being prepared for future tasks that are not yet given. Also, unlike transfer learning, meta-learning exploits the commonality of previous tasks rather than the similarity between some specific source and target tasks. We use these metadata to design a *meta-procedure* that adaptively learns a predictor for *any* new learning task that is a priori unknown, and the goal is to quickly learn to adapt a learning procedure from past experience. Meta-learning is therefore sometimes referred to as *learning-to-learn*, or *lifelong learning* in the online context. The implementation of this learning-to-learn mechanism can take different forms, which we briefly describe in the following paragraph.

As the name suggests, meta-learning involves two levels of abstraction to improve learning over time: a meta-level and a within-task level. At the within-task level, the new task of interest is presented and the corresponding pattern is learned from the training data set of the task at hand. This learning process is greatly accelerated by a meta-learner, which has distilled the knowledge accumulated in previous tasks into the within-task model. The meta-learning procedure can accelerate the within-task algorithm in various ways, and three main categories stand out in the literature: metric-based methods, which are based on non-parametric predictive models governed by a metric that is learned using the meta-training data set (Koch et al., 2015; Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018); model-based methods, which quickly update the parameters in a few learning steps, which can be achieved by the model's internal architecture or another meta-learning model (Santoro et al., 2016; Munkhdalai and Yu, 2017; Mishra et al., 2018); and optimisation-based

methods, which mainly involve learning the hyper-parameters of a within-task algorithm using the meta-training set for fast adaptation (Hochreiter et al., 2001; Ravi and Larochelle, 2017; Finn et al., 2017; Nichol et al., 2018; Qiao et al., 2018; Gidaris and Komodakis, 2018). Due to their performance and ease of implementation, the optimisation-based family is the dominant class in the recent literature, exploiting the idea that well-chosen hyperparameters can greatly speed up the learning process and allow model parameters to be quickly adapted to new tasks with little data. For example, it is possible to learn the task of interest using a gradient descent whose initialisation and learning rate would have been learned from the metadata. Among the best known meta-strategies is the model agnostic meta-learning procedure (MAML) (Finn et al., 2017) and its variants implicit MAML (Rajeswaran et al., 2019), Bayesian MAML (Grant et al., 2018; Yoon et al., 2018; Nguyen et al., 2020) and Reptile (Nichol et al., 2018). We refer the interested reader to the recent review by Chen et al. (2023) for more details.

Our results will turn out to be useful in the regime where the number of tasks T is much larger than the average sample size per task n. A typical example would be given by recommender systems (Candes and Plan, 2010) or toxicogenomics studies (Yamada et al., 2017). Recommendations are usually based on matrix completion of the user-movie matrix. The main limit of this approach is that the dimensions of the matrix are fixed, which means that we are working with a fixed number of users. Using meta learning, we can see each user as a separate task for which the objective is to build a model that would predict which movies will be liked by this user. The number of users T is much larger than the average number of movies n rated by each user. In this case, meta learning will use information from previous users to learn more efficiently the model of a new user from a small sample size.

2. Approach and Contributions

In this paper, we focus on the Gibbs algorithms within tasks, or their variational approximations. The Gibbs algorithm, also known as Gibbs posterior (Alquier et al., 2016) or exponentially weighted aggregation (Dalalyan and Tsybakov, 2008), can also be interpreted in the framework of Bayesian statistics as a kind of generalized posterior (Bissiri et al., 2016; Germain et al., 2016; Knoblauch et al., 2022). PAC-Bayes bounds were developed to control the risk and the excess risk of such procedures (Shawe-Taylor and Williamson, 1997; McAllester, 1998; Catoni, 2004; Zhang, 2006; Catoni, 2007; Yang et al., 2019), see Guedj (2019); Alquier (2024) for recent surveys. More recently, the related mutual information bounds (Russo and Zou, 2019; Haghifam et al., 2021) were also used to study the excess risk of the Gibbs algorithms (Xu and Raginsky, 2017). Gibbs posteriors are often intractable, and it is then easier to compute a variational approximation of such a posterior. It appears that PAC-Bayes bounds can also be used on such approximations (Alquier et al., 2016). Many recent publications built foundations of meta-learning through PAC-Bayes and information bounds (Pentina and Lampert, 2014; Amit and Meir, 2018; Ding et al., 2021; Liu et al., 2021; Rothfuss et al., 2021; Farid and Majumdar, 2021; Rothfuss et al., 2023; Guan and Lu, 2022a; Rezazadeh, 2022). These works and the related literature are discussed in detail in Section 6. Most of these papers proved empirical PAC-Bayes bounds for meta-learning. These bounds can be minimized, and we obtain both a practical meta-learning procedure, together with a numerical certificate on its generalization. However, these works did not focus on the rate of convergence of the excess risk.

Bernstein's condition is a low-noise assumption reflecting the inherent difficulty of the learning task (Mammen and Tsybakov, 1999; Tsybakov, 2004; Bartlett and Mendelson, 2006). While it was initially designed to study the ERM (Bartlett and Mendelson, 2006), it characterizes the learning rate of algorithms beyond the ERM. PAC-Bayes bounds and mutual information bounds show that the excess risk of the Gibbs algorithm is in $O(d_{\pi,t}/n)$ for a sample size n when Bernstein's condition is satisfied (Catoni, 2007; Grünwald and Mehta, 2020), as opposed to the slow rate $O((d_{\pi,t}/n)^{1/2})$ in the general case. The quantity $d_{\pi,t}$ measures the complexity of task t. Importantly, it also depends on the prior distribution π used in the algorithm. Similar results hold when we replace the Gibbs algorithm by a variational approximation (Alquier et al., 2016).

In the meta-learning setting, we are given T tasks simultaneously. Using the Gibbs algorithm with a fixed π in all tasks leads to an average excess risk in$O(\mathbb{E}_t[(d_{\pi,t}/n)^\gamma])$, where $\gamma = 1$ when Bernstein's condition holds for each task $t \in \{1, \dots, T\}$, and $\gamma = 1/2$ otherwise. Here, $\mathbb{E}_t$ denotes the expectation with respect to a future (out-of-sample) task t. This approach is referred to as "learning in isolation", because each task is solved regardless of the others. Of course, in meta-learning we want to take advantage of the multiple tasks. For example, Pentina and Lampert (2014) used the Gibbs algorithm at the meta-level, in order to learn a better prior. The expected benefit is to reduce the complexity term $d_{\pi,t}$.

2.1 Overview of the Paper

- In Section 3, we recall existing PAC-Bayes bounds on the excess risk of the Gibbs algorithm when learning tasks in isolation, and we introduce Bernstein's condition, a fundamental assumption under which the fast rate $O(\mathbb{E}_t[d_{\pi,t}/n])$ is achieved by the Gibbs algorithm.

- In Section 4, we prove that these PAC-Bayes bounds can be used to build a natural criterion to define a meta-learning Gibbs algorithm. This criterion takes the form of a log-loss, which, building on a classic analysis (Bartlett et al., 2006; Vijaykumar, 2021), satisfies Bernstein's condition. This leads to a striking result that the prior π can be learnt with fast rates, regardless of the rate withing tasks. In other words, the meta-learning Gibbs algorithm achieves the excess risk $O(\inf_{\pi\in\mathcal{M}} \mathbb{E}_t[(d_{\pi,t}/n)^\gamma] + 1/T)$ with $\gamma = 1$ if Bernstein's condition is satisfied within tasks, and $\gamma = 1/2$ otherwise. We further raise the open question of the generalization of this result to its variational approximations.

- In Section 5, we apply the previous results to various settings: learning a discrete prior, learning a Gaussian prior and learning a mixture of Gaussians prior. We show that the gain brought from the meta learning is blatant, as in some favorable situations, one can even have $\inf_{\pi\in\mathcal{M}} \mathbb{E}_t[d_{\pi,t}/n] = 0$.

- In Section 6, we provide a deeper comparison with the rich literature on meta-learning.

3. Problem Definition and Notations

Let $\mathcal{Z}$ be a space of observations, Θ a decision space and $\ell : \mathcal{Z} \times \Theta \to \mathbb{R}_+$ a bounded loss function defined on the previously defined sets. Let $\mathcal{P}(\Theta)$ denote the set of all probability distributions on Θ equipped with a suitable σ-field. The learner is given T tasks. For each task $t \in \{1, \dots, T\}$, the learner receives N_t observations $Z_{t,i}, i = 1, \dots, N_t$, assumed to be drawn independently from a

distribution P_t on the decision space $\mathcal{Z}$. Conditions on P_t and N_t will be discussed at the beginning of Section 4 as they are not required in the following discussion. The objective of the learner is to find a parameter θ in the parameter space Θ which minimizes the so-called prediction risk associated to P_t on $\mathcal{Z}$, defined as

$$R_{P_t}(\theta) = \mathbb{E}_{Z\sim P_t}[\ell(Z,\theta)].$$

We denote by $R^*_{P_t}$ the minimum of $R_{P_t}(\theta)$ and by θ^*_t a minimizer:[1]

$$R^*_{P_t} = \inf_{\theta\in\Theta} R_{P_t}(\theta) = R_{P_t}(\theta^*_t).$$

In Bayesian approaches, we seek for $\rho_t \in \mathcal{P}(\Theta)$ such that

$$\mathbb{E}_{\theta\sim\rho_t}[R_{P_t}(\theta)]$$

is small. Defining, for any $\theta \in \Theta$, the empirical risk as

$$\hat{R}_t(\theta) = \frac{1}{N_t}\sum_{i=1}^{N_t} \ell(Z_{t,i},\theta),$$

a standard choice for ρ_t is given by

$$\rho_t(\pi,\alpha,\mathcal{F}) = \arg\min_{\rho\in\mathcal{F}} \left\{ \mathbb{E}_{\theta\sim\rho}\left[\hat{R}_t(\theta)\right] + \frac{\mathrm{KL}(\rho\|\pi)}{\alpha N_t} \right\}, \tag{1}$$

where π is the prior distribution on the parameter θ, α is some temperature parameter which will be made explicit later, and $\mathcal{F} \subseteq \mathcal{P}(\Theta)$. We denote

$$\hat{B}_t(\rho,\pi,\alpha) = \mathbb{E}_{\theta\sim\rho}\left[\hat{R}_t(\theta)\right] + \frac{\mathrm{KL}(\rho\|\pi)}{\alpha N_t}. \tag{2}$$

Interestingly enough, in the unconstrained case where $\mathcal{F} = \mathcal{P}(\Theta)$, we recover the so-called Gibbs posterior or Gibbs algorithm $\rho_t(\pi,\alpha,\mathcal{P}(\Theta))$, which is given by

$$\rho_t(\pi,\alpha,\mathcal{P}(\Theta))(\mathrm{d}\theta) \propto \exp\left(-\alpha N_t \hat{R}_t(\theta)\right)\pi(\mathrm{d}\theta).$$

It is often interpreted as an extension of Bayesian inference to model-free statistical learning, where the likelihood is replaced by the loss function. This connection directly follows from Donsker and Varadhan's variational formula (see Lemma 16), see e.g. Lemma 2.2 of Alquier (2024) for a proof. In the sequel, we will use the following shortcut notation:

$$\rho_t(\pi,\alpha) = \rho_t(\pi,\alpha,\mathcal{P}(\Theta)).$$

In our study, we delve into the theoretical aspects of the Gibbs posterior in meta-learning, leaving aside the issue of its computational feasibility, which nonetheless often arises as a concern in practical scenarios. To briefly address this aspect in this introduction, tackling the computational

1. For the sake of simplicity, we will assume that such a minimizer always exists. By considering an ε-approximate minimizer instead, all our results can be directly extended to the general case.

challenges associated with posterior distributions typically involves employing Monte Carlo sampling techniques. This encompasses widely-used approaches such as Markov Chain Monte Carlo (MCMC) algorithms (Robert and Casella, 2004; Andrieu et al., 2003; Robert, 2007) and Sequential Monte Carlo (Doucet et al., 2001; Doucet and Lee, 2018). Despite their effectiveness, these methods may encounter slower performance, particularly when confronted with very large data sets. Consequently, there has been a notable surge in the development of faster approximation techniques leveraging optimization routines over the past 25 years. These include methods like Expectation Propagation (Minka, 2001) and variational inference (Jordan et al., 1999; Blei et al., 2017), which aim to ascertain a deterministic approximation of the posterior distribution. Note that variational inference can be recast as the minimization of (2) for a specific choice of $\mathcal{F}$. We refer the reader to Martin et al. (2020) for a recent review on Bayesian computation.

We give a complete list of all notations introduced throughout this paper in Appendix A.

3.1 Assumptions on the Loss and Bernstein's Condition

Recall that we assumed that the loss function is bounded: there exists a constant $C > 0$ such that

$$\forall (z, \theta) \in \mathcal{Z} \times \Theta,\ \ell(z, \theta) \leq C. \tag{3}$$

This is a classical assumption in machine learning, which is however restrictive. Here, we want to highlight that PAC-Bayes bounds for unbounded losses are well-known, see Section 5 in (Alquier, 2024). In particular, Theorem 1 below can be extended without modification to sub-Gaussian and sub-exponential variables. The extension to heavy-tailed variables requires adaptations, but is also possible. On the other hand, our results on meta learning rely explicitly on the boundedness assumption. We will thus make this assumption in the whole paper.

With such a bounded loss, in parametric settings, it is possible to prove generalization bounds in $1/\sqrt{N_t}$ for various methods, including the empirical risk minimizer or ERM (Devroye et al., 1996, Chapter 12). This is also true for the Gibbs posterior (Catoni, 2007, Chapter 1). However, in some specific settings, it is possible to achieve faster bounds. For example, it is standard that the ERM achieves a rate in $1/N_t$ when there is a perfect predictor, that is, $R^*_{P_t} = 0$ (Devroye et al., 1996, Section 12.1). It turns out that fast rates are possible under more general conditions. Bartlett et al. (2006) proved this holds when the loss function is Lipschitz and strongly convex. Mammen and Tsybakov (1999) proved this is also true for classification under margin conditions. It turns out that these assumptions are all special cases of the so-called Bernstein's condition (Bartlett and Mendelson, 2006, Definition 2.6 with $\beta = 1$). In other words, either $R^*_{P_t} = 0$, or Mammen and Tsybakov's margin assumption, or the smoothness and convexity assumption on the loss of Bartlett et al. (2006) all imply Bernstein's condition, which itself is enough to prove fast rates. We refer the reader to (Alquier, 2024, Section 4) for more details on this topic in the study of the Gibbs posterior. We now recall Bernstein's condition. It requires to introduce the following variance term for task t and for $\theta \in \Theta$:

$$V_t(\theta, \theta^*_t) := \mathbb{E}_{Z \sim P_t}\left[\left|\ell(Z, \theta) - \ell(Z, \theta^*_t)\right|^2\right].$$

Assumption 1 (Bernstein's condition) *There exists a constant $c > 0$ such that, for any $\theta \in \Theta$,*

$$V_t(\theta, \theta^*_t) \leq c\left(R_{P_t}(\theta) - R^*_{P_t}\right).$$

This assumption characterizes the excess risk of Gibbs posterior, see Theorem 1 below. In this paper, we will provide a bound on the excess risk both under this condition and without it.

3.2 Learning in Isolation

In the process of learning in isolation, we consider each of the tasks separately. We then fix some $t \in \{1, \dots, T\}$ and denote by $\mathcal{S}_t$ the sample of N_t i.i.d. observations from task t: $\mathcal{S}_t = (Z_{t,1}, \dots, Z_{t,N_t})$. We recall the following classic result (Alquier, 2024, Theorems 4.1 and 4.3), whose proof is recalled in Appendix C for the sake of completeness.

Theorem 1 *Assume that the loss ℓ satisfies the boundedness assumption* (3) *with constant C. Then, the following bound holds, for any $\alpha \in \left(0, \frac{1}{C}\right)$:*

$$\mathbb{E}_{\mathcal{S}_t}\mathbb{E}_{\theta\sim\rho_t(\pi,\alpha,\mathcal{F})}\left[R_{P_t}(\theta)\right] - R^*_{P_t} \leq \frac{1}{1-\frac{\alpha c \mathbb{I}_B}{2(1-C\alpha)}}\left(\mathbb{E}_{\mathcal{S}_t}\left[\inf_{\rho\in\mathcal{F}}\left\{\mathbb{E}_{\theta\sim\rho}\left[\hat{R}_t(\theta)\right] - \hat{R}_t(\theta^*_t) + \frac{\mathrm{KL}(\rho\|\pi)}{\alpha N_t}\right\}\right] + \frac{\alpha C^2(1-\mathbb{I}_B)}{8}\right),$$

where $\mathbb{E}_{\mathcal{S}_t}$ is a short notation for the expectation w.r.t. $\mathcal{S}_t \triangleq (Z_{t,1}, \dots, Z_{t,N_t})$ i.i.d. from P_t, and $\mathbb{I}_B$ is equal to 1 if Assumption 1 (Bernstein's condition) is satisfied, and 0 otherwise. In particular, if Assumption 1 is satisfied with constant c, the choice $\alpha = \frac{1}{c+C}$ yields the bound

$$\mathbb{E}_{\mathcal{S}_t}\mathbb{E}_{\theta\sim\rho_t(\pi,\alpha,\mathcal{F})}[R_{P_t}(\theta)] - R^*_{P_t} \leq 2\mathbb{E}_{\mathcal{S}_t}\left[\inf_{\rho\in\mathcal{F}}\left\{\mathbb{E}_{\theta\sim\rho}\left[\hat{R}_t(\theta)\right] - \hat{R}_t(\theta^*_t) + \frac{(C+c)\mathrm{KL}(\rho\|\pi)}{N_t}\right\}\right].$$

Theorem 1 not only justifies the choice of $\rho_t(\pi, \alpha, \mathcal{F})$ as the within-task algorithm, but also paves the way for its statistical analysis. Indeed, while it is known in the PAC-Bayes literature that the definition in (1) is actually a minimization program of a generalization bound over the risk $\mathbb{E}_{\theta\sim\rho}\left[R_{P_t}(\theta)\right]$ (see e.g. Corollary 2.3 in (Alquier, 2024) for an exact statement), this fact alone does not provide any statistical rate for the risk of $\rho_t(\pi, \alpha, \mathcal{F})$. Theorem 1 formulates an oracle-type inequality whose derivation, when specifying a model and a prior, leads to explicit convergence rates.

For example, a classic assumption in the Bayesian literature is that there are constants $\kappa, d \geq 0$ such that, for s small enough, $\pi(\{\theta : \delta(\theta, \theta^*_t) \leq s\}) \geq s^d/\kappa$, where $\delta(\theta, \theta^*_t)$ is a measure of the discrepancy between θ and θ^*_t. When studying the contraction of the posterior in Bayesian statistics, δ is usually the Kullback-Leibler divergence (Ghosal and Van der Vaart, 2017, Condition (i) in Theorem 8.9). However, more general δ can be considered, such as Rényi divergences or the Hellinger distance (Zhang and Gao, 2020, Condition C.3 in Theorem 2.1). In PAC-Bayesian bounds, where the prediction risk is the main focus, the condition is often used with $\delta = R_t(\theta) - R_t(\theta^*_t)$ (Alquier, 2024, Section 4.4). As this condition is usually applied to one task with a specific prior, the notation d does not reflect the dependence with respect to π or t. However, in our context, this dependence will be crucial, so we will write $d_{\pi,t}$ instead of d. Under such an assumption, the right-hand side in Theorem 1 can be made more explicit.

Corollary 2 *Recall that the loss ℓ satisfies the boundedness assumption* (3) *with constant C. Assume that, for any $0 < s < s_0$, $\pi(\{\theta : R_{P_t}(\theta) - R^*_{P_t} \leq s\}) \geq s^{d_{\pi,t}}/\kappa_{\pi,t}$. Then, as soon as*

$N_t \geq d_{\pi,t}/(\alpha s_0)$,

$$\mathbb{E}_{\mathcal{S}_t}\left[\inf_{\rho\in\mathcal{P}(\Theta)} \hat{B}_t(\rho,\pi,\alpha) - \hat{R}_t(\theta_t^*)\right] \leq \frac{d_{\pi,t}\log\frac{\alpha e N_t}{d_{\pi,t}} + \log\kappa_{\pi,t}}{\alpha N_t}.$$

In particular, if Bernstein's condition (Assumption 1) is satisfied with constant c, the choice $\alpha = 1/(c+C)$ *gives*

$$\mathbb{E}_{\mathcal{S}_t}\mathbb{E}_{\theta\sim\rho_t(\pi,\alpha)}[R_{P_t}(\theta)] \leq R^*_{P_t} + \frac{2(C+c)}{N_t}\left(d_{\pi,t}\log\frac{eN_t}{d_{\pi,t}(C+c)} + \log\kappa_{\pi,t}\right).$$

On the other hand, without Bernstein's condition (Assumption 1),

$$\mathbb{E}_{\mathcal{S}_t}\mathbb{E}_{\theta\sim\rho_t(\pi,\alpha)}[R_{P_t}(\theta)] \leq R^*_{P_t} + \frac{d_{\pi,t}\log\frac{\alpha e N_t}{d_{\pi,t}} + \log\kappa_{\pi,t}}{\alpha N_t} + \frac{\alpha C^2}{8},$$

and in particular, for $\alpha = 2\sqrt{2d_{\pi,t}}/(\sqrt{N_t}C)$*, we obtain*

$$\mathbb{E}_{\mathcal{S}_t}\mathbb{E}_{\theta\sim\rho_t(\pi,\alpha)}[R_{P_t}(\theta)] \leq R^*_{P_t} + \frac{C}{2}\sqrt{\frac{d_{\pi,t}}{2N_t}}\left(\frac{1}{2}\log\frac{8e^4N_t}{d_{\pi,t}C^2} + \frac{1}{d_{\pi,t}}\log\kappa_{\pi,t}\right).$$

Corollary 2 thus provides explicit rates of convergence for $\rho_t(\pi,\alpha)$: under Bernstein's condition, we recover the fast rate $d_{\pi,t}/N_t$ for a good choice of α, while we obtain the general rate $\sqrt{d_{\pi,t}/N_t}$ in the general case. Yet, in both situations, the rate of convergence depends on the choice of the prior π and on the given task through the complexity term $d_{\pi,t}$. In the following, we propose to improve this dependence by meta-learning the prior π using all the data from the many other observed tasks.

4. Main Results

From this section onward, we focus on the meta learning. As opposed to the learning in isolation, the meta learning considers all the tasks $t \in \{1,\dots,T\}$ and takes advantage of possible similarities between the T tasks to improve the learning in each task. More precisely, while assuming that for any $t \in \{1,\dots,T\}, \mathcal{S}_t = (Z_{t,1},\dots,Z_{t,N_T})$ is a sample of N_t i.i.d observations from P_t, we further assume that the couples $(P_1,N_1),\dots,(P_T,N_T)$ are drawn independently from a distribution $\mathcal{P}$ on distributions and sample sizes. A future (out-of-sample) task is a couple (P_{T+1},N_{T+1}) which is drawn from $\mathcal{P}$, independently from $(P_1,N_1),\dots,(P_T,N_T)$, and $\mathcal{S}_{T+1} = (Z_{T+1,1},\dots,Z_{T+1,N_{T+1}})$ will be a sample of N_{T+1} i.i.d draws from P_{T+1}, conditionally on (P_{T+1},N_{T+1}). This task will be solved by the Gibbs algorithm $\rho_{T+1}(\pi,\alpha)$. Our objective is to learn the prior π using the tasks $t \in \{1,\dots,T\}$, such that (P_{T+1},N_{T+1}) would be solved more efficiently. More formally, π should make the meta-risk

$$E(\pi) = \mathbb{E}_{P_{T+1},N_{T+1}}\mathbb{E}_{\mathcal{S}_{T+1}}\mathbb{E}_{\theta\sim\rho_{T+1}(\pi,\alpha)}[R_{P_{T+1}}(\theta)].$$

as small as possible, where $\mathbb{E}_{\mathcal{S}_{T+1}}$ denotes the expectation w.r.t. $Z_{T+1,1},\dots,Z_{T+1,N_{T+1}} \sim P_{T+1}$ conditionally on (P_{T+1},N_{T+1}), and $\mathbb{E}_{P_t,N_t}$ denotes the expectation w.r.t. $(P_t,N_t) \sim \mathcal{P}$ for any $t \in \{1,\dots,T+1\}$. For a more detailed list of notations including the expectations, please see Appendix A. We will compare the meta risk $\mathcal{E}(\pi)$ to the so-called oracle meta-risk

$$\mathcal{E}^* = \mathbb{E}_{P_{T+1},N_{T+1}}[R^*_{P_{T+1}}],$$

which can only be reached by an oracle who would know the best classifier in each task in advance.

4.1 Bernstein's Condition at the Meta Level

In this subsection, we prove a version of Bernstein's condition at the meta level. For any prior π, let

$$\widehat{\mathcal{L}}_t(\pi, \alpha) = \hat{B}_t(\rho_t(\pi, \alpha), \pi, \alpha), \tag{4}$$

and let

$$\mathcal{L}_t(\pi, \alpha) = \mathbb{E}_{\mathcal{S}_t}\left[\widehat{\mathcal{L}}_t(\pi, \alpha)\right],$$

where we recall that $\hat{B}_t(\rho, \pi, \alpha)$ is defined in (2). Let π^*_α be the distribution minimizing the expectation of the above quantity:

$$\pi^*_\alpha = \arg\min_\pi \mathbb{E}_{P_t, N_t}\left[\mathcal{L}_t(\pi, \alpha)\right]. \tag{5}$$

At the within-task level, Assumption 1 is the key hypothesis to determine the rate of convergence, as shown in Theorem 1. We would expect a similar assumption-theorem structure at the meta-level. Surprisingly enough, this is not the case. The following result proves that a condition, formally very similar to Bernstein's condition, holds unconditionally at the meta-level. This will be used in the proof of Theorem 5 below to show that one can always achieve fast rates when learning the prior.

Theorem 3 *Assume that the loss ℓ satisfies the boundedness assumption* (3) *with constant C. Recall that $\widehat{\mathcal{L}}_t(\pi, \alpha)$ and π^*_α are defined as in* (4) *and* (5) *above. Then, for any $\alpha \in \left(0, \frac{1}{C}\right)$ and $\pi \in \mathcal{P}(\Theta)$,*

$$\mathbb{E}_{P_t, N_t}\mathbb{E}_{\mathcal{S}_t}\left[\left(\widehat{\mathcal{L}}_t(\pi, \alpha) - \widehat{\mathcal{L}}_t(\pi^*_\alpha, \alpha)\right)^2\right] \leq 8eC\mathbb{E}_{P_t, N_t}\mathbb{E}_{\mathcal{S}_t}\left[\widehat{\mathcal{L}}_t(\pi, \alpha) - \widehat{\mathcal{L}}_t(\pi^*_\alpha, \alpha)\right].$$

Proof The proof relies on two arguments. First, we can rewrite the upper bound on the task risk as

$$\widehat{\mathcal{L}}_t(\pi, \alpha) = -\frac{1}{N_t\alpha}\log\mathbb{E}_{\theta\sim\pi}\left[\mathrm{e}^{-N_t\alpha\hat{R}_t(\theta)}\right] = -\frac{1}{\tau}\log\left(\mathbb{E}_{\theta\sim\pi}\left[\mathrm{e}^{-N_t\alpha\hat{R}_t(\theta)}\right]^{\frac{\tau}{N_t\alpha}}\right) \tag{6}$$

for any fixed $\tau > 0$ (using, e.g., Lemma 16 in the Appendix). Under this form, we will be able to use the arguments from Bartlett et al. (2006, Lemma 7) based on strong convexity. The strong convexity of the function $f(x) = -(1/\tau)\log(x)$ on a bounded interval is of course known, and was indeed used to derive fast rates in machine learning before (Vijaykumar, 2021, Lemma 10). We write it in the most convenient way for our proof in Lemma 4 (which we still prove in the appendix for the sake of completeness). First, observe that, by the boundedness assumption (3), it holds that, for any $\pi \in \mathcal{P}(\Theta)$,

$$\exp(-C\tau) \leq \mathbb{E}_{\theta\sim\pi}\left[\mathrm{e}^{-N_t\alpha\hat{R}_t(\theta)}\right]^{\frac{\tau}{N_t\alpha}} \leq 1.$$

Lemma 4 *Let $f : x \mapsto -\frac{1}{\tau}\log x$. Then, for any $x, y \in [\exp(-C\tau), 1]$,*

$$(f(x) - f(y))^2 \leq \frac{8\exp(2C\tau)}{\tau}\left(\frac{f(x) + f(y)}{2} - f\left(\frac{x+y}{2}\right)\right).$$

An application of Lemma 4 to $x = \mathbb{E}_{\theta\sim\pi}\left[\mathrm{e}^{-N_t\alpha\hat{R}_t(\theta)}\right]^{\frac{\tau}{N_t\alpha}}$ and $y = \mathbb{E}_{\theta\sim\pi^*_\alpha}\left[\mathrm{e}^{-N_t\alpha\hat{R}_t(\theta)}\right]^{\frac{\tau}{N_t\alpha}}$ gives

$$\begin{aligned}
&\left(\widehat{\mathcal{L}}_t(\pi,\alpha) - \widehat{\mathcal{L}}_t(\pi^*_\alpha,\alpha)\right)^2 \\
&= \left(\hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha) - \hat{B}_t(\rho_t(\pi^*_\alpha,\alpha),\pi^*_\alpha,\alpha)\right)^2 \\
&= (f(x) - f(y))^2 \\
&\leq \frac{8\exp(2C\tau)}{\tau}\left(\frac{f(x)+f(y)}{2} - f\left(\frac{x+y}{2}\right)\right) \\
&= \frac{8\exp(2C\tau)}{\tau}\left[\frac{\hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha) + \hat{B}_t(\rho_t(\pi^*_\alpha,\alpha),\pi^*_\alpha,\alpha)}{2} + \frac{1}{\tau}\log\left(\frac{x+y}{2}\right)\right].
\end{aligned}$$

Taking expectations with respect to $\mathcal{S}_t$ on both sides yields

$$\mathbb{E}_{\mathcal{S}_t}\left[\left(\widehat{\mathcal{L}}_t(\pi,\alpha) - \widehat{\mathcal{L}}_t(\pi^*_\alpha,\alpha)\right)^2\right] \leq \frac{8\exp(2C\tau)}{\tau}\left(\frac{\mathcal{L}_t(\pi,\alpha)+\mathcal{L}_t(\pi^*_\alpha,\alpha)}{2} - \mathcal{L}_t\left(\frac{\pi+\pi^*_\alpha}{2},\alpha\right)\right).$$

Integrating with respect to $(P_t, N_t) \sim \mathcal{P}$ yields

$$\begin{aligned}
&\mathbb{E}_{P_t,N_t}\mathbb{E}_{\mathcal{S}_t}\left[\left(\widehat{\mathcal{L}}_t(\pi,\alpha) - \widehat{\mathcal{L}}_t(\pi^*_\alpha,\alpha)\right)^2\right] \\
&\qquad\leq \frac{8\exp(2C\tau)}{\tau}\left(\frac{\mathbb{E}_{P_t,N_t}\mathcal{L}_t(\pi,\alpha)+\mathbb{E}_{P_t,N_t}\mathcal{L}_t(\pi^*_\alpha,\alpha)}{2} - \mathbb{E}_{P_t,N_t}\mathcal{L}_t\left(\frac{\pi+\pi^*_\alpha}{2},\alpha\right)\right). \quad (7)
\end{aligned}$$

By definition of π^*_α, it holds that, for any $\pi' \in \mathcal{P}(\Theta)$,

$$\mathbb{E}_{P_t,N_t}[\mathcal{L}_t(\pi^*_\alpha,\alpha)] \leq \mathbb{E}_{P_t,N_t}[\mathcal{L}_t(\pi',\alpha)].$$

In particular, this holds for $\pi' = (\pi+\pi^*_\alpha)/2$ and plugging this into the right hand side of (7) gives

$$\mathbb{E}_{P_t,N_t}\mathbb{E}_{\mathcal{S}_t}\left[\left(\widehat{\mathcal{L}}_t(\pi,\alpha) - \widehat{\mathcal{L}}_t(\pi^*_\alpha,\alpha)\right)^2\right] \leq \frac{4\exp(2C\tau)}{\tau}\left(\mathbb{E}_{P_t,N_t}\mathcal{L}_t(\pi,\alpha) - \mathbb{E}_{P_t,N_t}\mathcal{L}_t(\pi^*_\alpha,\alpha)\right).$$

The (optimal) choice $\tau = \frac{1}{2C}$ gives the desired bound

$$\mathbb{E}_{P_t,N_t}\mathbb{E}_{\mathcal{S}_t}\left[\left(\widehat{\mathcal{L}}_t(\pi,\alpha) - \widehat{\mathcal{L}}_t(\pi^*_\alpha,\alpha)\right)^2\right] \leq 8eC\mathbb{E}_{P_t,N_t}\mathbb{E}_{\mathcal{S}_t}\left[\widehat{\mathcal{L}}_t(\pi,\alpha) - \widehat{\mathcal{L}}_t(\pi^*_\alpha,\alpha)\right].$$

■

4.2 PAC-Bayes Bound for Meta-learning

We will now seek a distribution Π on the set of priors π which allows to obtain a small meta-risk

$$\mathbb{E}_{\pi\sim\Pi}[\mathcal{E}(\pi)].$$

In order to do so, we will fix a set $\mathcal{M}$ of possible priors, and a subset $\mathcal{G}$ of the set $\mathcal{P}(\mathcal{M})$ of distributions on these priors: $\mathcal{G} \subseteq \mathcal{P}(\mathcal{M})$.[2] Of course, $\mathcal{G} = \mathcal{P}(\mathcal{M})$ is a possible choice. However, smaller sets might be preferable for computational reasons (that is, we can use variational approximations at the meta-level). Given a probability distribution $\Lambda \in \mathcal{G}$ called "prior on priors", we define Gibbs distribution on priors similarly as in (1), but at the meta-level:

$$\hat{\Pi} = \arg\min_{\Pi \in \mathcal{G}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi \sim \Pi} \left[\widehat{\mathcal{L}}_t(\pi, \alpha) \right] + \frac{\mathrm{KL}(\Pi \| \Lambda)}{\beta T} \right\}, \tag{8}$$

where $\beta > 0$ is some parameter made explicit in the next theorem. As a consequence of Theorem 3 comes the next result, whose proof is given in Appendix F.

Theorem 5 *Assume that the loss ℓ satisfies the boundedness assumption* (3) *with constant C. The choice $\beta = \frac{1}{(1+8e)C}$ yields, for any $\mathcal{F} \subseteq \mathcal{P}(\Theta)$ and any $\alpha \in \left(0, \frac{1}{C}\right)$,*

$$\mathbb{E}_{P_{1:T}, N_{1:T}} \mathbb{E}_{\mathcal{S}_{1:T}} \mathbb{E}_{\pi \sim \hat{\Pi}} \left[\mathcal{E}(\pi) \right] - \mathcal{E}^* \leq \frac{2}{1 - \frac{\alpha c \mathbb{I}_B}{2(1 - C\alpha)}} \inf_{\Pi \in \mathcal{G}} \mathbb{E}_{P_{T+1}, N_{T+1}} \Bigg[$$
$$\mathbb{E}_{\pi \sim \Pi} \left[\inf_{\rho \in \mathcal{F}} \left\{ \mathbb{E}_{\theta \sim \rho}[R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} + \frac{\mathrm{KL}(\rho \| \pi)}{\alpha N_{T+1}} \right\} \right] + \frac{(1+8e)C \cdot \mathrm{KL}(\Pi \| \Lambda)}{T} + \frac{\alpha C^2 (1 - \mathbb{I}_B)}{8} \Bigg],$$

where $\mathbb{I}_B$ is equal to 1 if Assumption 1 holds and 0 otherwise, $\mathbb{E}_{\mathcal{S}_{1:T}}$ is a short notation for $\mathbb{E}_{\mathcal{S}_1} \ldots \mathbb{E}_{\mathcal{S}_T}$ and $\mathbb{E}_{P_{1:T}, N_{1:T}}$ is a short for $\mathbb{E}_{P_1, N_1} \ldots \mathbb{E}_{P_T, N_T}$. In particular, if Assumption 1 holds with constant c, the (optimal) choice $\alpha = \frac{1}{C+c}$ yields, for any $\mathcal{F} \subseteq \mathcal{P}(\Theta)$,

$$\mathbb{E}_{P_{1:T}, N_{1:T}} \mathbb{E}_{\mathcal{S}_{1:T}} \mathbb{E}_{\pi \sim \hat{\Pi}} \left[\mathcal{E}(\pi) \right] - \mathcal{E}^* \leq 4 \inf_{\Pi \in \mathcal{G}} \mathbb{E}_{P_{T+1}, N_{T+1}} \Bigg[$$
$$\mathbb{E}_{\pi \sim \Pi} \left[\inf_{\rho \in \mathcal{F}} \left\{ \mathbb{E}_{\theta \sim \rho}[R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} + \frac{(c+C)\mathrm{KL}(\rho \| \pi)}{N_{T+1}} \right\} \right] + \frac{(1+8e)C\mathrm{KL}(\Pi \| \Lambda)}{T} \Bigg].$$

Remark 6 *A special case of interest is when the sample size N_t and the distribution P_t of the observations are independent. Intuitively, this implies that the number of observations is independent from the value of the optimal parameter θ_t^*. In this case, the expectation $\mathbb{E}_{P_t, N_t}$ can be taken successively and independently w.r.t. P_t and to N_t, as $\mathbb{E}_{P_t, N_t} = \mathbb{E}_{P_t} \mathbb{E}_{N_t}$. In particular, we can define the harmonic expected sample size n by*

$$\frac{1}{n} = \mathbb{E}_{N_t} \left[\frac{1}{N_t} \right],$$

and by inverting $\mathbb{E}_{N_{T+1}}$ and the block $\mathbb{E}_{\pi \sim \Pi}[\inf_{\rho \in \mathcal{F}}\{\ldots\}]$ in the right-hand side above, this yields

$$\mathbb{E}_{P_{1:T}, N_{1:T}} \mathbb{E}_{\mathcal{S}_{1:T}} \mathbb{E}_{\pi \sim \hat{\Pi}} \left[\mathcal{E}(\pi) \right] - \mathcal{E}^* \leq \frac{2}{1 - \frac{\alpha c \mathbb{I}_B}{2(1 - C\alpha)}} \inf_{\Pi \in \mathcal{G}} \mathbb{E}_{P_{T+1}} \Bigg[$$

2. Note that measurability issues can arise when the set $\mathcal{F}$ is non parametric. However, in all our examples, the set $\mathcal{F}$ is parametric.

$$\mathbb{E}_{\pi\sim\Pi}\left[\inf_{\rho\in\mathcal{F}}\left\{\mathbb{E}_{\theta\sim\rho}[R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} + \frac{(c+C)\mathrm{KL}(\rho\|\pi)}{n}\right\}\right] + \frac{(1+8e)C\mathrm{KL}(\Pi\|\Lambda)}{T}\Bigg],$$

under Assumption 1. In this case, n quantifies the expected rate of convergence across tasks. Naturally, this also holds in the more specific case where the sample size is constant, i.e., $N_t = n$ almost surely (in this case, we will write $\mathbb{E}_{P_{1:T}}$ instead of $\mathbb{E}_{P_{1:T},N_{1:T}}$). From now on, we will present several applications of Theorem 5, and show how its bound improves upon the one from the learning in isolation. For the sake of clarity, in the coming applications, we will work in the simpler setting $N_t = n$ a.s. for any task $t \in \{1,\dots,T+1\}$. However, our results straightforwardly extend to the general setting considered in Theorem 5.

Up to a constant factor, recall that the learning in isolation achieves a bound

$$\mathbb{E}_{\mathcal{S}_{T+1}}\left[\inf_{\rho\in\mathcal{F}}\left\{\mathbb{E}_{\theta\sim\rho}\left[\hat{R}_{T+1}(\rho,\pi,\alpha)\right] - \hat{R}_{T+1}(\theta^*_{T+1}) + \left(\frac{\mathrm{KL}(\rho\|\pi)}{n}\right)^{\gamma}\right\}\right],$$

where $\gamma = 1$ if Assumption 1 is satisfied, and $\frac{1}{2}$ otherwise. While the bound in isolation depends on the choice of prior π, in contrast, the meta learning achieves the above bound with $\pi \sim \Pi$, for the best possible Π. This comes at the cost of an additional $\frac{(C+c)\mathrm{KL}(\Pi\|\Lambda)}{T}$ term, which is of order $O\left(\frac{1}{T}\right)$ and hence very small in the case $T \gg n$. Interestingly enough, when Assumption 1 is not satisfied, this additional term is very small in the more general $T \gg \sqrt{n}$ regime, further pleading in favor of the meta learning when the number of tasks T is large.

We point out that the bound in Theorem 5 is written as depending on an infimum on some subset $\mathcal{F} \subseteq \mathcal{P}(\Theta)$. Of course, the tightest bound is reached for $\mathcal{F} = \mathcal{P}(\Theta)$. However, in all the practical examples below, it is much easier to derive explicit rates by using a well-chosen $\mathcal{F} \subsetneq \mathcal{P}(\Theta)$. Thus, we preferred to state the result directly with $\mathcal{F}$. Note that this is simply a theoretical device helpful to make the bound more explicit, and has nothing to do with possible variational approximations within tasks. Indeed, the bound is on the excess risk of $\hat{\Pi}$, which itself is based on the exact Gibbs posterior $\rho_t(\pi,\alpha)$, and not on its variational approximation $\rho_t(\pi,\alpha,\mathcal{F})$.

In contrast, we can define a variational approximation of $\hat{\Pi}$ based on $\rho_t(\pi,\alpha,\mathcal{F})$, as

$$\hat{\Pi}(\mathcal{F}) := \arg\min_{\Pi\in\mathcal{G}}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi\sim\Pi}\left[\hat{B}_t\left(\rho_t(\pi,\alpha,\mathcal{F}),\pi,\alpha\right)\right] + \frac{\mathrm{KL}(\Pi\|\Lambda)}{\beta T}\right\}.$$

In some settings, $\hat{\Pi}(\mathcal{F})$ is tractable while its Gibbs-based counterpart $\hat{\Pi}$ is not, and it is a fundamental open question to determine under what condition on $\mathcal{F}$ we can replace $\hat{\Pi}$ by $\hat{\Pi}(\mathcal{F})$ in Theorem 5.

Open Question 1 *Under what conditions on $\mathcal{F}$ can we replace $\hat{\Pi}$ by $\hat{\Pi}(\mathcal{F})$ in the left-hand side of Theorem 5?*

We would like to emphasize the significance of this question. As indicated in Section 3, computing the within-task Gibbs posterior is typically challenging in many practical scenarios, and sampling methods like MCMC can be expensive. Using variational inference within-task could simultaneously address the intractability of both the within-task Gibbs posterior $\rho_t(\pi,\alpha)$ and the meta-Gibbs distribution $\hat{\Pi}$. Therefore, extending Theorem 5 to any variational family $\mathcal{F}$ is crucial, and will be the focus of future research.

4.3 A Toy Example: Concurrent Priors

This subsection gives a toy application of Theorem 5 just to fix ideas. Here, we study the case where M statisticians propose a different prior, all of which are assumed to satisfy a prior mass condition as in Corollary 2. We denote by $\mathcal{M} = \{\pi_1, \dots, \pi_M\}$ the set of priors. We choose Λ as the uniform distribution on $\mathcal{M}$ and $\mathcal{G} = \mathcal{P}(\mathcal{M})$. Here again, for the sake of simplicity, we assume that Bernstein's condition (see Assumption 1) is satisfied at the within-task level, with constant c. We also assume that $N_t = n$ almost surely. A direct application of Theorem 5 and Corollary 2 gives

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{\mathcal{S}_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}\left[\mathcal{E}(\pi)\right] - \mathcal{E}^* \leq 4(C+c)\min_{\pi\in\mathcal{M}}\mathbb{E}_{P_{T+1}}\frac{d_{\pi,T+1}\log\frac{n}{(C+c)d_{\pi,T+1}} + \log\kappa_{\pi,T+1}}{n} + \frac{4(1+8e)C\log M}{T}.$$

In other words, we obtain the rate of convergence provided by the best prior among $\{\pi_1, \dots, \pi_M\}$, at the price of an additional $O(\log(M)/T)$ term.

5. Applications of Theorem 5

In this section, by an application of Theorem 5, we derive explicit bounds on the excess risk of the Gibbs algorithm in the case of discrete priors (the parameter set Θ is finite; Subsection 5.1), Gaussian priors (Subsection 5.2) and mixtures of Gaussians priors (Subsection 5.3).

In all the cases below, Assumption 1 (Bernstein's condition) holds with constant c (by proposition or assumption), and we will study the excess risk of the meta predictor $\hat{\Pi}$ defined in (8) with the fixed choices $\alpha = \frac{1}{C+c}$ and $\beta = \frac{1}{(1+8e)C}$ throughout this section. Finally, in all this section, we assume that $N_t = n$ almost surely for the sake of simplicity (see Remark 6).

5.1 Learning Discrete Priors

In this subsection, we assume that $|\Theta| = M < \infty$. Following Meunier and Alquier (2021), we define A^* as the smallest possible subset of Θ such that

$$\forall P \sim \mathcal{P},\ \theta^* := \arg\min_{\theta} R_P(\theta) \in A^*, \tag{9}$$

and we denote $m^* := |A^*|$. In general, $A^* = \Theta$ and $m^* = M$. However, in some favorable situations, $A^* \neq \Theta$ and $m^* \ll M$, in which case, the meta-learning may improve upon the learning in isolation. In the setting considered, Bernstein's condition is trivially satisfied for some constant c, and the excess risk of the Gibbs algorithm is $\frac{4(C+c)\log(M)}{n}$.

We define our set of priors $\mathcal{M}$ as the set of probability distributions π_A which are uniform on a subset $A \subseteq \Theta$ and parameterized by A:

$$\mathcal{M} = \{\pi_A | A \subseteq \Theta\},$$

and $\mathcal{G}$ is the set of all distributions on $\mathcal{M}$. Our "prior on priors" Λ is then defined as follows: we draw $m \in \{1, \dots, M\}$ with probability $\frac{2^{M-m}}{2^M-1} \propto 2^{-m}$, then given m, draw a subset $A \subseteq \Theta$ of cardinality m uniformly at random, and take π_A. In other words, Λ is a distribution defined on $\mathcal{F}$ such that $P_{\pi\sim\Lambda}(\pi = \pi_A) = \frac{2^{M-m}}{2^M-1} \times \frac{1}{\binom{M}{m}}$.

Proposition 7 *The excess risk of the meta predictor $\hat{\Pi}$ defined in* (8) *is bounded as follows:*

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{\mathcal{S}_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] \leq \mathcal{E}^* + \frac{4(C+c)\log m^*}{n} + \frac{4(1+8e)Cm^*\log\frac{2eM}{m^*}}{T}.$$

Remark 8 *Let us compare the meta-learning rate above to the $\frac{4(C+c)\log M}{n}$ rate achieved by the learning in isolation. In the unfavorable case $m^* \sim M$, the meta-learning bound is sensibly larger than the learning in isolation one, by an $O(M/T)$ term which vanishes rapidly when $T \to +\infty$.*

However, as soon as $m^* < M$, there is an improvement at the task level, as $\frac{4\log M}{\alpha n}$ is replaced by $\frac{4\log m^\star}{\alpha n}$. This means that for T large enough, the meta learning will always bring an improvement over the learning in isolation. In the very favorable case $m^* \ll M$, the improvement might be huge, and in the extreme case where $m^* = 1$, we have

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{\mathcal{S}_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] \leq \mathcal{E}^* + \frac{4(1+8e)C\log(2eM)}{T}.$$

As expected, the benefits of the meta learning appear in the $T \gg n$ regime, with a potential gain of an $O(1/n)$ term in many different scenarios, at the cost of an additional $O(M/T)$ term in the least favorable of them. This is in line with Meunier and Alquier (2021, Theorem 3) in the online setting.

Proof We first consider the learning in isolation. The classical choice is to take π uniform in each task. Bernstein's condition (Assumption 1) holds with constant c, and an application of Theorem 1 with $\alpha = \frac{1}{C+c}$ gives

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_t}\left[\mathbb{E}_{\theta\sim\rho_t(\pi,\alpha)}[R_{P_t}(\theta)]\right] - \mathcal{E}^* &\leq 2\inf_{\rho\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim\rho}\left[R_{P_t}(\theta)\right] - R^*_{P_t} + \frac{(C+c)\mathrm{KL}(\rho\|\pi)}{n}\right\} \\
&\leq 2\inf_{\rho\in\{\delta_\vartheta|\vartheta\in\Theta\}}\left\{\mathbb{E}_{\theta\sim\rho}\left[R_{P_t}(\theta)\right] - R^*_{P_t} + \frac{(C+c)\mathrm{KL}(\rho\|\pi)}{n}\right\} \\
&= \frac{2(C+c)\log M}{n}.
\end{aligned}$$

In the meta-learning case, an application of Theorem 5 gives

$$\begin{aligned}
\mathbb{E}_{P_{1:T}}\mathbb{E}_{\mathcal{S}_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4\inf_{1\leq m\leq M}\inf_{|A|=m}\mathbb{E}_{P_{T+1}}\Bigg[&\inf_{\theta\in A}\left\{R_{P_{T+1}}(\theta) - R^*_{P_{t+1}}\right\} \\
&+ \frac{(C+c)\log(m)}{n} + (1+8e)C\frac{m\log 2 + \log\binom{M}{m}}{T}\Bigg].
\end{aligned}$$

Using (9), the choice $A = A^*$ of cardinality m^* yields

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{\mathcal{S}_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] \leq \mathcal{E}^* + \frac{4(C+c)\log(m^*)}{n} + (1+8e)C\frac{4m^*\log(2) + 4\log\binom{M}{m^*}}{T}.$$

We conclude by using the classic bound $\log\binom{M}{m} \leq m\log\frac{Me}{m}$. ■

5.2 Learning Gaussian priors

In the applications developed from here on (models of Gaussians and mixtures of Gaussians), we will further make the following assumption: there exists $L > 0$ such that, for any $P_t \sim \mathcal{P}$ and any $\theta \in \Theta$,

$$R_{P_t}(\theta) - R^*_{P_t} \leq L\|\theta - \theta^*_t\|^2. \tag{10}$$

Intuitively, Taylor's expansion of R_{P_t} gives

$$R_{P_t}(\theta) = R_{P_t}(\theta^*_t) + \underbrace{dR_{P_t}(\theta^*_t).(\theta - \theta^*_t)}_{0} + O(\|\theta - \theta^*_t\|^2) = R_{P_t}(\theta^*_t) + O(\|\theta - \theta^*_t\|^2),$$

and thus, we can expect (10) to be satisfied when the risk is smooth enough. However, note that this assumption is not necessary for the main results of this paper to hold.

In this subsection, we consider the set of all Gaussian distributions

$$\mathcal{M} = \left\{ p_{\mu,\sigma^2} = \bigotimes_{i=1}^{d} \mathcal{N}(\mu_i, \sigma_i^2), \mu = (\mu_1, \ldots, \mu_d) \in \mathbb{R}^d, \sigma^2 = (\sigma_1^2, \ldots, \sigma_d^2) \in (\mathbb{R}_+^*)^d \right\}. \tag{11}$$

Thus, distributions on priors are actually defined as priors on μ and σ^2, and we focus on the set $\mathcal{G}$ of distributions on priors, defined as

$$\mathcal{G} = \left\{ q_{\tau,\xi^2,b} = \bigotimes_{i=1}^{d} \mathcal{N}(\tau_i, \xi_i^2) \otimes \Gamma(2, b) \right\}, \tag{12}$$

and we choose the prior on priors as $\Lambda = q_{0,\bar{\bar{\xi}}^2,\bar{\bar{b}}}$, for some parameters $\bar{\bar{\xi}}^2, \bar{\bar{b}}$.

From now on and until the end of this section, we assume that both (10) and Assumption 1 (Bernstein's condition) hold, and we are looking for a distribution on priors Π from which to sample π, such that $\rho_{T+1}(\pi, \alpha)$ concentrates as much as possible to the best parameter. Denoting $\mu^* := \mathbb{E}_{P_{T+1}}[\theta^*_{T+1}]$ and $\Sigma(\mathcal{P}) := \mathbb{E}_{P_{T+1}}\left[\|\theta^*_{T+1} - \mu^*\|^2\right]$, the following holds.

Proposition 9 *Under Assumption 1,* (3) *and* (10)*, the excess risk of* $\hat{\Pi}$ *defined in* (8) *for* $\mathcal{M}$ *and* $\mathcal{G}$ *defined as in* (11) *and* (12) *is bounded as follows:*

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{\mathcal{S}_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq \mathrm{CV}_{\mathrm{Gaussian}}(d, \Sigma(\mathcal{P}), n, T) + O\left(\frac{d\log T}{T}\right),$$

where

$$\mathrm{CV}_{\mathrm{Gaussian}}(d, \Sigma, n, T) = \inf_{1\leq b\leq T}\left\{ \frac{2(c+C)d}{n}\log\left(\frac{4nL}{b(c+C)} + 1\right) + \frac{2(c+C)b\Sigma}{n} \right\}.$$

In particular, the following convergence regimes are identifiable:

- *if* $\Sigma(\mathcal{P}) \geq \frac{d}{n}$, $\mathrm{CV}_{\mathrm{Gaussian}}(d, \Sigma(\mathcal{P}), n, T) \leq \frac{2(c+C)d}{n}\log\left(\frac{4nL}{c+C} + 1\right) + \frac{2(c+C)\Sigma(\mathcal{P})}{n}$ *(obtained for* $b = 1$*);*
- *if* $\frac{d}{n} \geq \Sigma(\mathcal{P}) \geq \frac{dn}{T^2}$, $\mathrm{CV}_{\mathrm{Gaussian}}(d, \Sigma(\mathcal{P}), n, T) \leq (8L + 2(c+C))\sqrt{\frac{d\Sigma(\mathcal{P})}{n}}$ *(obtained for* $b = \sqrt{dn\Sigma(\mathcal{P})^{-1}}$*);*

- *if* $\frac{dn}{T^2} \geq \Sigma(\mathcal{P}),\ \mathrm{CV}_{\mathrm{Gaussian}}(d, \Sigma(\mathcal{P}), n, T) \leq \frac{8Ld}{T} + \frac{2(c+C)d}{T}$ *(obtained for* $b = T$*).*

The detailed proof of this proposition is given in Appendix G.

Let us briefly analyze the above bound. In the favorable case $\Sigma(\mathcal{P}) \leq \frac{dn}{T^2}$, a proper choice of meta predictor $\hat{\Pi}$ leads to the very fast rate of convergence $O\left(\frac{d}{T}\right)$, which considerably improves upon the fast rate $O\left(\frac{d}{n} + \frac{d}{T}\right)$ when $n \ll T$. On the other hand, when $\Sigma(\mathcal{P}) \geq \frac{d}{n}$, the gain of the meta learning is undermined due to many variations across the tasks. In that case, the excess risk of the meta learning and the learning in isolation are similar, up to a $O\left(\frac{d \log T}{T}\right)$ term, which we interpret as the cost of the meta learning, and which is much smaller than the main $O\left(\frac{d \log n}{n}\right)$ term of the excess risk.

5.3 Learning Mixtures of Gaussian Priors

In this subsection, we generalize the result of the previous subsection to priors that are mixtures of Gaussians. We still assume that Assumption 1 (Bernstein's condition), (3) and (10) hold. We first assume that the number of components K in the mixture is known. Under these hypotheses, the set of possible priors π is

$$\mathcal{M} = \left\{ p_{w,\mu,\sigma^2} = \sum_{k=1}^{K} w_k \bigotimes_{i=1}^{d} \mathcal{N}(\mu_{k,i}, \sigma_{k,i}^2) : \forall k \in [K], w_k \geq 0, 1^\top w = 1 \right\}, \tag{13}$$

where $[K] \triangleq \{1, \ldots, K\}$. We add a Dirichlet prior $\mathrm{Dir}(\delta)$ on the weights w of the components in the mixture, and the set of distributions on priors becomes

$$\mathcal{G} = \left\{ q_{\delta,\tau,\xi^2,b} = \mathrm{Dir}(\delta) \otimes \bigotimes_{\substack{k\in[K]\\ i\in[d]}} \mathcal{N}(\tau_{k,i}, \xi_k^2) \otimes \bigotimes_{k=1}^{K} \Gamma(2, b_k) : \delta = (\delta_1, \ldots, \delta_K) \in \mathbb{R}^K \right\}, \tag{14}$$

while the prior on priors is chosen as $\Lambda = q_{\mathbb{I}_K, 0, \bar{\bar{\xi}}^2, \bar{\bar{b}}}$, for some parameters $\bar{\bar{\xi}}^2, \bar{\bar{b}}$ and $\mathbb{I}_K \triangleq (1, \ldots, 1)^\top \in \mathbb{R}^K$. We define

$$\Sigma_K(\mathcal{P}) := \inf_{\tau_1,\ldots,\tau_K} \mathbb{E}_{P_{T+1}} \left[\min_{k\in[K]} \|\theta_{T+1}^* - \tau_k\|^2 \right].$$

Proposition 10 *Under Assumption 1,* (3) *and* (10)*, the excess risk of* $\hat{\Pi}$ *defined in* (8) *for* $\mathcal{M}$ *and* $\mathcal{G}$ *defined in* (13) *and* (14) *is bounded as follows:*

$$\mathbb{E}_{P_{1:T}} \mathbb{E}_{\mathcal{S}_{1:T}} \mathbb{E}_{\pi\sim\hat{\Pi}} [\mathcal{E}(\pi)] - \mathcal{E}^* \leq \mathrm{CV}_{\mathrm{finite}}(K, n) + \mathrm{CV}_{\mathrm{Gaussian}}\left(d, \Sigma_K(\mathcal{P}), n, T\right) + \mathrm{CV}_{\mathrm{meta}}(T, n, d, K, \bar{\bar{b}}, \bar{\bar{\xi}}^2),$$

$$\textit{where}\ \ \mathrm{CV}_{\mathrm{finite}}(K, n) = \frac{4(c+C)\log(2K)}{n}; \quad \mathrm{CV}_{\mathrm{meta}}(T, n, d, K, \bar{\bar{b}}, \bar{\bar{\xi}}^2) = O\left(\frac{dK \log T}{T}\right).$$

Let us analyze each of the terms of the above bound. The first term $\mathrm{CV}_{\mathrm{finite}}(K, n)$ is the bound we had in the finite case of Subsection 5.1. Visualizing our K mixtures as the K points in the finite

case, this term is the time required by our estimator to select the right mixture. While this term makes the convergence rate of $O\left(\frac{1}{n}+\frac{1}{T}\right)$ notably worse than the $O\left(\frac{1}{T}\right)$ we might hope for, it is essentially unavoidable as it appears in the much simpler model of a finite set of K parameters described in Subsection 5.1.

The next term $\mathrm{CV}_{\text{Gaussian}}\left(d, \Sigma_K(\mathcal{P}), n, T\right)$ is exactly the main term of the bound obtained in the Gaussian case of Subsection 5.2, with the exception that $\Sigma(\mathcal{P})$ is replaced by $\Sigma_K(\mathcal{P})$, and scales with the convergence time to the best Gaussian for every task t.

Eventually, the last term $\mathrm{CV}_{\text{meta}}(T, n, d, K, \bar{\bar{b}}, \bar{\bar{\xi}}^2, \tau)$ is the convergence term at the meta level and is a $O\left(\frac{dK\log T}{T}\right)$. This is the cost of the meta learning compared to the learning in isolation. Since this term is very small in the $T \gg n$ regime, it enables to sometimes shrink the rate of the Gaussian term from $O\left(\frac{1}{n}\right)$ down to $O\left(\frac{1}{T}\right)$ in the most favorable cases, hence justifying the use of the meta learning.

Remark 11 *One may think, looking at the bounds in the two previous cases considered, that the $O\left(\frac{1}{n}+\frac{1}{T}\right)$ convergence rate in the case of mixtures of Gaussians is slower than the one for Gaussians which can be as fast as $O\left(\frac{1}{T}\right)$. In reality, the rate of convergence is (naturally) faster for the model of mixtures of Gaussians, because in the case of mixtures of Gaussians, the convergence term $CV_{Gaussian}\left(d, \Sigma_K(\mathcal{P}), n, T\right)$ is $O\left(\frac{1}{T}\right)$ under the assumption that $\Sigma_K(\mathcal{P}) \leq \frac{dn}{T^2}$, while in the Gaussian case, the much stronger assumption $\Sigma(\mathcal{P}) = \Sigma_1(\mathcal{P}) \leq \frac{dn}{T^2}$ is required. Under this assumption, the similar rate $O\left(\frac{1}{T}\right)$ is naturally achieved.*

Remark 12 *We would like to draw a parallel between the Gaussian mixture prior approach and the conditional meta-learning (CML) framework (Denevi et al., 2020; Wang et al., 2020; Denevi et al., 2022). In contrast to standard meta-learning, CML enhances adaptability by conditioning learning on specific variables, such as task attributes or environmental contexts, allowing adaptive strategies to be tailored to task requirements. This guarantees not only generalization of the model across tasks, but also adaptation of strategies to unique environmental conditions, just as mixture models and their ability to represent complex data distributions through combinations of simpler distributions. Thus, learning a mixture of priors is conceptually similar to CML. Note however that the hypotheses used in theoretical works on CML are based on convexity (Denevi et al., 2020, 2022), which is fundamentally different from the ones used in our PAC-Bayes analysis. It might be interesting to develop a unified framework that encompasses both approaches and produces results for both settings.*

We now consider the case when the number of mixtures K is unknown. The set of priors hence becomes the set of all (finite) mixtures of Gaussians:

$$\mathcal{M} = \left\{ p_{w,\mu,\sigma^2} = \sum_{k=1}^{+\infty} w_k \bigotimes_{i=1}^{d} \mathcal{N}(\mu_{k,i}, \sigma_{k,i}^2) : \exists K \geq 1 : \forall k \geq K+1, w_k = 0 \right\}. \tag{15}$$

In the definition of the set of distributions on priors $\mathcal{G}$, we assume that $K \leq T$ (otherwise, there is a high chance of overfitting). We then set a Dirichlet prior q_x on the number of components K, and given K, set the same model as before, denoted by $q_{\delta,\tau,\xi^2,b|K}$. Formally,

$$\mathcal{G} = \left\{ q_{x,\delta,\tau,\xi^2,b} = q_x \times q_{\delta,\tau,\xi^2,b|K} \right\}, \tag{16}$$

and we set the prior on priors $\Lambda = q_{\frac{1}{T}\mathbb{I}_T, \mathbb{I}_K, 0, \bar{\bar{\xi}}^2, \bar{\bar{b}}}$. An application of Theorem 5 gives the next bound.

Proposition 13 *Under the same conditions and using the same notations as in Proposition 10, the excess risk of* $\hat{\Pi}$ *defined in* (8) *for* $\mathcal{M}$ *and* $\mathcal{G}$ *defined as in* (15) *and* (16) *is bounded as follows:*

$$\mathbb{E}_{P_{1:T}} \mathbb{E}_{\mathcal{S}_{1:T}} \mathbb{E}_{\pi \sim \hat{\Pi}} [\mathcal{E}(\pi)] - \mathcal{E}^* \\ \leq \inf_{K \in [T]} \left\{ \mathrm{CV}_{\mathrm{finite}}(K, n) + \mathrm{CV}_{\mathrm{Gaussian}}\left(d, \Sigma_K(\mathcal{P}), n, T\right) + \mathrm{CV}_{\mathrm{meta}}^{\mathrm{unknown}}(T, n, d, K, \bar{\bar{b}}, \bar{\bar{\xi}}^2) \right\},$$

where the convergence term at the meta level becomes

$$\mathrm{CV}_{\mathrm{meta}}^{\mathrm{unknown}}(T, n, d, K, \bar{\bar{b}}, \bar{\bar{\xi}}^2) = \mathrm{CV}_{\mathrm{meta}}(T, n, d, K, \bar{\bar{b}}, \bar{\bar{\xi}}^2) + \frac{2(1+8e)C \log T}{T}.$$

Our estimator takes $\frac{2(1+8e)C \log T}{T}$ to find the optimal number of mixtures at the meta level. This is the price to pay to have the infimum on K in the bound. When $T \gg n$ and when no prior information on K is available, this clearly improves upon the bound of Proposition 10, and hence justifies setting a prior on K at the meta level rather than choosing an isolated K, pleading again in favor of the meta learning. The proof of all the results of this subsection is given in Appendix H.

6. Discussion

In recent years, the statistical guarantees of meta-learning have received increasing attention. In the following paragraphs, we present a short review of the literature on the statistical theory of meta-learning, followed by a brief discussion of three papers that are closely related to our analysis.

6.1 Theoretical Bounds in Meta-Learning

The first theoretical analysis of meta-learning goes back to Baxter (2000), who introduced the notion of task environment and derived a uniform generalization bound based on the capacity and covering number of the model. Following this i.i.d. task environment setting, many other generalization bounds have since been provided for different strategies and proof techniques, including VC theory (Baxter, 2000; Ben-David and Schuller, 2003; Maurer, 2009; Maurer et al., 2016; Guan and Lu, 2022b), algorithmic stability (Maurer and Jaakkola, 2005; Chen et al., 2020; Al-Shedivat et al., 2019; Guan et al., 2022) and information theory (Jose and Simeone, 2021; Chen et al., 2021; Jose et al., 2021; Rezazadeh et al., 2021; Hellström and Durisi, 2022). A related approach for deriving such bounds is based on PAC-Bayes theory. First proposed in the meta-learning framework in the pioneering paper of Pentina and Lampert (2014), this idea of learning a hyper-posterior that generates a prior for the new task has been taken up several times in the recent years (Amit and Meir, 2018; Ding et al., 2021; Liu et al., 2021; Rothfuss et al., 2021; Farid and Majumdar, 2021; Rothfuss et al., 2023; Guan and Lu, 2022a; Rezazadeh, 2022). In particular, Amit and Meir (2018) derived a new PAC-Bayes bound, which they applied to the optimization of deep neural networks, albeit with computational limitations. This latter concern was partially addressed by Rothfuss et al. (2021), who also specified the hyper-posterior and extended the results to unbounded losses, and further investigated their study in (Rothfuss et al., 2023). Some papers combined ideas from different literatures, such as Farid and Majumdar (2021), who explored the link between PAC-Bayes and

uniform stability in meta-learning, and provided a precise analysis of stability and generalization. Excess risk bounds have also been provided in the i.i.d. task environment framework, see (Maurer et al., 2016; Denevi et al., 2018a,b, 2019a,b, 2020; Balcan et al., 2019; Bai et al., 2021; Chen and Chen, 2022). The task environment assumption has recently been challenged, for example by Du et al. (2020) and Tripuraneni et al. (2021), who proposed to use assumptions on the distributional similarity between the features and the diversity of tasks to control the excess risk, an idea further explored by Fallah et al. (2021) who exploited a notion of diversity between the new task and training tasks using the total variation distance. Finally, a detailed analysis of regret bounds in lifelong learning has been carried out in recent years (Alquier et al., 2017; Denevi et al., 2018b, 2019b; Balcan et al., 2019; Khodak et al., 2019; Finn et al., 2019; Meunier and Alquier, 2021).

6.2 Comparison to (Denevi et al., 2019a)

and (Denevi et al., 2019a) is probably the study that is the most related to our paper. The authors provide statistical guarantees for Ridge regression with a meta-learned bias, and focus on the usefulness of their strategy relative to single-task learning, proving that their method outperforms the standard ℓ_2-regularized empirical risk minimizer. In particular, they can achieve an excess risk rate of order $O\left(1/\sqrt{T}\right)$ in the favorable case $\Sigma(\mathcal{P}) \leq \frac{n}{T}$, where $\Sigma(\mathcal{P})$ is a variance term similar to the one we defined in our Gaussian example.

6.3 Comparison to (Guan et al., 2022)

To the best of our knowledge, Guan et al. (2022) is the only work in the meta-learning literature that addresses fast rates with respect to the number of tasks T under the task environment assumption. However, we actually show in our paper that there is no need to extend Bernstein's condition when using exact Bayesian inference and that the final posterior naturally satisfies the extended Bernstein assumption, thus giving fast rates with respect to T, while Guan et al. (2022) require an additional Polyak-Łojasiewicz condition to achieve fast rates. Furthermore, their analysis is very different in nature, relying on stability arguments to derive generalization bounds, while we use PAC-Bayes theory to control the excess risk.

6.4 Comparison to (Guan and Lu, 2022a) and (Rezazadeh, 2022)

Finally, Guan and Lu (2022a) and Rezazadeh (2022) provide fast rate generalization bounds based on another version of Catoni's PAC-Bayes bound. While these are indeed very nice results, the cost is that the empirical risk in the right-hand side of the bound is multiplied by a factor $c > 1$. In terms of excess risk, this leads to fast rates only in the case where the optimal risk is null. On the other hand, if the optimal risk is positive, this would not even prove consistency. To reduce the factor in front of the empirical risk to 1 would return a slow rate.

7. Conclusion and Open Problems

We provided an analysis of the excess risk in meta-learning the prior via PAC-Bayes bounds. Surprisingly, at the meta-level, conditions for fast rates are always satisfied if one uses exact Gibbs posteriors at the task level. An important problem is to extend this result to variational approximations of Gibbs posteriors.

Acknowledgments

The three authors thank the anonymous Reviewers and the AE for their very constructive comments. BECA acknowledges funding from the ANR grant project BACKUP ANR-23-CE40-0018-01.

Appendix A. Notations

The following is a complete list of all the notations introduced throughout the paper.

Name	Definition	Page
Ensembles		
Observation space	$\mathcal{Z}$	p4
Decision space	Θ	p4
Set of all probability distributions on Θ	$\mathcal{P}(\Theta)$	p4
Subset of potential posteriors (often chosen to be parametric)	$\mathcal{F} \subseteq \mathcal{P}(\Theta)$	p5
Set of priors π	$\mathcal{M}$	p11
Set of all probability distributions on $\mathcal{M}$	$\mathcal{P}(\mathcal{M})$	p11
Subset of potential posteriors on $\mathcal{P}(\mathcal{M})$	$\mathcal{G} \subseteq \mathcal{P}(\mathcal{M})$	p11
Loss function		
Loss function	$\ell : \mathcal{Z} \times \Theta \to \mathbb{R}_+$	p4
Observations and Data Distributions at the within-task level		
Sample size of task t	N_t	p4
Observations (i.i.d.) of task t	$Z_{t,1}, \ldots, Z_{t,N_t}$	p4
Sample of (i.i.d.) observations of task t	$\mathcal{S}_t \triangleq (Z_{t,1}, \ldots, Z_{t,N_t})$	p7
Distribution of the (i.i.d.) observations of task t	P_t	p5
Distributions at the meta-level		
Distribution of $(P_1, N_1), \ldots, (P_T, N_T)$ when they are assumed to be i.i.d.	$\mathcal{P}$	p8
Expected sample size	n	p11
Expectations		
Expectation with respect to one observation from task t: $Z \sim P_t$	$\mathbb{E}_{Z \sim P_t}$	p5
Expectation with respect to the observations of task t: $\mathcal{S}_t = (Z_{t,1}, \ldots, Z_{t,N_t})$ i.i.d. from P_t, conditionally on (P_t, N_t)	$\mathbb{E}_{\mathcal{S}_t}$	p7

Expectation with respect to the observations of all the tasks $t \in \{1, \dots, T\}$: $\mathcal{S}_1 \sim P_1, \dots, \mathcal{S}_T \sim P_T$, conditionally on $(P_1, N_1), \dots, (P_T, N_T)$	$\mathbb{E}_{\mathcal{S}_{1:T}} = \mathbb{E}_{\mathcal{S}_1} \dots \mathbb{E}_{\mathcal{S}_T}$	p11
Expectation with respect to the observations of an out-of-sample task $T+1$: $\mathcal{S}_{T+1} \sim P_{T+1}$, conditionally on (P_{T+1}, N_{T+1})	$\mathbb{E}_{\mathcal{S}_{T+1}}$	p8
Expectation with respect to a task $t \in \{1, \dots, T+1\}$: $(P_t, N_t) \sim \mathcal{P}$	$\mathbb{E}_{P_t, N_t}$	p8
Expectation with respect to all the tasks $t \in \{1, \dots, T\}$: $(P_1, n_1), \dots, (P_T, N_t) \sim \mathcal{P}$	$\mathbb{E}_{P_{1:T}, N_{1:T}}$	p11
Short for $\mathbb{E}_{P_{1:T}, N_{1:T}}$ when $N = 1 = \dots = N_T = n$ are deterministic	$\mathbb{E}_{P_{1:T}}$	p12
Expectation with respect to a parameter θ sampled from a distribution ρ: $\theta \sim \rho$	$\mathbb{E}_{\theta \sim \rho}$	p5
Expectation with respect to a prior π sampled from a prior on priors Π: $\pi \sim \Pi$	$\mathbb{E}_{\pi \sim \Pi}$	p10
Parameters		
Parameter of Gibbs posterior at the within-task level	α	p5
Parameter of Gibbs distribution on priors at the meta level	β	p11
Bound on the loss	C	p6
Bernstein's condition's constant (common to all tasks)	c	p6
Empirical and Expected Risks at the within-task level		
Prediction Risk of task t	$R_{P_t}(\theta) \triangleq \mathbb{E}_{Z \sim P_t}[\ell(Z, \theta)]$	p5
Minimizer of the prediction risk at task t (existence implicitly assumed)	$\theta_t^* \triangleq \arg\min_{\theta \in \Theta} R_{P_t}(\theta)$	p5
Optimal prediction risk of task t	$R_{P_t}^* \triangleq \inf_{\theta \in \Theta} R_{P_t}(\theta) = R_{P_t}(\theta_t^*)$	p5
Empirical Risk at task t	$\hat{R}_t(\theta) \triangleq \frac{1}{N_t} \sum_{i=1}^{N_t} \ell(Z_{t,i}, \theta)$	p5
Surrogate risk estimate of posterior ρ with prior π and parameter α at task t	$\hat{B}_t(\rho, \pi, \alpha) \triangleq \mathbb{E}_{\theta \sim \rho}\left[\hat{R}_t(\theta)\right] + \frac{\mathrm{KL}(\rho \Vert \pi)}{\alpha N_t}$	p5

Surrogate risk estimate of Gibbs posterior with prior π and parameter α at task t	$\widehat{\mathcal{L}}_t(\pi,\alpha) \triangleq \hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha)$	p9
Expected surrogate risk of Gibbs posterior with prior π and parameter α at task t	$\mathcal{L}_t(\pi,\alpha) \triangleq \mathbb{E}_{\mathcal{S}_t}\left[\widehat{\mathcal{L}}_t(\pi,\alpha)\right]$	p9
Minimizer of the expected surrogate risk of Gibbs posterior with prior π and parameter α	$\pi^*_\alpha \triangleq \arg\min_\pi \mathbb{E}_{P_t,N_t}\left[\mathcal{L}_t(\pi,\alpha)\right]$	p9
Expected variability of the loss at task t with respect to θ	$V_t(\theta,\theta^*_t) \triangleq \mathbb{E}_{Z\sim P_t}\left[\lvert\ell(Z,\theta)-\ell(Z,\theta^*_t)\rvert^2\right]$	p6
Empirical and Expected Risks at the meta level		
Meta risk	$\mathcal{E}(\pi) \triangleq \mathbb{E}_{P_{T+1},N_{T+1}}\mathbb{E}_{\mathcal{S}_{T+1}}\mathbb{E}_{\theta\sim\rho_{T+1}(\pi,\alpha)}[R_{P_{T+1}}$	p8
Oracle meta risk	$\mathcal{E}^* \triangleq \mathbb{E}_{P_{T+1},N_{T+1}}[R^*_{P_{T+1}}]$	p8
Prior and Bayesian Estimators at the within-task level		
Within-task Prior (common to all tasks)	π	p5
Gibbs posterior at task t with prior π and parameter α	$\rho_t(\pi,\alpha) \triangleq \operatorname{argmin}_{\rho\in\mathcal{P}(\Theta)} \hat{B}_t(\rho,\pi,\alpha)$	p5
Variational approximation on $\mathcal{F}$ of Gibbs posterior at task t with prior π and parameter α	$\rho_t(\pi,\alpha,\mathcal{F}) \triangleq \operatorname{argmin}_{\rho\in\mathcal{F}} \hat{B}_t(\rho,\pi,\alpha)$	p5
Prior and Bayesian Estimators at the meta level		
Prior on the set of priors	Λ	p11
Gibbs distribution on priors	$\hat{\Pi} \triangleq \operatorname{argmin}_{\Pi\in\mathcal{G}}\left\{\frac{\mathrm{KL}(\Pi\|\Lambda)}{\beta T}\right.$ $\left.+\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\pi\sim\Pi}\left[\hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha)\right]\right\}$	p11
Variational approximation of Gibbs distribution on priors based on the variational approximations $\rho_t(\pi,\alpha,\mathcal{F})$	$\hat{\Pi}(\mathcal{F}) \triangleq \operatorname{argmin}_{\Pi\in\mathcal{G}}\left\{\frac{\mathrm{KL}(\Pi\|\Lambda)}{\beta T}\right.$ $\left.+\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\pi\sim\Pi}\left[\hat{B}_t(\rho_t(\pi,\alpha,\mathcal{F}),\pi,\alpha)\right]\right\}$	p12
Applications to the Models of Gaussians and Mixtures of Gaussians		
Lipschitz constant of the loss (only used when the loss is assumed to be Lipschitz, in some applications)	L	p15
Parameter α of Gibbs posterior at the observation level	$\alpha = \frac{1}{c+C}$	p7
Parameter β of the prior on priors at the meta level	$\beta = \frac{1}{(1+8e)C}$	p11
Expectation of the optimal parameter	$\mu^* \triangleq \mathbb{E}_{P_{T+1}}[\theta^*_{T+1}]$	p15

Distributions and Parameters in the Gaussian Model		
Joint distribution of d i.i.d. Gaussians of expectations $\mu_1, \dots, \mu_d$ and variances $\sigma_1^2, \dots, \sigma_d^2$	$p_{\mu,\sigma^2} \triangleq \bigotimes_{i=1}^d \mathcal{N}(\mu_i, \sigma_i^2)$	p15
Set of priors	$\mathcal{M} \triangleq \left\{p_{\mu,\sigma^2} : \forall i \in [d], \mu_i \in \mathbb{R}, \sigma_i^2 > 0\right\}$	p15
General parametric form of prior on (μ, σ^2) as the joint distribution of d i.i.d. Gaussians of expectations $\tau_1, \dots, \tau_d$ and variances $\xi_1^2, \dots, \xi_d^2$, and an independent gamma distribution of parameters $(2, b)$	$q_{\tau,\xi^2,b} \triangleq \left(\bigotimes_{i=1}^d \mathcal{N}(\tau_i, \xi_i^2)\right) \otimes \Gamma(2, b)$	p15
Set of distributions on priors	$\mathcal{G} \triangleq \left\{q_{\tau,\xi^2,b} : \forall i \in [d], \tau_i \in \mathbb{R}, \xi_i^2 > 0, b > 0\right\}$	p15
Prior on priors	$\Lambda \triangleq q_{0,\bar{\xi}^2,\bar{b}}$	p15
Parameters of the prior on prior Λ	$\bar{\xi}^2, \bar{b}$	p15
Posterior of priors in the bound $\hat{\Pi}$	$\hat{\Pi} \triangleq q_{\hat{\tau},\hat{\xi}^2,\hat{b}}$	p34
Parameters of $\hat{\Pi}$	$\hat{\tau}, \hat{\xi}^2, \hat{b}$	p34
Variance of the optimal parameter	$\Sigma(\mathcal{P}) \triangleq \mathbb{E}_{P_{T+1}}\left[\|\theta^*_{T+1} - \mu^*\|^2\right]$	p15
Model of Mixtures of Gaussians with Known Number of Mixtures		
Number of mixtures	K	p16
Mixture of joint distribution of d i.i.d. Gaussians	$p_{w,\mu,\sigma^2} \triangleq \sum_{k=1}^K w_k \bigotimes_{i=1}^d \mathcal{N}(\mu_{k,i}, \sigma_{k,i}^2)$	p16
Set of priors	$\mathcal{M} \triangleq \{p_{w,\mu,\sigma^2} : \mu_{k,i} \in \mathbb{R}, \sigma_{k,i}^2 > 0, w_k \geq 0, 1^\top w = 1\}$	p16
General parametric form of prior on (w, μ, σ^2)	$q_{\delta,\tau,\xi^2,b} \triangleq \mathrm{Dir}(\delta) \otimes \left(\bigotimes_{\substack{i\in[d]\\k\in[K]}} \mathcal{N}(\tau_{k,i}, \xi_k^2)\right) \otimes \left(\bigotimes_{k\in[K]} \Gamma(2, b_k)\right)$	p16
Set of distributions on priors	$\mathcal{G} \triangleq \{q_{\delta,\tau,\xi^2,b} : \forall (i,k) \in [d] \times [K], \delta_k > 0, \tau_{k,i} \in \mathbb{R}, \xi_k^2 > 0, b_k > 0\}$	p16
Prior on priors	$\Lambda \triangleq q_{1_K,0,\bar{\xi}^2,\bar{b}}$	p16
Parameters of the prior on prior Λ	$\bar{\xi}^2 = (\bar{\xi}_1^2, \dots, \bar{\xi}_K^2), \bar{b} = (\bar{b}_1, \dots, \bar{b}_K)$	p16
K-Variance of the optimal parameter	$\Sigma_K(\mathcal{P}) \triangleq \inf_{\tau_1,\dots,\tau_K} \mathbb{E}_{P_{T+1}}\left[\min_{k\in[K]} \|\theta^*_{T+1} - \tau_k\|^2\right]$	p16
Model of Mixtures of Gaussians with Unknown Finite Number of Mixtures		
Mixture of joint distribution of d i.i.d. Gaussians	$p_{w,\mu,\sigma^2} \triangleq \sum_{k=1}^\infty \bigotimes_{i=1}^d w_k \mathcal{N}(\mu_{k,i}, \sigma_{k,i}^2)$	p17
Set of priors	$\mathcal{M} \triangleq \{p_{w,\mu,\sigma^2} : \exists K : \forall k \geq K+1, w_k = 0\}$	p17

Prior on the number of mixtures K	$q_x \triangleq \mathrm{Mult}(x_1, \ldots, x_T)$	p17
Prior on the parameters given the number of mixtures	$q_{\delta,\tau,\xi^2,b\|K} \triangleq \mathrm{Dir}(\delta) \otimes \left(\bigotimes_{\substack{i\in[d]\\k\in[K]}} \mathcal{N}(\tau_{k,i}, \xi_k^2)\right) \otimes \left(\bigotimes_{k\in[K]} \Gamma(2, b_k)\right)$	p17
General parametric form of prior on w, μ, σ^2	$q_{x,\delta,\tau,\xi^2,b} \triangleq q_x \times q_{\delta,\tau,\xi^2,b\|K}$	p17
Set of priors on priors	$\mathcal{G} \triangleq \{q_{x,\delta,\tau,\xi^2,b}\}$	p17
Prior on priors	$\Lambda \triangleq q_{\frac{1}{T}1_T,1_K,0,\bar{\xi}^2,\bar{b}}$	p18
Probability Distributions		
Multinomial distribution of support $[K]$ and parameters $x = (x_1, \ldots, x_K)$	$\mathrm{Mult}(x)$	p25
Gaussian distribution of mean μ and variance σ^2	$\mathcal{N}(\mu, \sigma^2)$	p25
Gamma distribution of parameters (a, b)	$\Gamma(a, b)$	p25
Dirichlet distribution of support $[K]$ and parameters $\delta = (\delta_1, \ldots, \delta_K)$	$\mathrm{Dir}(\delta)$	p25
Miscellaneous		
First K positive integers	$[K] \triangleq \{1, \ldots, K\}$	p16
Vector of ones	$1_K \triangleq (1, \ldots, 1)^\top \in \mathbb{R}^K$	p16
Kullback-Leibler divergence between distributions ρ and π	$\mathrm{KL}(\rho\|\pi)$	p25
Dirac distribution of parameter ϑ	$\delta_\vartheta(x) \triangleq \begin{cases} 1 & \text{if } x = \vartheta \\ 0 & \text{otherwise.} \end{cases}$	N/A
Indicator function of the validity of Bernstein's condition	$\mathbb{I}_B \triangleq \begin{cases} 1 & \text{if Bernstein's condition holds} \\ 0 & \text{otherwise.} \end{cases}$	p7
Entropy of distribution $\mathrm{Mult}(x)$	$H(x)$	p25
Gamma function	$\Gamma(x) \triangleq \int_0^\infty t^{x-1}\mathrm{e}^{-t}dt$	p25
Digamma function	$\psi(x) \triangleq \frac{\Gamma'(x)}{\Gamma(x)}$	p25

Appendix B. Some Useful Formulas

The following (known) results are used throughout the text. They are recalled here without proof.

B.1 Concentration Inequalities

For Hoeffding and Bernstein see (Boucheron et al., 2013). For Donsker and Varadhan, see for example (Catoni, 2007).

Lemma 14 (Hoeffding's inequality) *Let $U_1, \dots, U_n$ be i.i.d random variables taking values in an interval $[a, b]$. Then, for any $s > 0$,*

$$\mathbb{E}\left[\mathrm{e}^{s\sum_{i=1}^n [U_i - \mathbb{E}(U_i)]}\right] \leq \mathrm{e}^{\frac{ns^2(b-a)^2}{8}}.$$

Lemma 15 (Bernstein's inequality) *Let $U_1, \dots, U_n$ be i.i.d random variables such that for any $k \geq 2$,*

$$\mathbb{E}\left[|U_i|^k\right] \leq \frac{k!}{2} V C^{k-2}. \tag{17}$$

Then, for any $s \in (0, 1/C]$,

$$\mathbb{E}\left[\mathrm{e}^{s\sum_{i=1}^n [U_i - \mathbb{E}(U_i)]}\right] \leq \mathrm{e}^{\frac{ns^2 V}{2(1-sC)}}.$$

Note that in particular, if $|U_i| \leq C$ almost surely, (17) always holds with $V = \mathbb{E}(U_i^2)$.

B.2 Donsker and Varadhan's Lemma

Lemma 16 (Donsker and Varadhan's variational inequality, 1976) *Let μ be a probability measure on Θ. For any measurable, bounded function $h : \Theta \to \mathbb{R}$, we have:*

$$\log \mathbb{E}_{\theta\sim\mu}\left[\mathrm{e}^{h(\theta)}\right] = \sup_{\rho\in\mathcal{P}(\Theta)} \Big\{\mathbb{E}_{\theta\sim\rho}[h(\theta)] - \mathrm{KL}(\rho\|\mu)\Big\}.$$

Moreover, the supremum with respect to ρ in the right-hand side is reached for the Gibbs measure μ_h defined by its density with respect to μ

$$\frac{\mathrm{d}\mu_h}{\mathrm{d}\mu}(\vartheta) = \frac{\mathrm{e}^{h(\vartheta)}}{\mathbb{E}_{\theta\sim\mu}\left[\mathrm{e}^{h(\theta)}\right]}.$$

B.3 KL Divergence of Some Known Distributions

Denoting by $H(x)$ the entropy of $(x_1, \dots, x_T)$, recall that the KL divergence between a multinomial distribution of parameters $(x_1, \dots, x_T)$ and a multinomial distribution of parameters $\left(\frac{1}{T}, \dots, \frac{1}{T}\right)$ is

$$\mathrm{KL}\left(\mathrm{Mult}(x)\|\mathrm{Mult}\left(\frac{1}{T}\right)\right) = \log T - H(x). \tag{18}$$

Recall that the KL divergence between 2 normal distributions is

$$\mathrm{KL}\left(\mathcal{N}(\mu, \sigma^2)\|\mathcal{N}(\bar{\mu}, \bar{\sigma}^2)\right) = \frac{1}{2}\left(\frac{(\mu-\bar{\mu})^2}{\bar{\sigma}^2} + \frac{\sigma^2}{\bar{\sigma}^2} - 1 + \log\frac{\bar{\sigma}^2}{\sigma^2}\right). \tag{19}$$

Recall that the KL divergence between 2 Gamma distributions is

$$\mathrm{KL}\left(\Gamma(a, b)\|\Gamma\left(\bar{\bar{a}}, \bar{\bar{b}}\right)\right) = (a - \bar{\bar{a}})\,\psi(a) + \log\frac{\Gamma(\bar{\bar{a}})}{\Gamma(a)} + \bar{\bar{a}}\log\frac{b}{\bar{\bar{b}}} + a\frac{\bar{\bar{b}} - b}{b}, \tag{20}$$

where ψ denotes the digamma function, and Γ the gamma function. Recall that the KL divergence between a Dirichlet distribution of parameter δ and a Dirichlet distribution of parameter $1_K = (1, \dots, 1)$ is

$$\mathrm{KL}\left(\mathrm{Dir}(\delta)\|\mathrm{Dir}(1_K)\right) = \log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^K\Gamma(\delta_k)} + \sum_{k=1}^K(\delta_k - 1)\left(\psi(\delta_k) - \psi(1^\top\delta)\right). \tag{21}$$

Appendix C. Proof of Theorem 1

We mostly follow the proof technique developed in (Catoni, 2007). For any $s > 0$, fix $\theta \in \Theta$ and let $U_i = \mathbb{E}[\ell(Z_{t,i}, \theta)] - \ell(Z_{t,i}, \theta) - \mathbb{E}[\ell(Z_{t,i}, \theta_t^*)] + \ell(Z_{t,i}, \theta_t^*)$ for any $i \in \{1, \dots, n\}$. We are going to distinguish two cases, whether or not Assumption 1 (Bernstein's condition) is satisfied.

If Assumption 1 is satisfied, we apply Lemma 15 to U_i. Note that in this case, V is actually the variance term $V_t(\theta, \theta_t^*)$. So, for any $s > 0$,

$$\mathbb{E}_{\mathcal{S}_t}\left[\mathrm{e}^{sn\left(R_{P_t}(\theta) - \hat{R}_t(\theta) - R^*_{P_t} + \hat{R}_t(\theta_t^*)\right)}\right] \leq \mathrm{e}^{\frac{ns^2 V(\theta, \theta_t^*)}{2(1 - sC)}}.$$

We let $s = \lambda / n$, which gives

$$\mathbb{E}_{\mathcal{S}_t}\left[\mathrm{e}^{\lambda\left(R_{P_t}(\theta) - \hat{R}_t(\theta) - R^*_{P_t} + \hat{R}_t(\theta_t^*)\right)}\right] \leq \mathrm{e}^{\frac{\lambda^2 V(\theta, \theta_t^*)}{2(n - C\lambda)}}.$$

Making use of Assumption 1 gives

$$\mathbb{E}_{\mathcal{S}_t}\left[\mathrm{e}^{\lambda\left(R_{P_t}(\theta) - \hat{R}_t(\theta) - R^*_{P_t} + \hat{R}_t(\theta_t^*)\right)}\right] \leq \mathrm{e}^{\frac{\lambda^2 c\left(R_{P_t}(\theta) - R^*_{P_t}\right)}{2(n - C\lambda)}}.$$

If Assumption 1 is not satisfied, we apply Lemma 14 to U_i, which gives, for any $s > 0$,

$$\mathbb{E}_{\mathcal{S}_t}\left[\mathrm{e}^{sn\left(R_{P_t}(\theta) - \hat{R}_t(\theta) - R^*_{P_t} + \hat{R}_t(\theta_t^*)\right)}\right] \leq \mathrm{e}^{\frac{ns^2 C^2}{8}}.$$

Letting $s = \lambda / n$ gives

$$\mathbb{E}_{\mathcal{S}_t}\left[\mathrm{e}^{\lambda\left(R_{P_t}(\theta) - \hat{R}_t(\theta) - R^*_{P_t} + \hat{R}_t(\theta_t^*)\right)}\right] \leq \mathrm{e}^{\frac{\lambda^2 C^2}{8n}}.$$

Defining the general bound

$$W = \frac{\lambda^2 c\left(R_{P_t}(\theta) - R^*_{P_t}\right)}{2(n - C\lambda)} \mathbb{I}_B + \frac{\lambda^2 C^2}{8n}(1 - \mathbb{I}_B), \tag{22}$$

where $\mathbb{I}_B$ is equal to 1 if Assumption 1 is satisfied, and 0 otherwise, it holds in either case that

$$\mathbb{E}_{\mathcal{S}_t}\left[\mathrm{e}^{\lambda\left(R_{P_t}(\theta) - \hat{R}_t(\theta) - R^*_{P_t} + \hat{R}_t(\theta_t^*)\right)}\right] \leq \mathrm{e}^{W}.$$

Rearranging the terms gives

$$\mathbb{E}_{\mathcal{S}_t}\left[\mathrm{e}^{\lambda\left(R_{P_t}(\theta) - R^*_{P_t} - \frac{W}{\lambda} - \hat{R}_t(\theta) + \hat{R}_t(\theta_t^*)\right)}\right] \leq 1.$$

Next, integrating this bound with respect to π and using Fubini's theorem to exchange both integrals gives

$$\mathbb{E}_{\mathcal{S}_t}\mathbb{E}_{\theta \sim \pi}\left[\mathrm{e}^{\lambda\left(R_{P_t}(\theta) - R^*_{P_t} - \frac{W}{\lambda} - \hat{R}_t(\theta) + \hat{R}_t(\theta_t^*)\right)}\right] \leq 1.$$

We then apply Lemma 16 to the argument of the expectation with respect to the sample, and we have

$$\mathbb{E}_{\mathcal{S}_t}\left[\mathrm{e}^{\sup_{\rho\in\mathcal{P}(\Theta)}\left\{\lambda\mathbb{E}_{\theta\sim\rho}\left[R_{P_t}(\theta)-R^*_{P_t}-\frac{W}{\lambda}-\hat{R}_t(\theta)+\hat{R}_t(\theta^*_t)\right]-\mathrm{KL}(\rho\|\pi)\right\}}\right]\leq 1.$$

Jensen's inequality implies

$$\mathrm{e}^{\lambda\mathbb{E}_{\mathcal{S}_t}\left[\sup_{\rho\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim\rho}\left[R_{P_t}(\theta)-R^*_{P_t}-\frac{W}{\lambda}-\hat{R}_t(\theta)+\hat{R}_t(\theta^*_t)\right]-\frac{\mathrm{KL}(\rho\|\pi)}{\lambda}\right\}\right]}\leq 1,$$

in other words,

$$\mathbb{E}_{\mathcal{S}_t}\left[\sup_{\rho\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim\rho}\left[R_{P_t}(\theta)-R^*_{P_t}-\frac{W}{\lambda}-\hat{R}_t(\theta)+\hat{R}_t(\theta^*_t)\right]-\frac{\mathrm{KL}(\rho\|\pi)}{\lambda}\right\}\right]\leq 0.$$

At this stage, we can replace W by its value given in (22) to obtain the bound:

$$\mathbb{E}_{\mathcal{S}_t}\left[\sup_{\rho\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim\rho}\left[\left(1-\frac{\lambda c\mathbb{I}_B}{2(n-C\lambda)}\right)\left(R_{P_t}(\theta)-R^*_{P_t}\right)\right.\right.\right.$$
$$\left.\left.\left.-\frac{\lambda C^2(1-\mathbb{I}_B)}{8n}-\hat{R}_t(\theta)+\hat{R}_t(\theta^*_t)\right]-\frac{\mathrm{KL}(\rho\|\pi)}{\lambda}\right\}\right]\leq 0.$$

Next, we rearrange the terms and replace the supremum on ρ by $\rho_t(\pi,\alpha)$:

$$\mathbb{E}_{\mathcal{S}_t}\mathbb{E}_{\theta\sim\rho_t(\pi,\alpha)}\left[R_{P_t}(\theta)\right]-R^*_{P_t}\leq\frac{1}{1-\frac{\lambda c\mathbb{I}_B}{2(n-C\lambda)}}$$
$$\times\left(\mathbb{E}_{\mathcal{S}_t}\left[\mathbb{E}_{\theta\sim\rho_t(\pi,\alpha)}\left[\hat{R}_t(\theta)\right]-\hat{R}_t(\theta^*_t)+\frac{\mathrm{KL}(\rho_t(\pi,\alpha)\|\pi)}{\lambda}\right]+\frac{\lambda C^2(1-\mathbb{I}_B)}{8n}\right).$$

We then replace λ by αn and by definition of Gibbs posterior $\rho_t(\pi,\alpha)$, the above bound is the same as

$$\mathbb{E}_{\mathcal{S}_t}\mathbb{E}_{\theta\sim\rho_t(\pi,\alpha)}\left[R_{P_t}(\theta)\right]-R^*_{P_t}\leq\frac{1}{1-\frac{\alpha c\mathbb{I}_B}{2(1-C\alpha)}}$$
$$\times\left(\mathbb{E}_{\mathcal{S}_t}\left[\inf_{\rho\in\mathcal{F}}\left\{\mathbb{E}_{\theta\sim\rho}\left[\hat{R}_t(\theta)\right]-\hat{R}_t(\theta^*_t)+\frac{\mathrm{KL}(\rho\|\pi)}{\alpha n}\right\}\right]+\frac{\alpha C^2(1-\mathbb{I}_B)}{8}\right).$$

In particular, under Assumption 1, i.e., if $\mathbb{I}_B=1$, the choice $\alpha=\frac{1}{C+c}$ gives

$$\mathbb{E}_{\mathcal{S}_t}\left[\mathbb{E}_{\theta\sim\rho_t(\pi,\alpha)}[R_{P_t}(\theta)]\right]-R^*_{P_t}\leq 2\mathbb{E}_{\mathcal{S}_t}\left[\inf_{\rho\in\mathcal{F}}\left\{\mathbb{E}_{\theta\sim\rho}\left[\hat{R}_t(\theta)\right]-\hat{R}_t(\theta^*_t)+\frac{\mathrm{KL}(\rho\|\pi)}{\alpha n}\right\}\right].$$

Without Assumption 1, i.e., if $\mathbb{I}_B=0$, rewriting the bound and taking the minimum over α yields

$$\mathbb{E}_{\mathcal{S}_t}\left[\mathbb{E}_{\theta\sim\rho_t(\pi,\alpha)}[R_{P_t}(\theta)]\right]-R^*_{P_t}\leq\inf_{\alpha}\mathbb{E}_{\mathcal{S}_t}\left[\inf_{\rho\in\mathcal{F}}\left\{\mathbb{E}_{\theta\sim\rho}\left[\hat{R}_t(\theta)\right]-\hat{R}_t(\theta^*_t)+\frac{\mathrm{KL}(\rho\|\pi)}{\alpha n}+\frac{\alpha C^2}{8}\right\}\right],$$

and this concludes the proof. ■

Appendix D. Proof of Corollary 2

As this corollary is stated for one task only, let's put $n := N_t$ for the sake of clarity. First,

$$\begin{aligned}\mathbb{E}_{\mathcal{S}_t}\left[\inf_{\rho\in\mathcal{P}(\Theta)} \hat{B}_t(\rho,\pi,\alpha) - \hat{R}_t(\theta_t^*)\right] &= \mathbb{E}_{\mathcal{S}_t}\left[\inf_{\rho\in\mathcal{P}(\Theta)} \left\{\mathbb{E}_{\theta\sim\rho}[\hat{R}_t(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\alpha n}\right\} - \hat{R}_t(\theta_t^*)\right] \\ &\leq \inf_{\rho\in\mathcal{P}(\Theta)} \mathbb{E}_{\mathcal{S}_t}\left[\mathbb{E}_{\theta\sim\rho}[\hat{R}_t(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\alpha n} - \hat{R}_t(\theta_t^*)\right] \\ &= \inf_{\rho\in\mathcal{P}(\Theta)} \left[\mathbb{E}_{\theta\sim\rho}[R_{P_t}(\theta)] - R_{P_t}^* + \frac{\mathrm{KL}(\rho\|\pi)}{\alpha n}\right].\end{aligned}$$

Now, for any $s \in (0, s_0]$, let ρ_s be the restriction of π to the set $\{\theta : R_{P_t}(\theta) - R_{P_t}^* \leq s\}$.

$$\begin{aligned}\mathbb{E}_{\mathcal{S}_t}\left[\inf_{\rho\in\mathcal{P}(\Theta)} \hat{B}_t(\rho,\pi,\alpha) - \hat{R}_t(\theta_t^*)\right] &\leq \inf_{0<s\leq s_0}\left[\mathbb{E}_{\theta\sim\rho_s}[R_{P_t}(\theta)] - R_{P_t}^* + \frac{\mathrm{KL}(\rho_s\|\pi)}{\alpha n}\right] \\ &\leq \inf_{0<s\leq s_0}\left[s + \frac{\log\frac{1}{\pi(\{\theta:R_t(\theta)-R_t(\theta_t^*)\leq s\})}}{\alpha n}\right] \\ &\leq \inf_{0<s\leq s_0}\left[s + \frac{d_{\pi,t}\log\frac{1}{s} + \log\kappa_{\pi,t}}{\alpha n}\right]\end{aligned}$$

by assumption. An optimization with respect to s leads to $s = d_{\pi,t}/(\alpha n) \leq s_0$ as soon as $n \geq d_{\pi,t}/(\alpha s_0)$ and we obtain the first statement:

$$\mathbb{E}_{\mathcal{S}_t}\left[\inf_{\rho\in\mathcal{P}(\Theta)} \hat{B}_t(\rho,\pi,\alpha) - \hat{R}_t(\theta_t^*)\right] \leq \frac{d_{\pi,t}\log\frac{ne\alpha}{d_{\pi,t}} + \log\kappa_{\pi,t}}{\alpha n}.$$

Plugging this into Theorem 1 leads immediately to the other statements.

Appendix E. Proof of Lemma 4

First, we note that $f : x \mapsto -\frac{1}{\tau}\log(x)$ is differentiable on $[\exp(-C\tau), 1]$ and $f'(x) = -\frac{1}{\tau x}$. As a consequence, $|f'(x)| = 1/(\tau x)$ is maximized at $x = \exp(-C\tau)$ and its maximum is $\exp(C\tau)/\tau$. This implies that f is $\exp(C\tau)/\tau$-Lipschitz, that is, for any $(x, y) \in [\exp(-C\tau), 1]^2$,

$$|f(x) - f(y)| \leq \frac{\exp(C\tau)}{\tau}|x - y|.$$

Taking the square of both sides of the inequality yields

$$\forall (x, y) \in [\exp(-C\tau), 1]^2,\ (f(x) - f(y))^2 \leq \frac{\exp(2C\tau)}{\tau^2}(x - y)^2. \tag{23}$$

Then, $f''(x) = 1/(\tau x^2)$ and thus, $|f''(x)| \geq 1/\tau$ (minimum reached for $x = 1$). This implies that f is $1/\tau$-strongly convex, that is, for any $(x, y) \in [\exp(-C\tau), 1]^2$ we have, for any $\theta \in [0, 1]$,

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) - \frac{\theta(1 - \theta)}{2\tau}(x - y)^2.$$

We apply this inequality to $\theta = 1/2$ and rearrange terms, and we obtain

$$\forall (x,y) \in [\exp(-C\tau), 1]^2,\ \frac{1}{8\tau}(x-y)^2 \leq \frac{f(x)+f(y)}{2} - f\left(\frac{x+y}{2}\right). \tag{24}$$

Finally, combining (23) with (24) yields

$$\forall (x,y) \in [\exp(-C\tau), 1]^2, [f(x)-f(y)]^2 \leq \frac{8\exp(2C\tau)}{\tau}\left[\frac{f(x)+f(y)}{2} - f\left(\frac{x+y}{2}\right)\right],$$

concluding the proof of the lemma. ■

Appendix F. Proof of Theorem 5

The proof of Theorem 5 is structured as follows: we first bound the excess risk by the expectation of the infimum of the empirical risk $\hat{B}_t(\rho, \pi, \alpha) - \hat{R}(\theta_t^*)$ in Lemma 17. Using classic techniques, we turn this bound into the prediction risk.

F.1 Lemma

Lemma 17 *Assume that the loss ℓ satisfies the boundedness assumption* (3) *with constant C. Then, the following bound holds with the choice $\beta = \frac{1}{(1+8e)C}$:*

$$\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}\left[\mathcal{E}(\pi)\right] - \mathcal{E}^* \leq \frac{2}{1-\frac{\alpha c \mathbb{I}_B}{2(1-C\alpha)}}\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\Bigg[$$
$$\inf_{\Pi\in\mathcal{G}}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}_{\pi\sim\Pi}\left[\hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha)\right] - \hat{R}_t(\theta_t^*)\right) + \frac{8eC\mathrm{KL}(\Pi\|\Lambda)}{T}\right\} + \frac{\alpha C^2(1-\mathbb{I}_B)}{8}\Bigg],$$

Proof For any $t \in [T]$, let

$$U_t := \hat{B}_t(\rho_t(\pi_\alpha^*,\alpha),\pi_\alpha^*,\alpha) - \hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha).$$

Note that

$$\mathbb{E}[U_t] = \mathbb{E}_{P_t,N_t}\left[\mathcal{L}_t(\pi_\alpha^*,\alpha)\right] - \mathbb{E}_{P_t,N_t}\left[\mathcal{L}_t(\pi,\alpha)\right],$$

where $\mathbb{E}[U_t]$ is a short notation for $\mathbb{E}_{P_t,N_t}\mathbb{E}_{S_t}[U_t]$. Besides, note that, by the assumption on the boundedness of ℓ, it a.s. holds that $|U_t| \leq C$. Applying Lemma 15 to U_t gives, for any $\beta > 0$,

$$\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\left[e^{\beta\sum_{t=1}^{T}(U_t - \mathbb{E}[U_t])}\right] \leq \mathrm{e}^{\frac{\beta^2 T\tilde{V}(\pi)}{2(1-\beta C)}},$$

and

$$\tilde{V}(\pi) = \mathbb{E}_{P_{T+1},N_{T+1}}\mathbb{E}_{S_{T+1}}\left[\left(\hat{B}_{T+1}(\rho_{T+1}(\pi,\alpha),\pi,\alpha) - \hat{B}_{T+1}(\rho_{T+1}(\pi_\alpha^*,\alpha),\pi_\alpha^*,\alpha)\right)^2\right].$$

This factor can be bounded as $\tilde{V}(\pi) \leq 8eC\mathbb{E}[-U_{T+1}]$ by Theorem 3, which states that Bernstein's condition is satisfied at the meta level, so that the bound becomes

$$\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\left[\mathrm{e}^{\beta\sum_{t=1}^{T}(U_t-\mathbb{E}[U_t]) + \frac{8eC\beta^2 T\mathbb{E}[U_{T+1}]}{2(1-\beta C)}}\right] \leq 1.$$

Integrating with respect to the prior $\pi \sim \Lambda$ and using Fubini's theorem yields

$$\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\Lambda}\left[\mathrm{e}^{\beta\sum_{t=1}^{T}(U_t-\mathbb{E}[U_t])+\frac{8eC\beta^2T\mathbb{E}[U_{T+1}]}{2(1-\beta C)}}\right]\leq 1.$$

Next, by an application of Lemma 16, the left-hand side becomes

$$\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\left[\mathrm{e}^{\sup_{\Pi\in\mathcal{P}(\mathcal{P}(\theta))}\left\{\mathbb{E}_{\pi\sim\Pi}\left[\beta\sum_{t=1}^{T}(U_t-\mathbb{E}[U_t])+\frac{8eC\beta^2T\mathbb{E}[U_{T+1}]}{2(1-\beta C)}\right]-\mathrm{KL}(\Pi\|\Lambda)\right\}}\right]\leq 1.$$

We then make use of Jensen's inequality and arrange terms, so that the bound becomes

$$\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\left[\sup_{\Pi\in\mathcal{P}(\mathcal{P}(\theta))}\left\{\mathbb{E}_{\pi\sim\Pi}\left[\frac{1}{T}\sum_{t=1}^{T}(U_t-\mathbb{E}[U_t])+\frac{8eC\beta\mathbb{E}\left[U_{T+1}\right]}{2(1-\beta C)}\right]-\frac{\mathrm{KL}(\Pi\|\Lambda)}{\beta T}\right\}\right]\leq 0.$$

We replace the supremum on Π by an evaluation of the term in $\hat{\Pi}$ and arrange terms, so that the bound becomes

$$\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}\left[-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[U_t]+\frac{8eC\beta\mathbb{E}\left[U_{T+1}\right]}{2(1-\beta C)}\right]$$
$$\leq\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\left[-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi\sim\hat{\Pi}}[U_t]+\frac{\mathrm{KL}(\hat{\Pi}\|\Lambda)}{\beta T}\right],$$

which is identical to

$$\left(-1+\frac{8eC\beta}{2(1-C\beta)}\right)\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}\mathbb{E}_{P_{T+1},N_{T+1}}\mathbb{E}_{S_{T+1}}\left[U_{T+1,N_{T+1}}\right]$$
$$\leq\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\left[-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi\sim\hat{\Pi}}[U_t]+\frac{\mathrm{KL}(\hat{\Pi}\|\Lambda)}{\beta T}\right].$$

Then, we replace U_t by its value for $t\in\{1,\dots,T+1\}$, yielding the bound

$$\left(1-\frac{8eC\beta}{2(1-C\beta)}\right)\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}\mathbb{E}_{P_{T+1},N_{T+1}}\mathbb{E}_{S_{T+1}}\Bigg[$$
$$\hat{B}_{T+1}(\rho_{T+1}(\pi,\alpha),\pi,\alpha)-\hat{B}_{T+1}(\rho_{T+1}(\pi^*_\alpha,\alpha),\pi^*_\alpha,\alpha)\Bigg]$$
$$\leq\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi\sim\hat{\Pi}}\left[\hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha)-\hat{B}_t(\rho_t(\pi^*_\alpha,\alpha),\pi^*_\alpha,\alpha)\right]+\frac{\mathrm{KL}(\hat{\Pi}\|\Lambda)}{\beta T}\right].$$

Since the term $\hat{B}_t(\rho_t(\pi^*_\alpha, \alpha), \pi^*_\alpha, \alpha)$ does not depend on $\pi \sim \hat{\Pi}$, we can simplify it on both sides of the inequality, which gives

$$\left(1 - \frac{8eC\beta}{2(1-C\beta)}\right) \mathbb{E}_{P_{1:T},N_{1:T}} \mathbb{E}_{S_{1:T}} \mathbb{E}_{\pi\sim\hat{\Pi}} \mathbb{E}_{P_{T+1},N_{T+1}} \mathbb{E}_{S_{T+1}} \left[\hat{B}_{T+1}(\rho_{T+1}(\pi,\alpha),\pi,\alpha)\right] + \frac{8eC\beta}{2(1-C\beta)} \mathbb{E}_{P_{1:T},N_{1:T}} \mathbb{E}_{S_{1:T}} \left[\hat{B}_{T+1}(\rho_{T+1}(\pi^*_\alpha,\alpha),\pi^*_\alpha,\alpha)\right] \leq \mathbb{E}_{P_{1:T},N_{1:T}} \mathbb{E}_{S_{1:T}} \left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\pi\sim\hat{\Pi}} \left[\hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha)\right] + \frac{\mathrm{KL}(\hat{\Pi}\|\Lambda)}{\beta T}\right]. \quad (25)$$

Theorem 1 provides the following lower bound for any π', and any (P_{T+1}, N_{T+1}):

$$\mathbb{E}_{S_{T+1}} \left[\hat{B}_{T+1}(\rho_{T+1}(\pi',\alpha),\pi',\alpha)\right] \geq \left(1 - \frac{\alpha c \mathbb{I}_B}{2(1-C\alpha)}\right) \mathbb{E}_{S_{T+1}} \mathbb{E}_{\theta\sim\rho_{T+1}(\pi',\alpha)} \left[R_{P_{T+1}}(\theta)\right] + \frac{\alpha c \mathbb{I}_B}{2(1-C\alpha)} R^*_{P_{T+1}} - \frac{\alpha C^2(1-\mathbb{I}_B)}{8}. \quad (26)$$

This further implies that

$$\mathbb{E}_{S_{T+1}} \left[\hat{B}_{T+1}(\rho_{T+1}(\pi',\alpha),\pi',\alpha)\right] \geq \left(1 - \frac{\alpha c \mathbb{I}_B}{2(1-C\alpha)}\right) R^*_{P_{T+1}} + \frac{\alpha c \mathbb{I}_B}{2(1-C\alpha)} R^*_{P_{T+1}} - \frac{\alpha C^2(1-\mathbb{I}_B)}{8} \quad (27)$$

for any π', and (P_{T+1}, N_{T+1}). In particular, applying (26) to $\pi' = \pi$ and (27) to $\pi = \pi^*_\alpha$, and injecting the results in the left-hand side of (25) gives

$$\left(1 - \frac{\alpha c \mathbb{I}_B}{2(1-C\alpha)}\right) \left(\frac{8eC\beta}{2(1-C\beta)} \mathbb{E}_{P_{T+1},N_{T+1}} \left[R^*_{P_{T+1}}\right] + \left(1 - \frac{8eC\beta}{2(1-C\beta)}\right) \mathbb{E}_{P_{1:T},N_{1:T}} \mathbb{E}_{S_{1:T}} \mathbb{E}_{\pi\sim\hat{\Pi}} \mathbb{E}_{P_{T+1},N_{T+1}} \mathbb{E}_{S_{T+1}} \mathbb{E}_{\theta\sim\rho_{T+1}(\pi',\alpha)} \left[R_{P_{T+1}}(\theta)\right]\right) + \frac{\alpha c \mathbb{I}_B}{2(1-C\alpha)} \mathbb{E}_{P_{T+1},N_{T+1}} \left[R^*_{P_{T+1}}\right] - \frac{\alpha C^2(1-\mathbb{I}_B)}{8} \leq \mathbb{E}_{P_{1:T},N_{1:T}} \mathbb{E}_{S_{1:T}} \left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\pi\sim\hat{\Pi}} \left[\hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha)\right] + \frac{\mathrm{KL}(\hat{\Pi}\|\Lambda)}{\beta T}\right].$$

We remove $\mathcal{E}^* = \mathbb{E}_{P_{T+1},N_{T+1}} \left[R^*_{P_{T+1}}\right]$ from both sides of the inequality and arrange terms, so that the bound becomes

$$\left(1 - \frac{\alpha c \mathbb{I}_B}{2(1-C\alpha)}\right) \left(1 - \frac{8eC\beta}{2(1-C\beta)}\right) \left(\mathbb{E}_{P_{1:T},N_{1:T}} \mathbb{E}_{S_{1:T}} \mathbb{E}_{\pi\sim\hat{\Pi}} \left[\mathcal{E}(\pi)\right] - \mathcal{E}^*\right) - \frac{\alpha C^2(1-\mathbb{I}_B)}{8}$$

$$\leq \mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi\sim\hat{\Pi}}\left[\hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha)\right]-\mathcal{E}^*+\frac{\mathrm{KL}(\hat{\Pi}\|\Lambda)}{\beta T}\right].$$

By definition, $\hat{\Pi}$ is the minimizer of the integrand of the right-hand side, and therefore,

$$\begin{aligned}\left(1-\frac{\alpha c\mathbb{I}_B}{2(1-C\alpha)}\right)\left(1-\frac{8eC\beta}{2(1-C\beta)}\right)&\left(\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}\left[\mathcal{E}(\pi)\right]-\mathcal{E}^*\right)\\ &\leq \mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\left[\inf_{\Pi\in\mathcal{G}}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}_{\pi\sim\Pi}\left[\hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha)\right]-\hat{R}_t(\theta_t^*)\right)+\frac{\mathrm{KL}(\Pi\|\Lambda)}{\beta T}\right\}\right]\\ &\qquad+\frac{\alpha C^2(1-\mathbb{I}_B)}{8},\end{aligned}$$

and the choice $\beta=\frac{1}{(1+8e)C}$ yields

$$\begin{aligned}\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}\left[\mathcal{E}(\pi)\right]-\mathcal{E}^*\leq\frac{2}{1-\frac{\alpha c\mathbb{I}_B}{2(1-C\alpha)}}&\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\Bigg[\\ &\inf_{\Pi\in\mathcal{G}}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}_{\pi\sim\Pi}\left[\hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha)\right]-\hat{R}_t(\theta_t^*)\right)+\frac{(1+8e)C\mathrm{KL}(\Pi\|\Lambda)}{T}\right\}\\ &\qquad+\frac{\alpha C^2(1-\mathbb{I}_B)}{8}\Bigg],\end{aligned}$$

which concludes the proof of the lemma. ■

F.2 Proof of Theorem 5

From Lemma 17,

$$\begin{aligned}&\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}\left[\mathcal{E}(\pi)\right]-\mathcal{E}^*\\ &\leq\frac{2}{1-\frac{\alpha c\mathbb{I}_B}{2(1-C\alpha)}}\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\left[\inf_{\Pi\in\mathcal{G}}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}_{\pi\sim\Pi}\left[\hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha)\right]-\hat{R}_t(\theta_t^*)\right)\right.\right.\\ &\qquad\left.\left.+\frac{(1+8e)C\mathrm{KL}(\Pi\|\Lambda)}{T}\right\}+\frac{\alpha C^2(1-\mathbb{I}_B)}{8}\right]\\ &\leq\frac{2}{1-\frac{\alpha c\mathbb{I}_B}{2(1-C\alpha)}}\inf_{\Pi\in\mathcal{G}}\mathbb{E}_{P_{1:T},N_{1:T}}\mathbb{E}_{S_{1:T}}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}_{\pi\sim\Pi}\left[\hat{B}_t(\rho_t(\pi,\alpha),\pi,\alpha)\right]-\hat{R}_t(\theta_t^*)\right)\right.\\ &\qquad\left.+\frac{(1+8e)C\mathrm{KL}(\Pi\|\Lambda)}{T}+\frac{\alpha C^2(1-\mathbb{I}_B)}{8}\right\}\\ &=\frac{2}{1-\frac{\alpha c\mathbb{I}_B}{2(1-C\alpha)}}\inf_{\Pi\in\mathcal{G}}\mathbb{E}_{P_{T+1},N_{T+1}}\mathbb{E}_{S_{T+1}}\left\{\left(\mathbb{E}_{\pi\sim\Pi}\left[\hat{B}_{T+1}(\rho_{T+1}(\pi,\alpha),\pi,\alpha)\right]-\hat{R}_{T+1}(\theta_{T+1}^*)\right)\right.\end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{(1+8e)C\mathrm{KL}(\Pi\|\Lambda)}{T} + \frac{\alpha C^2(1-\mathbb{I}_B)}{8}\Bigg\} \\
&\leq \frac{2}{1-\frac{\alpha c\mathbb{I}_B}{2(1-C\alpha)}} \inf_{\Pi\in\mathcal{G}} \mathbb{E}_{P_{T+1},N_{T+1}} \mathbb{E}_{S_{T+1}} \Bigg\{ \mathbb{E}_{\pi\sim\Pi} \inf_{\rho\in\mathcal{P}(\Theta)} \Bigg[\mathbb{E}_{\theta\sim\rho}[\hat{R}_{T+1}(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\alpha N_{T+1}} \\
&\qquad - \hat{R}_{T+1}(\theta^*_{T+1}) \Bigg] + \frac{(1+8e)C\mathrm{KL}(\Pi\|\Lambda)}{T} + \frac{\alpha C^2(1-\mathbb{I}_B)}{8} \Bigg\} \\
&\leq \frac{2}{1-\frac{\alpha c\mathbb{I}_B}{2(1-C\alpha)}} \inf_{\Pi\in\mathcal{G}} \mathbb{E}_{P_{T+1},N_{T+1}} \Bigg\{ \mathbb{E}_{\pi\sim\Pi} \inf_{\rho\in\mathcal{P}(\Theta)} \mathbb{E}_{S_{T+1}} \Bigg[\mathbb{E}_{\theta\sim\rho}[\hat{R}_{T+1}(\theta)] + \frac{\mathrm{KL}(\rho\|\pi)}{\alpha N_{T+1}} \\
&\qquad - \hat{R}_{T+1}(\theta^*_{T+1}) \Bigg] + \frac{(1+8e)C\mathrm{KL}(\Pi\|\Lambda)}{T} + \frac{\alpha C^2(1-\mathbb{I}_B)}{8} \Bigg\} \\
&\leq \frac{2}{1-\frac{\alpha c\mathbb{I}_B}{2(1-C\alpha)}} \inf_{\Pi\in\mathcal{G}} \mathbb{E}_{P_{T+1},N_{T+1}} \Bigg\{ \mathbb{E}_{\pi\sim\Pi} \inf_{\rho\in\mathcal{P}(\Theta)} \Bigg[\mathbb{E}_{\theta\sim\rho}[R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} + \frac{\mathrm{KL}(\rho\|\pi)}{\alpha N_{T+1}} \Bigg] \\
&\qquad + \frac{(1+8e)C\mathrm{KL}(\Pi\|\Lambda)}{T} + \frac{\alpha C^2(1-\mathbb{I}_B)}{8} \Bigg\}.
\end{aligned}
$$

This ends the proof. ■

Appendix G. Application of Theorem 5 to the Gaussian Case

In this section and until the end of this paper, we assume that Bernstein's condition (Assumption 1), (3) (the loss is bounded) and (10) are satisfied. For the sake of clarity, we recall (10): for any P_t and any $\theta \in \Theta$,

$$R_{P_t}(\theta) - R^*_{P_t} \leq L\|\theta - \theta^*_t\|^2.$$

In this section and in the next one, we set $\alpha = \frac{1}{c+C}$ and $\beta = \frac{1}{(1+8e)C}$ in order to compactify the equations.

In this section, we assume that priors follow Gaussian distributions, and we consider the set of all Gaussian distributions

$$\mathcal{M} = \left\{ p_{\mu,\sigma^2} = \bigotimes_{i=1}^{d} \mathcal{N}(\mu_i, \sigma_i^2) : \forall i \in [d], \mu_i \in \mathbb{R}, \sigma_i^2 \in \mathbb{R}_+^* \right\}.$$

We choose the prior $p_{\bar{\mu},(\bar{\sigma}^2,\dots,\bar{\sigma}^2)} \in \mathcal{F}$ (and hence, the variances are the same for all coordinates). A straightforward application of (19) gives

$$\mathrm{KL}(p_{\mu,\sigma^2}|p_{\bar{\mu},(\bar{\sigma}^2,\dots,\bar{\sigma}^2)}) = \frac{1}{2}\sum_{i=1}^{d}\left[\frac{(\mu_i-\bar{\mu}_i)^2}{\bar{\sigma}^2} + \frac{\sigma_i^2}{\bar{\sigma}^2} - 1 + \log\left(\frac{\bar{\sigma}^2}{\sigma_i^2}\right)\right].$$

G.1 Bound in Isolation

We start from the bound in isolation from Theorem 1 at $t = T+1$:

$$\mathbb{E}_{S_{T+1}}\mathbb{E}_{\theta\sim\rho_{T+1}(\pi,\alpha)}[R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} \leq 2\inf_{\mu,\sigma^2}\left\{\mathbb{E}_{\theta\sim p_{\mu,\sigma^2}}[R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} + \frac{1}{2\alpha n}\sum_{i=1}^{d}\left(\frac{(\mu_i-\bar{\mu}_i)^2}{\bar{\sigma}^2} + \frac{\sigma_i^2}{\bar{\sigma}^2} - 1 + \log\frac{\bar{\sigma}^2}{\sigma_i^2}\right)\right\}.$$

Using the assumption made in (10), the bound becomes

$$\mathbb{E}_{S_{T+1}}\mathbb{E}_{\theta\sim\rho_{T+1}(\pi,\alpha)}[R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} \leq 2\inf_{\mu,\sigma^2}\left\{L\mathbb{E}_{\theta\sim p_{\mu,\sigma^2}}\left[\|\theta-\theta^*_{T+1}\|^2\right] + \frac{\|\mu-\bar{\mu}\|^2}{2\alpha n} + \frac{1}{2\alpha n}\sum_{i=1}^{d}\left(\frac{\sigma_i^2}{\bar{\sigma}^2} - 1 + \log\frac{\bar{\sigma}^2}{\sigma_i^2}\right)\right\}.$$

We can then perform an exact optimization on σ^2 which yields $\sigma_i^2 = \frac{\bar{\sigma}^2}{1+2\alpha Ln\bar{\sigma}^2}$, and after simplifications, the bound becomes

$$\mathbb{E}_{S_{T+1}}\mathbb{E}_{\theta\sim\rho_{T+1}(\pi,\alpha)}[R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} \leq 2\inf_{\mu,\sigma^2}\left\{L\mathbb{E}_{\theta\sim p_{\mu,\sigma^2}}\left[\|\theta-\theta^*_{T+1}\|^2\right] + \frac{\|\mu-\bar{\mu}\|^2}{2\alpha n} + \frac{d}{2\alpha n}\log(1+2\alpha Ln\bar{\sigma}^2)\right\}.$$

We can easily compute the expectation

$$\mathbb{E}_{\theta\sim p_{\mu,\sigma^2}}\left[\|\theta-\theta^*_{T+1}\|^2\right] = \|\mu-\theta^*_{T+1}\|^2 + \|\sigma\|^2,$$

and then perform an exact optimization in μ, which gives $\mu_i = \frac{2L\theta^*_{T+1,i} + \frac{1}{\alpha n\bar{\sigma}^2}\bar{\mu}_i}{2L + \frac{1}{\alpha n\bar{\sigma}^2}}$, and replacing in the bound yields, after simplifications,

$$\mathbb{E}_{S_{T+1}}\mathbb{E}_{\theta\sim\rho_{T+1}(\pi,\alpha)}[R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} \leq \frac{2L}{2L\alpha n\bar{\sigma}^2+1}\|\bar{\mu}-\theta^*_{T+1}\|^2 + \frac{d}{\alpha n}\log(1+2\alpha Ln\bar{\sigma}^2). \quad (28)$$

The objective is to see if we can achieve a better bound than the above with the meta-learning.

G.2 Bound for the Meta-Learning

For the meta-learning, we need to define our set of priors on the priors $\mathcal{G}$, which we choose as the family of distributions $q_{\tau,\xi^2,b}$ on $(\bar{\mu},\bar{\sigma}^2)$, where

$$q_{\tau,\xi^2,b}(\bar{\mu},\bar{\sigma}^2) = \left[\bigotimes_{i=1}^{d}\mathcal{N}(\bar{\mu}_i;\tau_i,\xi_i^2)\right]\otimes\Gamma(\bar{\sigma}^2;2,b).$$

Fix a prior on priors $\Lambda = q_{0,\bar{\xi}^2,\bar{b}}$. We choose $\hat{\Pi} = q_{\hat{\tau},\hat{\xi}^2,\hat{b}}$, where

$$(\hat{\tau}, \hat{\xi}^2, \hat{b}) = \operatorname{argmin}_{\tau,\xi^2,b}\Bigg\{\mathbb{E}_{(\bar{\mu},\bar{\sigma}^2)\sim q_{\tau,\xi^2,b}}\Bigg[\frac{1}{T}\sum_{t=1}^{T}\min_{\mu(t),\sigma^2(t)}\Bigg\{\mathbb{E}_{\theta\sim\mathcal{N}(\mu(t),\sigma^2(t))}\left[\hat{R}_t(\theta)\right] + \frac{1}{2\alpha n}\sum_{i=1}^{d}\left[\frac{(\mu_i(t)-\bar{\mu}_i)^2}{\bar{\sigma}^2} + \frac{\sigma_i^2(t)}{\bar{\sigma}^2} - 1 + \log\frac{\bar{\sigma}^2}{\sigma_i^2(t)}\right]\Bigg\}\Bigg] + \frac{1}{2\beta T}\sum_{i=1}^{d}\left[\frac{\tau_i^2}{\bar{\xi}^2} + \frac{\xi_i^2}{\bar{\xi}^2} - 1 + \log\frac{\bar{\xi}^2}{\xi_i^2}\right] + \frac{\log\frac{b}{\bar{b}} + \frac{\bar{\bar{b}}-b}{b}}{\beta T}\Bigg\},$$

where $\Gamma(\cdot)$ is the gamma function and $\psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is the digamma function. By an application of Theorem 5,

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4\inf_{\tau,\xi^2,b}\Bigg\{\mathbb{E}_{(\bar{\mu},\bar{\sigma}^2)\sim q_{\tau,\xi^2,b}}\mathbb{E}_{P_{T+1}}\Bigg[\min_{\mu,\sigma^2}\Bigg\{\mathbb{E}_{\theta\sim p_{\mu,\sigma^2}}[R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} + \frac{1}{2\alpha n}\sum_{i=1}^{d}\left[\frac{(\mu_i-\bar{\mu}_i)^2}{\bar{\sigma}^2} + \frac{\sigma_i^2}{\bar{\sigma}^2} - 1 + \log\frac{\bar{\sigma}^2}{\sigma_i^2}\right]\Bigg\}\Bigg] + \frac{1}{2\beta T}\sum_{i=1}^{d}\left[\frac{\tau_i^2}{\bar{\xi}^2} + \frac{\xi_i^2}{\bar{\xi}^2} - 1 + \log\frac{\bar{\xi}^2}{\xi_i^2}\right] + \frac{\log\frac{b}{\bar{b}} + \frac{\bar{\bar{b}}-b}{b}}{\beta T}\Bigg\}.$$

The assumption made in (10) implies that

$$\mathbb{E}_{\theta\sim p_{\mu,\sigma^2}}[R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} \leq L\mathbb{E}_{\theta\sim p_{\mu,\sigma^2}}\left[\|\theta - \theta^*_{T+1}\|^2\right].$$

With the choice $\mu_1 = \cdots = \mu_K = \theta^*_{T+1}$, the previous bound becomes

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4\inf_{\tau,\xi^2,b}\Bigg\{\mathbb{E}_{(\bar{\mu},\bar{\sigma}^2)\sim q_{\tau,\xi^2,b}}\mathbb{E}_{P_{T+1}}\Bigg[\min_{\mu,\sigma^2}\Bigg\{L\|\sigma\|^2 + \frac{1}{2\alpha n}\sum_{i=1}^{d}\left[\frac{(\mu_i-\bar{\mu}_i)^2}{\bar{\sigma}^2} + \frac{\sigma_i^2}{\bar{\sigma}^2} - 1 + \log\frac{\bar{\sigma}^2}{\sigma_i^2}\right]\Bigg\}\Bigg] + \frac{1}{2\beta T}\sum_{i=1}^{d}\left[\frac{\tau_i^2}{\bar{\xi}^2} + \frac{\xi_i^2}{\bar{\xi}^2} - 1 + \log\frac{\bar{\xi}^2}{\xi_i^2}\right] + \frac{\log\frac{b}{\bar{b}} + \frac{\bar{\bar{b}}-b}{b}}{\beta T}\Bigg\}.$$

An exact optimization in σ^2 gives $\sigma_i^2 = \frac{\bar{\sigma}^2}{2\alpha L\bar{\sigma}^2 n+1}$, and after simplifications,

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4\inf_{\tau,\xi^2,b}\Bigg\{\mathbb{E}_{(\bar{\mu},\bar{\sigma}^2)\sim q_{\tau,\xi^2,b}}\mathbb{E}_{P_{T+1}}\Bigg[\min_{\mu}\left\{\frac{d}{2\alpha n}\log\left(2\alpha n L\bar{\sigma}^2 + 1\right) + \frac{1}{2\alpha n}\frac{\|\mu-\bar{\mu}\|^2}{\bar{\sigma}^2}\right\}\Bigg]$$

$$+ \frac{1}{2\beta T}\sum_{i=1}^{d}\left[\frac{\tau_i^2}{\bar{\xi}^2} + \frac{\xi_i^2}{\bar{\xi}^2} - 1 + \log\frac{\bar{\xi}^2}{\xi_i^2}\right] + \frac{\log\frac{b}{\bar{b}} + \frac{\bar{b}-b}{b}}{\beta T}\Bigg\}.$$

At this stage, we choose $\mu = \theta^*_{T+1}$ which, by definition, satisfies $R_{P_{T+1}}\left(\theta^*_{T+1}\right) = R^*_{P_{T+1}}$, and choose $\tau = \mathbb{E}_{P_{T+1}}[\theta^*_{T+1}]$. If $\theta^*_{T+1} = \tau$ $\mathcal{P}$-a.s., all the task have the same solution. On the other hand, if θ^*_{T+1} has a lot of variations, then the tasks have very unrelated solutions. Replacing in the infimum above yields

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4\inf_{\xi^2,b}\Bigg\{ \mathbb{E}_{(\bar{\mu},\bar{\sigma}^2)\sim q_{\tau,\xi^2,b}}\left[\frac{d}{2\alpha n}\log\left(2\alpha n L\bar{\sigma}^2 + 1\right) + \frac{1}{2\alpha n}\frac{\mathbb{E}_{P_{T+1}}\left[\|\theta^*_{T+1} - \bar{\mu}\|^2\right]}{\bar{\sigma}^2}\right]$$
$$+ \frac{1}{2\beta T}\sum_{i=1}^{d}\left[\frac{\tau_i^2}{\bar{\xi}^2} + \frac{\xi_i^2}{\bar{\xi}^2} - 1 + \log\frac{\bar{\xi}^2}{\xi_i^2}\right] + \frac{\log\frac{b}{\bar{b}} + \frac{\bar{b}-b}{b}}{\beta T}\Bigg\}.$$

Let $\Sigma(\mathcal{P}) = \mathbb{E}_{P_{T+1}}\left[\|\theta^*_{T+1} - \mu^*\|^2\right]$, this quantity will be very important in the rate. Using Fubini's theorem to invert the expectation w.r.t. $\bar{\mu}$ and the expectation w.r.t. P_{T+1} yields

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4\inf_{\xi^2,b}\Bigg\{ \mathbb{E}_{\bar{\sigma}^2\sim\Gamma(2,b)}\left[\frac{d}{2\alpha n}\log\left(2\alpha n L\bar{\sigma}^2 + 1\right) + \frac{\Sigma(\mathcal{P}) + \|\xi\|^2}{2\alpha n\bar{\sigma}^2}\right]$$
$$+ \frac{1}{2\beta T}\sum_{i=1}^{d}\left[\frac{\tau_i^2}{\bar{\xi}^2} + \frac{\xi_i^2}{\bar{\xi}^2} - 1 + \log\frac{\bar{\xi}^2}{\xi_i^2}\right] + \frac{\log\frac{b}{\bar{b}} + \frac{\bar{b}-b}{b}}{\beta T}\Bigg\}.$$

We can then bound the expectation w.r.t. $\bar{\sigma}^2$, as, by Jensen's inequality,

$$\mathbb{E}_{\bar{\sigma}^2\sim\Gamma(2,b)}\left[\frac{d}{2\alpha n}\log\left(2\alpha n L\bar{\sigma}^2 + 1\right)\right] \leq \frac{d}{2\alpha n}\log\left(2\alpha n L\mathbb{E}[\bar{\sigma}^2] + 1\right)$$
$$= \frac{d}{2\alpha n}\log\left(\frac{4\alpha n L}{b} + 1\right),$$

and

$$\mathbb{E}_{\bar{\sigma}^2\sim\Gamma(2,b)}\left[\frac{\Sigma(\mathcal{P}) + \|\xi\|^2}{2\alpha n\bar{\sigma}^2}\right] = \frac{(\Sigma(\mathcal{P}) + \|\xi\|^2)b}{2\alpha n}.$$

We can then replace in the computation:

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4\inf_{\xi^2,b}\Bigg\{ \frac{d}{2\alpha n}\log\left(\frac{4\alpha n L}{b} + 1\right) + \frac{(\Sigma(\mathcal{P}) + \|\xi\|^2)b}{2\alpha n}$$
$$+ \frac{1}{2\beta T}\sum_{i=1}^{d}\left[\frac{\tau_i^2}{\bar{\xi}^2} + \frac{\xi_i^2}{\bar{\xi}^2} - 1 + \log\frac{\bar{\xi}^2}{\xi_i^2}\right] + \frac{\log\frac{b}{\bar{b}} + \frac{\bar{b}-b}{b}}{\beta T}\Bigg\}.$$

An exact optimization in ξ^2 yields $\xi_i^2 = \frac{\bar{\bar{\xi}}^2}{1+\frac{db\bar{\bar{\xi}}^2\beta T}{\alpha n}}$, and replacing ξ^2 in the previous bound gives

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4\inf_b \left\{ \frac{d}{2\alpha n}\log\left(\frac{4\alpha nL}{b}+1\right) + \frac{b\Sigma(\mathcal{P})}{2\alpha n} \right.$$
$$\left. + \frac{\|\tau\|^2}{2\beta\bar{\bar{\xi}}^2 T} + \frac{d}{2\beta T}\log\left(\frac{db\bar{\bar{\xi}}^2\beta T}{\alpha n}\right) + \frac{\log\frac{b}{\bar{\bar{b}}} + \frac{\bar{\bar{b}}-b}{b}}{\beta T} \right\}.$$

t We then restrict the above infimum to the values of b in $[1, T]$, and by noting that, for any $1 \leq b \leq T$,

$$\frac{d}{2\beta T}\log\left(\frac{db\bar{\bar{\xi}}^2\beta T}{\alpha n}\right) \leq \frac{d}{2\beta T}\log\left(\frac{d\bar{\bar{\xi}}^2\beta T^2}{\alpha n}\right)$$

and

$$\frac{\log\frac{b}{\bar{\bar{b}}} + \frac{\bar{\bar{b}}-b}{b}}{\beta T} \leq \frac{\log\frac{T}{\bar{\bar{b}}} + \bar{\bar{b}} - 1}{\beta T},$$

we can replace those terms by their respective bounds in the above computation and extract them from the infimum in $1 \leq b \leq T$, and this yields

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq \inf_{1\leq b\leq T}\left\{ \frac{2d}{\alpha n}\log\left(\frac{4\alpha nL}{b}+1\right) + \frac{2b\Sigma(\mathcal{P})}{\alpha n} \right\}$$
$$+ \frac{2\|\tau\|^2}{\beta\bar{\bar{\xi}}^2 T} + \frac{2d}{\beta T}\log\left(\frac{d\bar{\bar{\xi}}^2\beta T^2}{\alpha n}\right) + 4\frac{\log\frac{T}{\bar{\bar{b}}} + \bar{\bar{b}} - 1}{\beta T}.$$

In the specific regime $\Sigma(\mathcal{P}) \leq \frac{dn}{T^2}$ (d is here for dimension reasons), then the Gaussian is very concentrated around its mean, and its variance is smaller than $\frac{n}{T^2}$ on each axis. This implies that the optimal parameter of the new task $T+1$ is going to close to τ, and we can benefit from the previous tasks $t = 1, \ldots, T$ to infer it. Hence, we expect a significant improvement over the learning in isolation in this regime.

For our bound, this means that in the infimum in b, the term $\frac{2b\Sigma(\mathcal{P})}{\alpha n}$ is very small, and we will choose b large so that it minimizes the first term of the sum $\frac{2d}{\alpha n}\log\left(\frac{4\alpha nL}{b}\right)$, and we choose $b = T$. We then bound the infimum on b by

$$\frac{8dL}{T} + \frac{2d}{\alpha T},$$

where we used the majoration $\log(1+x) \leq x$. This yields the bound

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq \frac{8dL}{T} + \frac{2d}{\alpha T} + \frac{2\|\tau\|^2}{\beta\bar{\bar{\xi}}^2 T} + \frac{2d}{\beta T}\log\left(\frac{d\bar{\bar{\xi}}^2\beta T^2}{\alpha n}\right) + 4\frac{\log\frac{T}{\bar{\bar{b}}} + \bar{\bar{b}} - 1}{\beta T}$$

in the advantageous regime $\Sigma(\mathcal{P}) \leq \frac{dn}{T^2}$.

Appendix H. Application of Theorem 5 to the Case of Mixtures of Gaussians

Similarly as in the previous section, in this section, we assume that Bernstein's condition (Assumption 1), (3) (the loss is bounded) and (10) are satisfied. For the sake of clarity, we recall (10): for any P_t and any $\theta \in \Theta$,

$$R_{P_t}(\theta) - R^*_{P_t} \leq L\|\theta - \theta^*_t\|^2.$$

Recall that we set $\alpha = \frac{1}{c+C}$ and $\beta = \frac{1}{(1+8e)C}$ in order to compactify the equations.

H.1 Case where the Number of Mixtures is Known

We first assume that priors that are mixtures of K Gaussians, where K is known:

$$\mathcal{M} = \Bigg\{ p_{w,\mu,\sigma^2} = \sum_{k=1}^{K} w_k \bigotimes_{i=1}^{d} \mathcal{N}(\mu_{k,i}, \sigma^2_{k,i}) : \\ \forall (i,k) \in [d] \times [K], \mu_{k,i} \in \mathbb{R}, \sigma^2_{k,i} \in \mathbb{R}^*_+, w_k \geq 0, 1^\top w = 1 \Bigg\}.$$

We set the prior $\pi = \sum_{k=1}^{K} \bar{w}_k \mathcal{N}(\bar{\mu}_k, \bar{\sigma}^2_k I_d)$. Then, denoting by $g(x; \mu, \sigma^2)$ the pdf of the normal distribution $\mathcal{N}(\mu, \sigma^2)$, (19) implies, for any w, μ, σ^2,

$$\begin{aligned}
\mathrm{KL}(p_{w,\mu,\sigma^2} \| \pi) &= \int_{\mathbb{R}^d} \log \frac{\sum_{k=1}^{K} w_k g(x; \mu_k, \sigma^2_k)}{\sum_{k=1}^{K} \bar{w}_k g(x; \bar{\mu}_k, \bar{\sigma}^2_k I_d)} \sum_{k=1}^{K} w_k g(x; \mu_k, \sigma^2_k) dx \\
&\leq \int_{\mathbb{R}^d} \sum_{k=1}^{K} \log \frac{w_k g(x; \mu_k, \sigma^2_k)}{\bar{w}_k g(x; \bar{\mu}_k, \bar{\sigma}^2_k I_d)} w_k g(x; \mu_k, \sigma^2_k) dx \\
&= \sum_{k=1}^{K} w_k \log \frac{w_k}{\bar{w}_k} + \sum_{k=1}^{K} w_k \mathrm{KL}(\mathcal{N}(\mu_k, \sigma^2_k) \| \mathcal{N}(\bar{\mu}_k, \bar{\sigma}^2_k I_d)) \\
&= \mathrm{KL}(w \| \bar{w}) + \frac{1}{2} \sum_{k=1}^{K} w_k \sum_{i=1}^{d} \left(\frac{(\mu_{k,i} - \bar{\mu}_{k,i})^2}{\bar{\sigma}^2_k} + \frac{\sigma^2_{k,i}}{\bar{\sigma}^2_k} - 1 + \log \frac{\bar{\sigma}^2_k}{\sigma^2_{k,i}} \right),
\end{aligned}$$

where the inequality on the second line follows from the log sum inequality from Cover and Thomas (2006), and the bound from Theorem 5 becomes, at $t = T + 1$,

$$\begin{aligned}
\mathbb{E}_{S_{T+1}} \mathbb{E}_{\theta \sim \rho_{T+1}(\pi,\alpha)} [R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} \leq 2 \inf_{w,\mu,\sigma^2} \Bigg\{ \mathbb{E}_{\theta \sim p_{w,\mu,\sigma^2}} [R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} + \frac{\mathrm{KL}(w \| \bar{w})}{\alpha n} \\
+ \frac{1}{2\alpha n} \sum_{k=1}^{K} w_k \sum_{i=1}^{d} \left(\frac{(\mu_{k,i} - \bar{\mu}_{k,i})^2}{\bar{\sigma}^2_k} + \frac{\sigma^2_{k,i}}{\bar{\sigma}^2_k} - 1 + \log \frac{\bar{\sigma}^2_k}{\sigma^2_{k,i}} \right) \Bigg\}.
\end{aligned}$$

The assumption made in (10) implies that

$$\mathbb{E}_{\theta \sim \mathcal{N}(\mu,\sigma^2)} [R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} \leq L \mathbb{E}_{\theta \sim \mathcal{N}(\mu,\sigma^2)} \left[\|\theta - \theta^*_{T+1}\|^2 \right].$$

It follows that the previous bound with the choice $\mu_1 = \cdots = \mu_K = \theta^*_{T+1}$ becomes

$$\mathbb{E}_{S_{T+1}}\mathbb{E}_{\theta\sim\rho_{T+1}(\pi,\alpha)}[R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} \leq 2\inf_{w,\sigma^2}\left\{ L\sum_{k=1}^K w_k\|\sigma_k\|^2 + \frac{\mathrm{KL}(w\|\bar{w})}{\alpha n} \right.$$
$$\left. + \frac{1}{2\alpha n}\sum_{k=1}^K w_k \sum_{i=1}^d \left(\frac{(\theta^*_{T+1,i}-\bar{\mu}_{k,i})^2}{\bar{\sigma}_k^2} + \frac{\sigma_{k,i}^2}{\bar{\sigma}_k^2} - 1 + \log\frac{\bar{\sigma}_k^2}{\sigma_{k,i}^2}\right)\right\}.$$

While the choice $\mu_1 = \cdots = \mu_K = \theta^*_{T+1}$ may seem less meaningful than in the Gaussian case (with one single component), it is completely natural as the best possible choice for the parameter θ is $\mu_{P_{T+1}}$. In the computation, each component $\mathcal{N}(\mu_k, \sigma_k^2)$ of the mixture brings an error term which can be decomposed between a bias term and a variance term,

$$\mathbb{E}_{\theta\sim\mathcal{N}(\mu_k,\sigma_k^2)}\left[\|\theta-\theta^*_{T+1}\|^2\right] = \underbrace{\|\mu_k-\theta^*_{T+1}\|^2}_{\text{bias term (first order)}} + \underbrace{\sigma_k^2}_{\text{variance term (second order)}},$$

for which the choice $\mu_k = \theta^*_{T+1}$ minimizes the first order error term. Next, we set the family $\mathcal{G}$ of distributions on $\mathcal{F}$:

$$\mathcal{G} = \left\{ q_{\delta,\tau,\xi^2,b} = \mathrm{Dir}(\delta)\otimes\bigotimes_{\substack{k\in[K]\\ i\in[d]}}\mathcal{N}(\tau_{k,i},\xi_k^2)\otimes\bigotimes_{k=1}^K\Gamma(2,b_k) : \right.$$
$$\left. \delta=(\delta_1,\ldots,\delta_K)\in\mathbb{R}^K, \forall(k,i), \xi_k^2>0, \tau_{k,i}\in\mathbb{R}, b_k>0, \delta_k>0 \right\},$$

where $\mathrm{Dir}(\delta)$ is the Dirichlet distribution of parameter δ. We set the prior on priors $\Lambda = q_{1_K,0,\bar{\bar{\xi}}^2,\bar{\bar{b}}}$, where $1_K = (1,\ldots,1)$ and $\bar{\bar{\xi}}^2 = \left(\bar{\bar{\xi}}_1^2,\ldots,\bar{\bar{\xi}}_K^2\right)$. Then, using (19), (20) and (21),

$$\mathrm{KL}(q_{\delta,\tau,\xi^2,b}\|\Lambda) = \log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^K\Gamma(\delta_k)} + \sum_{k=1}^K(\delta_k-1)\left(\psi(\delta_k)-\psi(1^\top\delta)\right)$$
$$+\frac{1}{2}\sum_{k,i}\left(\frac{\tau_{k,i}^2}{\bar{\bar{\xi}}_k^2}+\frac{\xi_k^2}{\bar{\bar{\xi}}_k^2}-1+\log\frac{\bar{\bar{\xi}}_k^2}{\xi_k^2}\right)+2\sum_{k=1}^K\left(\log\frac{b_k}{\bar{\bar{b}}_k}+\frac{\bar{\bar{b}}_k-b_k}{b_k}\right),$$

where ψ is the digamma function. We can next use the bound from Theorem 5 and we have

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^*$$
$$\leq 4\inf_{\delta,\tau,\xi^2,b}\left\{\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{\delta,\tau,\xi^2,b}}\mathbb{E}_{P_{T+1}}\left[\inf_{w,\sigma^2}\left\{L\sum_{k=1}^K w_k\|\sigma_k\|^2+\frac{\mathrm{KL}(w\|\bar{w})}{\alpha n}\right.\right.\right.$$
$$\left.\left.\left.+\frac{1}{2\alpha n}\sum_{k=1}^K w_k\sum_{i=1}^d\left(\frac{(\theta^*_{T+1,i}-\bar{\mu}_{k,i})^2}{\bar{\sigma}_k^2}+\frac{\sigma_{k,i}^2}{\bar{\sigma}_k^2}-1+\log\frac{\bar{\sigma}_k^2}{\sigma_{k,i}^2}\right)\right\}\right]$$
$$+\frac{1}{2\beta T}\log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^K\Gamma(\delta_k)}+\frac{1}{2\beta T}\sum_{k=1}^K(\delta_k-1)\left(\psi(\delta_k)-\psi(1^\top\delta)\right)$$

$$+ \frac{1}{4\beta T} \sum_{k,i} \left(\frac{\tau_{k,i}^2}{\bar{\xi}_k^2} + \frac{\xi_k^2}{\bar{\xi}_k^2} - 1 + \log \frac{\bar{\xi}_k^2}{\xi_k^2} \right) + \frac{1}{\beta T} \sum_{k=1}^{K} \left(\log \frac{b_k}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_k}{b_k} \right) \Bigg\}.$$

Next, minimizing over $\sigma_{k,i}^2$ gives the optimal value $\frac{\bar{\sigma}_k^2}{2\alpha n L \bar{\sigma}_k^2 + 1}$, and replacing in the above bound gives

$$\begin{aligned}
\mathbb{E}_{P_{1:T}} \mathbb{E}_{S_{1:T}} \mathbb{E}_{\pi \sim \hat{\Pi}} [\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4 \inf_{\delta, \tau, \xi^2, b} \Bigg\{ & \mathbb{E}_{(\bar{w}, \bar{\mu}, \bar{\sigma}^2) \sim q_{\delta, \tau, \xi^2, b}} \mathbb{E}_{P_{T+1}} \Bigg[\\
& \inf_{w} \left\{ \frac{\mathrm{KL}(w \| \bar{w})}{\alpha n} + \frac{d}{2\alpha n} \sum_{k=1}^{K} w_k \log \left(2\alpha n L \bar{\sigma}_k^2 + 1 \right) + \frac{1}{2\alpha n} \sum_{k=1}^{K} w_k \frac{\| \theta_{T+1}^* - \bar{\mu}_k \|^2}{\bar{\sigma}_k^2} \right\} \Bigg] \\
& + \frac{1}{2\beta T} \log \frac{\Gamma(1^\top \delta)}{\Gamma(K) \times \prod_{k=1}^{K} \Gamma(\delta_k)} + \frac{1}{2\beta T} \sum_{k=1}^{K} (\delta_k - 1) \left(\psi(\delta_k) - \psi(1^\top \delta) \right) \\
& + \frac{1}{4\beta T} \sum_{k,i} \left(\frac{\tau_{k,i}^2}{\bar{\xi}_k^2} + \frac{\xi_k^2}{\bar{\xi}_k^2} - 1 + \log \frac{\bar{\xi}_k^2}{\xi_k^2} \right) + \frac{1}{\beta T} \sum_{k=1}^{K} \left(\log \frac{b_k}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_k}{b_k} \right) \Bigg\}.
\end{aligned}$$

We are going to restrict the infimum $\inf_w$ to the set of w such that $w_k \in \{0, 1\}$ for any $k \in [K]$. In other words, we are selecting only the best component of the mixture in the optimization bound. The reader can check that this is actually the exact solution to the minimization problem in the above bound. As a result of this minimization, the bound becomes

$$\begin{aligned}
\mathbb{E}_{P_{1:T}} \mathbb{E}_{S_{1:T}} \mathbb{E}_{\pi \sim \hat{\Pi}} [\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4 \inf_{\delta, \tau, \xi^2, b} \Bigg\{ & \mathbb{E}_{(\bar{w}, \bar{\mu}, \bar{\sigma}^2) \sim q_{\delta, \tau, \xi^2, b}} \mathbb{E}_{P_{T+1}} \Bigg[\\
& \min_{k \in [K]} \left\{ \frac{1}{\alpha n} \log \frac{1}{\bar{w}_k} + \frac{d}{2\alpha n} \log \left(2\alpha n L \bar{\sigma}_k^2 + 1 \right) + \frac{1}{2\alpha n} \frac{\| \theta_{T+1}^* - \bar{\mu}_k \|^2}{\bar{\sigma}_k^2} \right\} \Bigg] \\
& + \frac{1}{2\beta T} \log \frac{\Gamma(1^\top \delta)}{\Gamma(K) \times \prod_{k=1}^{K} \Gamma(\delta_k)} + \frac{1}{2\beta T} \sum_{k=1}^{K} (\delta_k - 1) \left(\psi(\delta_k) - \psi(1^\top \delta) \right) \\
& + \frac{1}{4\beta T} \sum_{k,i} \left(\frac{\tau_{k,i}^2}{\bar{\xi}_k^2} + \frac{\xi_k^2}{\bar{\xi}_k^2} - 1 + \log \frac{\bar{\xi}_k^2}{\xi_k^2} \right) + \frac{1}{\beta T} \sum_{k=1}^{K} \left(\log \frac{b_k}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_k}{b_k} \right) \Bigg\}.
\end{aligned}$$

Please note that the term inside the expectation is, up to the minimum on $k \in [K]$, identical to the one we had in the case of one single Gaussian mixture, except for the term $\frac{1}{\alpha n} \log \frac{1}{\bar{w}_k}$, which may be seen as a penalty for the choice of the component $k \in [K]$ in the mixture. We then bound the expectation term in the above bound by first using Fubini's theorem, and then inverting the minimum and the second expectation:

$$\begin{aligned}
\mathbb{E}_{P_{1:T}} \mathbb{E}_{S_{1:T}} \mathbb{E}_{\pi \sim \hat{\Pi}} [\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4 \inf_{\delta, \tau, \xi^2, b} \Bigg\{ & \mathbb{E}_{P_{T+1}} \Bigg[\min_{k \in [K]} \Bigg\{ \\
& \mathbb{E}_{(\bar{w}, \bar{\mu}, \bar{\sigma}^2) \sim q_{\delta, \tau, \xi^2, b}} \left[\frac{1}{\alpha n} \log \frac{1}{\bar{w}_k} + \frac{d}{2\alpha n} \log \left(2\alpha n L \bar{\sigma}_k^2 + 1 \right) + \frac{1}{2\alpha n} \frac{\| \theta_{T+1}^* - \bar{\mu}_k \|^2}{\bar{\sigma}_k^2} \right] \Bigg\} \Bigg]
\end{aligned}$$

$$+\frac{1}{2\beta T}\log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^K\Gamma(\delta_k)}+\frac{1}{2\beta T}\sum_{k=1}^K(\delta_k-1)\left(\psi(\delta_k)-\psi(1^\top\delta)\right)$$
$$+\frac{1}{4\beta T}\sum_{k,i}\left(\frac{\tau_{k,i}^2}{\bar{\bar{\xi}}_k^2}+\frac{\xi_k^2}{\bar{\bar{\xi}}_k^2}-1+\log\frac{\bar{\bar{\xi}}_k^2}{\xi_k^2}\right)+\frac{1}{\beta T}\sum_{k=1}^K\left(\log\frac{b_k}{\bar{\bar{b}}_k}+\frac{\bar{\bar{b}}_k-b_k}{b_k}\right)\Bigg\}. \quad (29)$$

We can then bound the expectation term, which we decompose as

$$\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{\delta,\tau,\xi^2,b}}\left[\frac{1}{\alpha n}\log\frac{1}{\bar{w}_k}+\frac{d}{2\alpha n}\log\left(2\alpha nL\bar{\sigma}_k^2+1\right)+\frac{1}{2\alpha n}\frac{\|\theta^*_{T+1}-\bar{\mu}_k\|^2}{\bar{\sigma}_k^2}\right]$$
$$=\frac{1}{\alpha n}\mathbb{E}_{\bar{w}\sim\mathrm{Dir}(\delta)}\left[\log\frac{1}{\bar{w}_k}\right]+\frac{d}{2\alpha n}\mathbb{E}_{\bar{\sigma}_k^2\sim\Gamma(2,b_k)}\left[\log\left(2\alpha nL\bar{\sigma}_k^2+1\right)\right]$$
$$+\frac{1}{2\alpha n}\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{\delta,\tau,\xi^2,b}}\left[\frac{\|\theta^*_{T+1}-\bar{\mu}_k\|^2}{\bar{\sigma}_k^2}\right].$$

Jensen's inequality helps to bound both the first term

$$\frac{1}{\alpha n}\mathbb{E}_{\bar{w}\sim\mathrm{Dir}(\delta)}\left[\log\frac{1}{\bar{w}_k}\right]\le\frac{1}{\alpha n}\log\mathbb{E}_{\bar{w}\sim\mathrm{Dir}(\delta)}\left[\frac{1}{\bar{w}_k}\right]$$
$$=\frac{1}{\alpha n}\log\frac{1^\top\delta-1}{\delta_k-1}$$

and the second term

$$\frac{d}{2\alpha n}\mathbb{E}_{\bar{\sigma}_k^2\sim\Gamma(2,b_k)}\left[\log\left(2\alpha nL\bar{\sigma}_k^2+1\right)\right]\le\frac{d}{2\alpha n}\log\left(2\alpha nL\mathbb{E}_{\bar{\sigma}_k^2\sim\Gamma(2,b_k)}\left[\bar{\sigma}_k^2\right]+1\right)$$
$$=\frac{d}{2\alpha n}\log\left(\frac{4L\alpha n}{b_k}+1\right)$$

in the decomposition. The third term can be computed as follows

$$\frac{1}{2\alpha n}\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{\delta,\tau,\xi^2,b}}\left[\frac{\|\theta^*_{T+1}-\bar{\mu}_k\|^2}{\bar{\sigma}_k^2}\right]=\frac{b_k}{2\alpha n}\mathbb{E}_{\bar{\mu}_k\sim\mathcal{N}(\tau_k,\xi_k^2I_d)}\left[\|\theta^*_{T+1}-\bar{\mu}_k\|^2\right]$$
$$=\frac{b_k}{2\alpha n}\left(\|\theta^*_{T+1}-\tau_k\|^2+d\xi_k^2\right).$$

The bound on the expectation then becomes

$$\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{\delta,\tau,\xi^2,a,b}}\left[\frac{1}{\alpha n}\log\frac{1}{\bar{w}_k}+\frac{d}{2\alpha n}\log\left(2\alpha nL\bar{\sigma}_k^2+1\right)+\frac{1}{2\alpha n}\frac{\|\theta^*_{T+1}-\bar{\mu}_k\|^2}{\bar{\sigma}_k^2}\right]$$
$$\le\frac{1}{\alpha n}\log\frac{1^\top\delta-1}{\delta_k-1}+\frac{d}{2\alpha n}\log\left(\frac{4L\alpha n}{b_k}+1\right)+\frac{b_k}{2\alpha n}\left(\|\theta^*_{T+1}-\tau_k\|^2+d\xi_k^2\right).$$

In our final bound, we wish to have as few terms as possible in $O\left(\frac{1}{n}\right)$ while the terms in $O\left(\frac{1}{T}\right)$ are not so problematic, because they correspond to the fast convergence rate at the meta-level. For this reason, we are going to take out of the infimum:

- the term $\frac{d}{2\alpha n}\log\left(\frac{4L\alpha n}{b_k}+1\right)$, which is unavoidable and corresponds to the main term of the bound in the worst case, with a $O\left(\frac{1}{n}\right)$ speed of convergence;
- the term $\frac{b_k d\xi_k^2}{2\alpha n}$, which will be handled through an optimization in ξ_k^2 and will be a $O\left(\frac{1}{T}\right)$ term.

As a consequence, we bound the minimum on $[K]$ by

$$\min_{k\in[K]}\left\{\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{\delta,\tau,\xi^2,b}}\left[\frac{1}{\alpha n}\log\frac{1}{\bar{w}_k}+\frac{d}{2\alpha n}\log\left(2\alpha nL\bar{\sigma}_k^2+1\right)+\frac{1}{2\alpha n}\frac{\|\theta_{T+1}^*-\bar{\mu}_k\|^2}{\bar{\sigma}_k^2}\right]\right\}\leq$$
$$\frac{d}{2\alpha n}\max_{k\in[K]}\left\{\log\left(\frac{4\alpha Ln}{b_k}+1\right)\right\}+\frac{1}{2\alpha n}\sum_{k=1}^K b_k d\xi_k^2+\frac{1}{\alpha n}\min_{k\in[K]}\left\{\frac{b_k}{2}\|\theta_{T+1}^*-\tau_k\|^2+\log\frac{1^\top\delta-1}{\delta_k-1}\right\}, \tag{30}$$

and plugging this result in (29) gives

$$\begin{aligned}\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)]-\mathcal{E}^*\leq 4\inf_{\delta,\tau,\xi^2,b}\Bigg\{&\frac{d}{2\alpha n}\max_{k\in[K]}\left\{\log\left(\frac{4\alpha Ln}{b_k}+1\right)\right\}\\&+\frac{1}{2\alpha n}\sum_{k=1}^K b_k d\xi_k^2+\frac{1}{\alpha n}\mathbb{E}_{P_{T+1}}\left[\min_{k\in[K]}\left\{\frac{b_k}{2}\|\theta_{T+1}^*-\tau_k\|^2+\log\frac{1^\top\delta-1}{\delta_k-1}\right\}\right]\\&+\frac{1}{2\beta T}\log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^K\Gamma(\delta_k)}+\frac{1}{2\beta T}\sum_{k=1}^K(\delta_k-1)\left(\psi(\delta_k)-\psi(1^\top\delta)\right)\\&+\frac{1}{4\beta T}\sum_{k=1}^K\frac{\|\tau_k\|^2}{\bar{\bar{\xi}}_k^2}+\frac{d}{4\beta T}\sum_{k=1}^K\left(\frac{\xi_k^2}{\bar{\bar{\xi}}_k^2}-1+\log\frac{\bar{\bar{\xi}}_k^2}{\xi_k^2}\right)+\frac{1}{\beta T}\sum_{k=1}^K\left(\log\frac{b_k}{\bar{\bar{b}}_k}+\frac{\bar{\bar{b}}_k-b_k}{b_k}\right)\Bigg\}.\end{aligned}$$

An exact optimization in ξ_k^2 gives

$$\xi_k^2=\frac{\bar{\bar{\xi}}_k^2}{1+\frac{2b_k\bar{\bar{\xi}}_k^2\beta T}{\alpha n}},$$

and replacing in the bound yields

$$\begin{aligned}\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)]-\mathcal{E}^*\leq 4\inf_{\delta,\tau,b}\Bigg\{&\frac{d}{2\alpha n}\max_{k\in[K]}\left\{\log\left(\frac{4\alpha Ln}{b_k}+1\right)\right\}\\&+\frac{1}{\alpha n}\mathbb{E}_{P_{T+1}}\left[\min_{k\in[K]}\left\{\frac{b_k}{2}\|\theta_{T+1}^*-\tau_k\|^2+\log\frac{1^\top\delta-1}{\delta_k-1}\right\}\right]\\&+\frac{1}{2\beta T}\log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^K\Gamma(\delta_k)}+\frac{1}{2\beta T}\sum_{k=1}^K(\delta_k-1)\left(\psi(\delta_k)-\psi(1^\top\delta)\right)\\&+\frac{1}{4\beta T}\sum_{k=1}^K\frac{\|\tau_k\|^2}{\bar{\bar{\xi}}_k^2}+\frac{d}{4\beta T}\sum_{k=1}^K\log\left(1+\frac{2b_k\bar{\bar{\xi}}_k^2\beta T}{\alpha n}\right)+\frac{1}{\beta T}\sum_{k=1}^K\left(\log\frac{b_k}{\bar{\bar{b}}_k}+\frac{\bar{\bar{b}}_k-b_k}{b_k}\right)\Bigg\}.\end{aligned}$$

From here, we set $\delta_k = 2$ for any $k \in [K]$, which implies

$$\frac{1}{2\beta T}\sum_{k=1}^{K}(\delta_k - 1)\left(\psi(\delta_k) - \psi(1^\top \delta)\right) \leq 0$$

because ψ is increasing. Please also note that

$$\log \frac{\Gamma(1^\top \delta)}{\Gamma(K) \times \prod_{k=1}^{K}\Gamma(\delta_k)} = \log \frac{\Gamma(2K)}{\Gamma(K)} \leq K\log(2K).$$

We can then deduce the bound

$$\begin{aligned}\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4\inf_{\tau,b}\Bigg\{ &\frac{d}{2\alpha n}\max_{k\in[K]}\left\{\log\left(\frac{4\alpha L n}{b_k}+1\right)\right\} + \frac{\log(2K)}{\alpha n} \\ &+ \frac{1}{2\alpha n}\mathbb{E}_{P_{T+1}}\left[\min_{k\in[K]}\left\{b_k\|\theta^*_{T+1}-\tau_k\|^2\right\}\right] + \frac{K\log(2K)}{2\beta T} + \frac{1}{4\beta T}\sum_{k=1}^{K}\frac{\|\tau_k\|^2}{\bar{\bar{\xi}}_k^2} \\ &+ \frac{d}{4\beta T}\sum_{k=1}^{K}\log\left(1+\frac{2b_k\bar{\bar{\xi}}_k^2\beta T}{\alpha n}\right) + \frac{1}{\beta T}\sum_{k=1}^{K}\left(\log\frac{b_k}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_k}{b_k}\right)\Bigg\}.\end{aligned}$$

Let

$$\Sigma_K(\mathcal{P}) := \inf_{\tau_1,\dots,\tau_K}\mathbb{E}_{P_{T+1}}\left[\min_{k\in[K]}\|\theta^*_{T+1}-\tau_k\|^2\right],$$

it is clear that

$$\mathbb{E}_{P_{T+1}}\left[\min_{k\in[K]}\left\{b_k\|\theta^*_{T+1}-\tau_k\|^2\right\}\right] \leq \Sigma_K(\mathcal{P})\max_{k\in[K]} b_k.$$

By choosing $\tau_1,\dots,\tau_K$ minimizing $\Sigma_K(\mathcal{P})$, the previous bound becomes

$$\begin{aligned}\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4\inf_{b}\Bigg\{ &\frac{\log(2K)}{\alpha n} + \frac{d}{2\alpha n}\max_{k\in[K]}\left\{\log\left(\frac{4\alpha L n}{b_k}+1\right)\right\} \\ &+ \frac{\Sigma_K(\mathcal{P})}{2\alpha n}\max_{k\in[K]} b_k + \frac{K\log(2K)}{2\beta T} + \frac{1}{4\beta T}\sum_{k=1}^{K}\frac{\|\tau_k\|^2}{\bar{\bar{\xi}}_k^2} \\ &+ \frac{d}{4\beta T}\sum_{k=1}^{K}\log\left(1+\frac{2b_k\bar{\bar{\xi}}_k^2\beta T}{\alpha n}\right) + \frac{1}{\beta T}\sum_{k=1}^{K}\left(\log\frac{b_k}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_k}{b_k}\right)\Bigg\}.\end{aligned}$$

Please note that $\tau_1,\dots,\tau_K$ are characteristic of the distribution $\mathcal{P}$. Intuitively, if the distribution $\mathcal{P}$ has K modes, or K Gaussian mixtures, $\tau_1,\dots,\tau_K$ correspond to the centers of these modes or mixtures up to a permutation. Consequently, they do not scale with n or T, but can be regarded as problem parameters of constant order.

We now restrict the infimum in the above bound to all $(b_k)_{1\leq k\leq K}$ such that $b_1 = \dots = b_K$ and $1 \leq b_1 \leq T$. Replacing in the above bound gives

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4 \inf_{1\leq b_1\leq T}\left\{\frac{\log(2K)}{\alpha n} + \frac{d}{2\alpha n}\log\left(\frac{4\alpha L n}{b_1}+1\right)\right.$$
$$+ \frac{b_1\Sigma_K(\mathcal{P})}{2\alpha n} + \frac{K\log(2K)}{2\beta T} + \frac{1}{4\beta T}\sum_{k=1}^{K}\frac{\|\tau_k\|^2}{\bar{\bar{\xi}}_k^2}$$
$$\left. + \frac{d}{4\beta T}\sum_{k=1}^{K}\log\left(1+\frac{2b_1\bar{\bar{\xi}}_k^2\beta T}{\alpha n}\right) + \frac{1}{\beta T}\sum_{k=1}^{K}\left(\log\frac{b_1}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_1}{b_1}\right)\right\}.$$

Please note that the last two terms of this bound can be bounded, for any $1 \leq b_1 \leq T$, as

$$\frac{d}{4\beta T}\sum_{k=1}^{K}\log\left(1+\frac{2b_1\bar{\bar{\xi}}_k^2\beta T}{\alpha n}\right) + \frac{1}{\beta T}\sum_{k=1}^{K}\left(\log\frac{b_1}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_1}{b_1}\right)$$
$$\leq \frac{d}{4\beta T}\sum_{k=1}^{K}\log\left(1+\frac{2\bar{\bar{\xi}}_k^2\beta T^2}{\alpha n}\right) + \frac{1}{\beta T}\sum_{k=1}^{K}\left(\log\frac{T}{\bar{\bar{b}}_k} + \bar{\bar{b}}_k - 1\right).$$

Replacing in the above bound and extracting from the infimum the terms which do not depend on b_1 gives

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq \frac{4\log(2K)}{\alpha n} + \inf_{1\leq b_1\leq T}\left\{\frac{2d}{\alpha n}\log\left(\frac{4\alpha L n}{b_1}+1\right) + \frac{2b_1\Sigma_K(\mathcal{P})}{\alpha n}\right\}$$
$$+ \frac{2K\log(2K)}{\beta T} + \frac{1}{\beta T}\sum_{k=1}^{K}\frac{\|\tau_k\|^2}{\bar{\bar{\xi}}_k^2} + \frac{d}{\beta T}\sum_{k=1}^{K}\log\left(1+\frac{2\bar{\bar{\xi}}_k^2\beta T^2}{\alpha n}\right) + \frac{4}{\beta T}\sum_{k=1}^{K}\left(\log\frac{T}{\bar{\bar{b}}_k} + \bar{\bar{b}}_k - 1\right).$$

Let

$$\mathrm{CV}_{\text{finite}}(K,n) = \frac{4\log(2K)}{\alpha n},$$

be the bound obtained in the finite case (when learning discrete priors) and let

$$\mathrm{CV}_{\text{Gaussian}}(d,\Sigma_K(\mathcal{P}),n,T) = \inf_{1\leq b_1\leq T}\left\{\frac{2d}{\alpha n}\log\left(\frac{4\alpha L n}{b_1}+1\right) + \frac{2b_1\Sigma_K(\mathcal{P})}{\alpha n}\right\}$$

be the bound in the Gaussian case (with one component). Let also

$$\mathrm{CV}_{\text{meta}}(T,n,d,K,\bar{\bar{b}},\bar{\bar{\xi}}^2) = \frac{2K\log(2K)}{\beta T} + \frac{1}{\beta T}\sum_{k=1}^{K}\frac{\|\tau_k\|^2}{\bar{\bar{\xi}}_k^2} + \frac{d}{\beta T}\sum_{k=1}^{K}\log\left(1+\frac{2\bar{\bar{\xi}}_k^2\beta T^2}{\alpha n}\right)$$
$$+ \frac{4}{\beta T}\sum_{k=1}^{K}\left(\log\frac{T}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - T}{T}\right).$$

Then, we can write the above bound as the sum of the three defined terms:

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^*$$
$$\leq \mathrm{CV}_{\text{finite}}(K,n) + \mathrm{CV}_{\text{Gaussian}}(d,\Sigma_K(\mathcal{P}),n,T) + \mathrm{CV}_{\text{meta}}(T,n,d,K,\bar{\bar{b}},\bar{\bar{\xi}}^2), \quad (31)$$

where an interpretation of each of the terms of the above bound is given in Remark 19.

Let us now identify one regime where the meta-learning brings a considerable improvement over the learning in isolation. Assume that $\Sigma_K(\mathcal{P}) \leq \frac{dn}{T^2}$, where d is simply here for dimensionality reasons. In this regime, the distribution is concentrated around $\tau_1, \dots, \tau_K$ and the variance of the local distribution around each of those points is smaller than $\frac{n}{T^2}$. As a result, the optimal parameter in the new task $T+1$ is going to be closed to one of $\tau_1, \dots, \tau_K$ and we can infer it from the previous tasks. For that reason, we expect a significant improvement from the meta-learning over the learning in isolation in this regime.

In this case, in the decomposition of $\mathrm{CV}_{\text{Gaussian}}(d, \Sigma_K(\mathcal{P}), n, T)$, the term $\frac{2b_1\Sigma_K(\mathcal{P})}{\alpha n}$ is very small compared to the other term $\frac{d}{2\alpha n}\log\left(\frac{4\alpha L n}{b_1}+1\right)$. Therefore, we will choose b_1 large so that it minimizes the latter. The choice $b_1 = T$ provides the bound

$$\mathrm{CV}_{\text{Gaussian}}(d, \Sigma_K(\mathcal{P}), n, T) \leq \frac{8Ld}{T} + \frac{2d}{\alpha T}.$$

Besides the $\mathrm{CV}_{\text{finite}}(K, n)$ fast term required to find the K centers of the Gaussians, the convergence at the Gaussian level is done in $O\left(\frac{\log T}{T}\right)$, which is a clear improvement in the regime $T \gg n$, even over the fast rate in the learning in isolation.

Remark 18 *Please note that the general bound* (31) *comes as no surprise, because the process of learning the parameter θ in the mixture of Gaussians framework consists of three different steps:*

- *first, identifying the K centers of the mixtures which, similarly to the finite case, is captured in the $\mathrm{CV}_{\text{finite}}(K, n) = \frac{4\log(2K)}{\alpha n}$ term;*
- *then, identifying the right parameters of the Gaussian components centered on the points identified in the previous step. Similarly to the Gaussian case, this is captured in the term $\mathrm{CV}_{\text{Gaussian}}(d, \Sigma_K(\mathcal{P}), n, T)$, and can be as small as a $O\left(\frac{1}{T}\right)$ in some favorable cases.*
- *eventually, the convergence at the meta level is a $\mathrm{CV}_{\text{meta}}(T, n, d, K, \bar{\bar{b}}, \bar{\bar{\xi}}^2) = O\left(\frac{\log T}{T}\right)$ is a small penalty in $O\left(\frac{\log T}{T}\right)$ to use the meta-learning, thanks to which the previous term $\mathrm{CV}_{\text{Gaussian}}(d, \Sigma_K(\mathcal{P}), n, T)$ can potentially achieve the very fast rate $O\left(\frac{1}{T}\right)$ instead of the much slower rate $O\left(\frac{1}{n}\right)$, which is the best one can hope for when learning in isolation.*

Remark 19 *Please note that the meta-learning penalty, $\mathrm{CV}_{\text{meta}}(T, n, d, K, \bar{\bar{b}}, \bar{\bar{\xi}}^2)$, is a very small term. Indeed, up to a logarithmic factor, it follows the very fast rate $\tilde{O}\left(\frac{1}{T}\right)$. Besides, in the $T \gg n$ regime, it is even much smaller than the fast rate $O\left(\frac{1}{n}\right)$ at the within-task level. Recall that, as described in the introduction part, the regime $T \gg n$ is very common in many applications and is one of the motivations for doing meta-learning.*

Remark 20 *It may seem paradoxical that the model of mixtures of Gaussians achieves, in the best possible case, a rate of convergence $O\left(\frac{\log K}{n}\right)$ slower than the $O\left(\frac{\log T}{T}\right)$ rate achieved by a single Gaussian in the regime $n << T$ under optimal conditions. In reality, the latter rate of convergence is also achievable in the model of mixtures of Gaussians, but similarly as in the case of a single Gaussian component, it requires the strong assumption that*

$$\Sigma_1(\mathcal{P}) \leq \frac{dn}{T^2},$$

which is much more restrictive. On the other hand, many distributions only satisfy

$$\Sigma_K(\mathcal{P}) \leq \frac{dn}{T^2}$$

for some $K \geq 2$, in which case the rate of convergence $O\left(\frac{\log K}{n}\right)$ achieved here is much faster than $O\left(\frac{d \log n}{n}\right)$, which is the best possible rate achieved in the single Gaussian model in general.

H.2 What if the Number of Components in the Mixture is Unknown?

In practice, we do not know in advance how to choose the number of components K in the prior. In this case, we are going to include inside $\mathcal{M}$ all the mixtures of Gaussians, i.e.,

$$\mathcal{M} = \left\{ p_{w,\mu,\sigma^2} = \sum_{k=1}^{+\infty} w_k \bigotimes_{i=1}^{d} \mathcal{N}(\mu_{k,i}, \sigma_{k,i}^2) : \exists K \geq 1 : \forall k \geq K+1, w_k = 0 \right\}.$$

Note that the sum inside the definition of $\mathcal{F}$ is finite, since $w_k = 0$ for any k beyond a certain rank K. We still denote by $\pi = \sum_{k=1}^{K} \bar{w}_k \mathcal{N}(\bar{\mu}_k, \bar{\sigma}_k^2 I_d)$ the prior in each task. By definition, for any $k \geq K+1, \bar{w}_k = 0$. It still holds that, for any w, μ, σ^2,

$$\mathrm{KL}(p_{w,\mu,\sigma^2} \| \pi) \leq \mathrm{KL}(w \| \bar{w}) + \frac{1}{2} \sum_{k=1}^{\infty} w_k \sum_{i=1}^{d} \left(\frac{(\mu_{k,i} - \bar{\mu}_{k,i})^2}{\bar{\sigma}_k^2} + \frac{\sigma_{k,i}^2}{\bar{\sigma}_k^2} - 1 + \log \frac{\bar{\sigma}_k^2}{\sigma_{k,i}^2} \right),$$

where we denoted

$$\mathrm{KL}(w \| \bar{w}) = \sum_{k=1}^{\infty} w_k \log \frac{w_k}{\bar{w}_k}.$$

To put things clearly, the KL remains identical to the case where K is known except for the fact that the sums on k are no longer stopping at a pre-determined K. This difference aside, the bound remains identical to the one in the case where K is known, and the bound from Theorem 1 becomes, at $t = T+1$,

$$\mathbb{E}_{S_{T+1}} \mathbb{E}_{\theta \sim \rho_{T+1}(\pi,\alpha)} [R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} \leq 2 \inf_{w,\mu,\sigma^2} \left\{ \mathbb{E}_{\theta \sim p_{w,\mu,\sigma^2}} [R_{P_{T+1}}(\theta)] - R^*_{P_{T+1}} + \frac{\mathrm{KL}(w \| \bar{w})}{\alpha n} \right.$$
$$\left. + \frac{1}{2\alpha n} \sum_{k=1}^{\infty} w_k \sum_{i=1}^{d} \left(\frac{(\mu_{k,i} - \bar{\mu}_{k,i})^2}{\bar{\sigma}_k^2} + \frac{\sigma_{k,i}^2}{\bar{\sigma}_k^2} - 1 + \log \frac{\bar{\sigma}_k^2}{\sigma_{k,i}^2} \right) \right\}.$$

Next, we are going to define a prior on K within the prior of priors as follows. We assume that the number of mixtures K is smaller than T, because even if it were not, it would be impossible to identify them with enough confidence. We define the set of priors on priors

$$\mathcal{G} = \left\{ q_{x,\delta,\tau,\xi^2,b} = q_x \times q_{\delta,\tau,\xi^2,b|K} \right\},$$

where $q_x = \mathrm{Mult}(x_1, \dots, x_T)$ is the prior distribution on K and

$$q_{\delta,\tau,\xi^2,b|K} = \mathrm{Dir}(\delta_1, \dots, \delta_K) \otimes \bigotimes_{\substack{k\in[K]\\ i\in[d]}} \mathcal{N}(\tau_{k,i}, \xi_k^2) \otimes \bigotimes_{k=1}^{K} \Gamma(2, b_k),$$

and we set the prior of prior as $\Lambda = q_{\frac{1}{T}1_T, 1_K, 0, \bar{\bar{\xi}}^2, \bar{\bar{b}}}$. We also need to re-compute the KL divergence between the priors of priors, which becomes

$$\mathrm{KL}(q_{x,\delta,\tau,\xi^2,b} \| \Lambda) = \log T - H(x) + \mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\sum_{k=1}^{K}(\delta_k - 1)\left(\psi(\delta_k) - \psi(1^\top\delta)\right)\right.$$
$$\left. + \log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^{K}\Gamma(\delta_k)} + \frac{1}{2}\sum_{k,i}\left(\frac{\tau_{k,i}^2}{\bar{\bar{\xi}}_k^2} + \frac{\xi_k^2}{\bar{\bar{\xi}}_k^2} - 1 + \log\frac{\bar{\bar{\xi}}_k^2}{\xi_k^2}\right) + 2\sum_{k=1}^{K}\left(\log\frac{b_k}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_k}{b_k}\right)\right],$$

using (18). Please note that in any optimization on $\mathcal{G}$, we optimize first in $(x_1, \dots, x_T)$ and then on δ, τ, ξ^2, b conditionally on K. This means that the latter parameters are allowed to depend on K. While the infimum on $\mathcal{G}$ of any quantity should be written $\inf_x \inf_{\delta,\tau,\xi^2,b\in\sigma(K)}$, we will adopt the shortcut notation $\inf_{x,\delta,\tau,\xi^2,b}$. We can next use the bound from Theorem 5 and (10), and we have

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^*$$
$$\leq 4\inf_{x,\delta,\tau,\xi^2,b}\left\{\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{x,\delta,\tau,\xi^2,b}}\mathbb{E}_{P_{T+1}}\left[\inf_{w,\sigma^2}\left\{L\sum_{k=1}^{\infty}w_k\|\sigma_k\|^2 + \frac{\mathrm{KL}(w\|\bar{w})}{\alpha n}\right.\right.\right.$$
$$\left.\left.\left. + \frac{1}{2\alpha n}\sum_{k=1}^{\infty}w_k\sum_{i=1}^{d}\left(\frac{(\mu_{k,i}-\bar{\mu}_{k,i})^2}{\bar{\sigma}_k^2} + \frac{\sigma_{k,i}^2}{\bar{\sigma}_k^2} - 1 + \log\frac{\bar{\sigma}_k^2}{\sigma_{k,i}^2}\right)\right\}\right]$$
$$+ \frac{1}{2\beta T}\left(\log T - H(x)\right) + \frac{1}{2\beta T}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\sum_{k=1}^{K}(\delta_k - 1)\left(\psi(\delta_k) - \psi(1^\top\delta)\right)\right.$$
$$\left.\left. + \log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^{K}\Gamma(\delta_k)} + \frac{1}{2}\sum_{k,i}\left(\frac{\tau_{k,i}^2}{\bar{\bar{\xi}}_k^2} + \frac{\xi_k^2}{\bar{\bar{\xi}}_k^2} - 1 + \log\frac{\bar{\bar{\xi}}_k^2}{\xi_k^2}\right) + 2\sum_{k=1}^{K}\left(\log\frac{b_k}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_k}{b_k}\right)\right]\right\}.$$

The optimization on $\sigma_{k,i}^2$ may be performed exactly by setting $\sigma_{k,i}^2 = \frac{\bar{\sigma}_k^2}{2\alpha L n\bar{\sigma}_k^2 + 1}$, and the bound becomes

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4\inf_{x,\delta,\tau,\xi^2,b}\left\{\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{x,\delta,\tau,\xi^2,b}}\mathbb{E}_{P_{T+1}}\left[\right.\right.$$
$$\left.\inf_{w}\left\{\frac{\mathrm{KL}(w\|\bar{w})}{\alpha n} + \frac{d}{2\alpha n}\sum_{k=1}^{\infty}w_k\log\left(2\alpha L n\bar{\sigma}_k^2 + 1\right) + \frac{1}{2\alpha n}\sum_{k=1}^{\infty}w_k\frac{\|\theta_{T+1}^* - \bar{\mu}_k\|^2}{\bar{\sigma}_k^2}\right\}\right]$$
$$+ \frac{1}{2\beta T}\left(\log T - H(x)\right) + \frac{1}{2\beta T}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\sum_{k=1}^{K}(\delta_k - 1)\left(\psi(\delta_k) - \psi(1^\top\delta)\right)\right.$$

$$+\log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^K\Gamma(\delta_k)}+\frac{1}{2}\sum_{k,i}\left(\frac{\tau_{k,i}^2}{\bar{\bar{\xi}}_k^2}+\frac{\xi_k^2}{\bar{\bar{\xi}}_k^2}-1+\log\frac{\bar{\bar{\xi}}_k^2}{\xi_k^2}\right)+2\sum_{k=1}^K\left(\log\frac{b_k}{\bar{\bar{b}}_k}+\frac{\bar{\bar{b}}_k-b_k}{b_k}\right)\Bigg]\Bigg\}.$$

We restrict the optimization in w to the set $\{(w_k)_{k\geq 1}:\exists k_0\leq K: w_{k_0}=1,\forall k\neq k_0, w_k=0\}$, and the bound becomes

$$\begin{aligned}\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)]-\mathcal{E}^*\leq 4\inf_{x,\delta,\tau,\xi^2,b}\Bigg\{&\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{x,\delta,\tau,\xi^2,b}}\mathbb{E}_{P_{T+1}}\Bigg[\\&\min_{k\in[K]}\left\{\frac{1}{\alpha n}\log\frac{1}{\bar{w}_k}+\frac{d}{2\alpha n}\log\left(2\alpha Ln\bar{\sigma}_k^2+1\right)+\frac{1}{2\alpha n}\frac{\|\theta_{T+1}^*-\bar{\mu}_k\|^2}{\bar{\sigma}_k^2}\right\}\Bigg]\\&+\frac{1}{2\beta T}\left(\log T-H(x)\right)+\frac{1}{2\beta T}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\sum_{k=1}^K(\delta_k-1)\left(\psi(\delta_k)-\psi(1^\top\delta)\right)\right.\\&\left.+\log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^K\Gamma(\delta_k)}+\frac{1}{2}\sum_{k,i}\left(\frac{\tau_{k,i}^2}{\bar{\bar{\xi}}_k^2}+\frac{\xi_k^2}{\bar{\bar{\xi}}_k^2}-1+\log\frac{\bar{\bar{\xi}}_k^2}{\xi_k^2}\right)+2\sum_{k=1}^K\left(\log\frac{b_k}{\bar{\bar{b}}_k}+\frac{\bar{\bar{b}}_k-b_k}{b_k}\right)\right]\Bigg\}.\end{aligned}$$

Next, we classically decompose the expectation $\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{x,\delta,\tau,\xi^2,b}}[X]$ as

$$\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{x,\delta,\tau,\xi^2,b}}[X]=\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{\delta,\tau,\xi^2,b|K}}[X]\right].$$

Applying Fubini's theorem and inverting the infimum and the expectation yields the bound

$$\begin{aligned}\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)]-\mathcal{E}^*\leq 4\inf_{x,\delta,\tau,\xi^2,b}\Bigg\{&\mathbb{E}_{K\sim\mathrm{Mult}(x)}\mathbb{E}_{P_{T+1}}\Bigg[\min_{k\in[K]}\Bigg\{\\&\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{\delta,\tau,\xi^2,b|K}}\left[\frac{1}{\alpha n}\log\frac{1}{\bar{w}_k}+\frac{d}{2\alpha n}\log\left(2\alpha Ln\bar{\sigma}_k^2+1\right)+\frac{1}{2\alpha n}\frac{\|\theta_{T+1}^*-\bar{\mu}_k\|^2}{\bar{\sigma}_k^2}\right]\Bigg\}\Bigg]\\&+\frac{1}{2\beta T}\left(\log T-H(x)\right)+\frac{1}{2\beta T}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\sum_{k=1}^K(\delta_k-1)\left(\psi(\delta_k)-\psi(1^\top\delta)\right)\right.\\&\left.+\log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^K\Gamma(\delta_k)}+\frac{1}{2}\sum_{k,i}\left(\frac{\tau_{k,i}^2}{\bar{\bar{\xi}}_k^2}+\frac{\xi_k^2}{\bar{\bar{\xi}}_k^2}-1+\log\frac{\bar{\bar{\xi}}_k^2}{\xi_k^2}\right)+2\sum_{k=1}^K\left(\log\frac{b_k}{\bar{\bar{b}}_k}+\frac{\bar{\bar{b}}_k-b_k}{b_k}\right)\right]\Bigg\}.\end{aligned}$$

The bound (30) from the previous section still holds:

$$\begin{aligned}&\min_{k\in[K]}\left\{\mathbb{E}_{(\bar{w},\bar{\mu},\bar{\sigma}^2)\sim q_{\delta,\tau,\xi^2,b|K}}\left[\frac{1}{\alpha n}\log\frac{1}{\bar{w}_k}+\frac{d}{2\alpha n}\log\left(2\alpha Ln\bar{\sigma}_k^2+1\right)+\frac{1}{2\alpha n}\frac{\|\theta_{T+1}^*-\bar{\mu}_k\|^2}{\bar{\sigma}_k^2}\right]\right\}\\&\leq\frac{d}{2\alpha n}\max_{k\in[K]}\left\{\log\left(\frac{4\alpha Ln}{b_k}+1\right)\right\}+\sum_{k=1}^K\frac{b_k d\xi_k^2}{2\alpha n}+\frac{1}{\alpha n}\min_{k\in[K]}\left\{\frac{b_k}{2}\|\theta_{T+1}^*-\tau_k\|^2+\log\frac{1^\top\delta-1}{\delta_k-1}\right\},\end{aligned}$$

and we can inject it in the computation so that the bound becomes

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4 \inf_{x,\delta,\tau,\xi^2,b} \left\{ \frac{d}{2\alpha n}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\max_{k\in[K]} \log\left(\frac{4\alpha L n}{b_k}+1\right)\right] \right.$$
$$+\frac{d}{2\alpha n}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\sum_{k=1}^K b_k\xi_k^2\right] + \frac{1}{\alpha n}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\mathbb{E}_{P_{T+1}}\left[\min_{k\in[K]}\left\{\frac{b_k}{2}\|\theta^*_{T+1}-\tau_k\|^2 + \log\frac{1^\top\delta-1}{\delta_k-1}\right\}\right]$$
$$+\frac{1}{2\beta T}\left(\log T - H(x)\right) + \frac{1}{2\beta T}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\sum_{k=1}^K(\delta_k-1)\left(\psi(\delta_k)-\psi(1^\top\delta)\right)\right.$$
$$\left.\left. +\log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^K\Gamma(\delta_k)} + \frac{1}{2}\sum_{k,i}\left(\frac{\tau_{k,i}^2}{\bar{\bar{\xi}}_k^2}+\frac{\xi_k^2}{\bar{\bar{\xi}}_k^2}-1+\log\frac{\bar{\bar{\xi}}_k^2}{\xi_k^2}\right) + 2\sum_{k=1}^K\left(\log\frac{b_k}{\bar{\bar{b}}_k}+\frac{\bar{\bar{b}}_k-b_k}{b_k}\right)\right]\right\}.$$

An exact optimization in ξ_k^2 yields $\xi_k^2 = \frac{\bar{\bar{\xi}}_k^2}{1+\frac{2b_k\bar{\bar{\xi}}_k^2\beta T}{\alpha n}}$ and we can replace in the bound

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4 \inf_{x,\delta,\tau,b} \left\{ \frac{d}{2\alpha n}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\max_{k\in[K]} \log\left(\frac{4\alpha L n}{b_k}+1\right)\right] \right.$$
$$+\frac{1}{\alpha n}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\mathbb{E}_{P_{T+1}}\left[\min_{k\in[K]}\left\{\frac{b_k}{2}\|\theta^*_{T+1}-\tau_k\|^2 + \log\frac{1^\top\delta-1}{\delta_k-1}\right\}\right] + \frac{1}{2\beta T}\left(\log T - H(x)\right)$$
$$+\frac{1}{2\beta T}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\sum_{k=1}^K(\delta_k-1)\left(\psi(\delta_k)-\psi(1^\top\delta)\right) + \log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^K\Gamma(\delta_k)}\right.$$
$$\left.\left. + \frac{1}{2}\sum_{k=1}^K\frac{\|\tau_k\|^2}{\bar{\bar{\xi}}_k^2} + \frac{d}{2}\sum_{k=1}^K\log\left(1+\frac{2b_k\bar{\bar{\xi}}_k^2\beta T}{\alpha n}\right) + 2\sum_{k=1}^K\left(\log\frac{b_k}{\bar{\bar{b}}_k}+\frac{\bar{\bar{b}}_k-b_k}{b_k}\right)\right]\right\}.$$

We choose $\delta_k = 2$ for any $k \geq 1$ and noting that both

$$\sum_{k=1}^K(\delta_k-1)\left(\psi(\delta_k)-\psi(1^\top\delta)\right) \leq 0$$

and

$$\log\frac{\Gamma(1^\top\delta)}{\Gamma(K)\times\prod_{k=1}^K\Gamma(\delta_k)} = \log\frac{\Gamma(2K)}{\Gamma(K)} \leq K\log(2K),$$

we deduce the following bound:

$$\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \leq 4 \inf_{x,\tau,b} \left\{ \frac{d}{2\alpha n}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\max_{k\in[K]} \log\left(\frac{4\alpha L n}{b_k}+1\right)\right] \right.$$
$$+\frac{1}{2\alpha n}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\mathbb{E}_{P_{T+1}}\left[\min_{k\in[K]}\left\{b_k\|\theta^*_{T+1}-\tau_k\|^2\right\}\right]$$
$$+\frac{1}{\alpha n}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\log(2K)\right] + \frac{1}{2\beta T}\left(\log T - H(x)\right) + \frac{1}{2\beta T}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[K\log(2K)\right.$$
$$\left.\left. + \frac{1}{2}\sum_{k=1}^K\frac{\|\tau_k\|^2}{\bar{\bar{\xi}}_k^2} + \frac{d}{2}\sum_{k=1}^K\log\left(1+\frac{2b_k\bar{\bar{\xi}}_k^2\beta T}{\alpha n}\right) + 2\sum_{k=1}^K\left(\log\frac{b_k}{\bar{\bar{b}}_k}+\frac{\bar{\bar{b}}_k-b_k}{b_k}\right)\right]\right\}.$$

Recall the (unchanged) definition of $\Sigma_K(\mathcal{P})$:

$$\Sigma_K(\mathcal{P}) = \inf_{\tau_1,\dots,\tau_K} \mathbb{E}_{P_{T+1}}\left[\min_{k\in[K]} \|\theta^*_{T+1} - \tau_k\|^2\right].$$

Recalling that τ (as well as b) is allowed to depend on K, we define $(\tau_1, \dots, \tau_K)$ as the argument (up to a permutation) of $\Sigma_K(\mathcal{P})$. It follows that the bound becomes

$$\begin{aligned}\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \le 4 \inf_{x,b}\Bigg\{ &\frac{d}{2\alpha n}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\max_{k\in[K]}\log\left(\frac{4\alpha L n}{b_k}+1\right)\right] \\ &+ \frac{1}{2\alpha n}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\left[\Sigma_K(\mathcal{P})\max_{k\in[K]} b_k + 2\log(2K)\right] + \frac{1}{2\beta T}\left(\log T - H(x)\right) \\ &+ \frac{1}{2\beta T}\mathbb{E}_{K\sim\mathrm{Mult}(x)}\Bigg[K\log(2K) + \frac{1}{2}\sum_{k=1}^K \frac{\|\tau_k\|^2}{\bar{\bar{\xi}}_k^2} + \frac{d}{2}\sum_{k=1}^K \log\left(1 + \frac{2b_k\bar{\bar{\xi}}_k^2\beta T}{\alpha n}\right) \\ &\qquad + 2\sum_{k=1}^K\left(\log\frac{b_k}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_k}{b_k}\right)\Bigg]\Bigg\}.\end{aligned}$$

At this stage, the proof slightly differs from the case where the number of mixtures is known. Now, we choose to restrict the infimum on all the multinomial distributions $\mathrm{Mult}(x_1, \dots, x_T)$ to all the Dirac masses, i.e., all the $(x_1, \dots, x_T)$ such that there exists $K \in \{1, \dots, T\}$ such that $x_K = 1$. It follows that $H(x) = 0$ and we deduce that

$$\begin{aligned}\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \le 4 \inf_{K\in[T]}\inf_{b}\Bigg\{ &\frac{d}{2\alpha n}\max_{k\in[K]}\log\left(\frac{4\alpha L n}{b_k}+1\right) \\ &+ \frac{1}{2\alpha n}\left(\Sigma_K(\mathcal{P})\max_{k\in[K]} b_k + 2\log(2K)\right) + \frac{\log T}{2\beta T} \\ &+ \frac{K\log(2K)}{2\beta T} + \frac{1}{4\beta T}\sum_{k=1}^K \frac{\|\tau_k\|^2}{\bar{\bar{\xi}}_k^2} + \frac{d}{4\beta T}\sum_{k=1}^K \log\left(1 + \frac{2b_k\bar{\bar{\xi}}_k^2\beta T}{\alpha n}\right) \\ &\qquad + \frac{1}{\beta T}\sum_{k=1}^K\left(\log\frac{b_k}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_k}{b_k}\right)\Bigg\}.\end{aligned}$$

Similarly as in the case of known number of mixtures, we restrict the infimum on $(b_k)_{1\le k\le K}$ to the sequences such that $b_1 = \cdots = b_K$ and $1 \le b_1 \le T$. We can then replace in the above equation, and it yields

$$\begin{aligned}\mathbb{E}_{P_{1:T}}\mathbb{E}_{S_{1:T}}\mathbb{E}_{\pi\sim\hat{\Pi}}[\mathcal{E}(\pi)] - \mathcal{E}^* \le 4 \inf_{K\in[T]}\inf_{1\le b_1\le T}\Bigg\{ &\frac{d}{2\alpha n}\log\left(\frac{4\alpha L n}{b_1}+1\right) \\ &+ \frac{b_1\Sigma_K(\mathcal{P})}{2\alpha n} + \frac{\log(2K)}{\alpha n} + \frac{\log T}{2\beta T} \\ &+ \frac{K\log(2K)}{2\beta T} + \frac{1}{4\beta T}\sum_{k=1}^K \frac{\|\tau_k\|^2}{\bar{\bar{\xi}}_k^2} + \frac{d}{4\beta T}\sum_{k=1}^K \log\left(1 + \frac{2b_1\bar{\bar{\xi}}_k^2\beta T}{\alpha n}\right)\end{aligned}$$

$$+ \frac{1}{\beta T} \sum_{k=1}^{K} \left(\log \frac{b_1}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_1}{b_1} \right) \Bigg\}.$$

Similarly as before, we bound the last two terms of the sum as follows:

$$\frac{d}{4\beta T} \sum_{k=1}^{K} \log \left(1 + \frac{2b_1 \bar{\bar{\xi}}_k^2 \beta T}{\alpha n} \right) + \frac{1}{\beta T} \sum_{k=1}^{K} \left(\log \frac{b_1}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_1}{b_1} \right) \leq$$
$$\frac{d}{4\beta T} \sum_{k=1}^{K} \log \left(1 + \frac{2\bar{\bar{\xi}}_k^2 \beta T^2}{\alpha n} \right) + \frac{1}{\beta T} \sum_{k=1}^{K} \left(\log \frac{T}{\bar{\bar{b}}_k} + \bar{\bar{b}}_k - 1 \right),$$

and replacing in the bound gives

$$\mathbb{E}_{P_{1:T}} \mathbb{E}_{S_{1:T}} \mathbb{E}_{\pi \sim \hat{\Pi}} [\mathcal{E}(\pi)] - \mathcal{E}^* \leq \inf_{K \in [T]} \Bigg\{ \frac{4 \log(2K)}{\alpha n}$$
$$+ \inf_{1 \leq b_1 \leq T} \left\{ \frac{2d}{\alpha n} \log \left(\frac{4\alpha L n}{b_1} + 1 \right) + \frac{2b_1 \Sigma_K(\mathcal{P})}{\alpha n} \right\} + \frac{2 \log T}{\beta T}$$
$$+ \frac{2K \log(2K)}{\beta T} + \frac{1}{\beta T} \sum_{k=1}^{K} \frac{\|\tau_k\|^2}{\bar{\bar{\xi}}_k^2} + \frac{d}{\beta T} \sum_{k=1}^{K} \log \left(1 + \frac{2b_1 \bar{\bar{\xi}}_k^2 \beta T^2}{\alpha n} \right)$$
$$+ \frac{4}{\beta T} \sum_{k=1}^{K} \left(\log \frac{b_1}{\bar{\bar{b}}_k} + \frac{\bar{\bar{b}}_k - b_1}{b_1} \right) \Bigg\}.$$

With the exact same notations as in the case where the number of mixtures K is known, we can rewrite the bound as

$$\mathbb{E}_{P_{1:T}} \mathbb{E}_{S_{1:T}} \mathbb{E}_{\pi \sim \hat{\Pi}} [\mathcal{E}(\pi)] - \mathcal{E}^*$$
$$\leq \inf_{K \in [T]} \left\{ \mathrm{CV}_{\text{finite}}(K, n) + \mathrm{CV}_{\text{Gaussian}}\left(d, \Sigma_K(\mathcal{P}), n, T\right) + \mathrm{CV}_{\text{meta}}^{\text{unknown}}(T, n, d, K, \bar{\bar{b}}, \bar{\bar{\xi}}^2) \right\},$$

where $\mathrm{CV}_{\text{finite}}(K, n)$ and $\mathrm{CV}_{\text{Gaussian}}\left(d, K, \Sigma_K(\mathcal{P}), n, T\right)$ are exactly the same terms as in the case where the number of mixtures K is known, and the convergence term at the meta level becomes

$$\mathrm{CV}_{\text{meta}}^{\text{unknown}}(T, n, d, K, \bar{\bar{b}}, \bar{\bar{\xi}}^2) = \mathrm{CV}_{\text{meta}}(T, n, d, K, \bar{\bar{b}}, \bar{\bar{\xi}}^2) + \frac{2 \log T}{\beta T}.$$

Even when the number of mixtures K is unknown, the same bound as in the case of K known can be achieved up to a $\frac{2 \log T}{\beta T}$ term, which is the order of the time required to find the optimal number of components in the mixture at the meta level.

References

A. Achille, M. Lam, R. Tewari, A. Ravichandran, S. Maji, C. C. Fowlkes, S. Soatto, and P. Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6430–6439, 2019.

M. Al-Shedivat, L. Li, E. Xing, and A. Talwalkar. On data efficiency of meta-learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1566–1575. PMLR, 09–15 Jun 2019.

P. Alquier. User-friendly Introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024.

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(239):1–41, 2016.

P. Alquier, T. T. Mai, and M. Pontil. Regret Bounds for Lifelong Learning. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 261–269, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

R. Amit and R. Meir. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 205–214. PMLR, 10–15 Jul 2018.

C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50:5–43, 2003.

Y. Bai, M. Chen, Pa. Zhou, T. Zhao, J. Lee, S. M. Kakade, H. Wang, and C. Xiong. How important is the train-validation split in meta-learning? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 543–553. PMLR, 18–24 Jul 2021.

M.-F. Balcan, M. Khodak, and A. Talwalkar. Provable guarantees for gradient-based meta-learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 424–433. PMLR, 09–15 Jun 2019.

P. L. Bartlett and S. Mendelson. Empirical minimization. *Probability theory and related fields*, 135 (3):311–334, 2006.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

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

S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 567–580. Springer, 2003.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities*. Oxford University Press, 2013.

E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics. Springer, 2004.

O. Catoni. *PAC-Bayesian Supervised Classification: the Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics Lecture Notes – Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007.

J. Chen, X.-M. Wu, Y. Li, Q. Li, L.-M. Zhan, and F.-L. Chung. A closer look at the training strategy for modern meta-learning. *Advances in Neural Information Processing Systems*, 33:396–406, 2020.

L. Chen and T. Chen. Is Bayesian model-agnostic meta learning better than model-agnostic meta learning, provably? In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 1733–1774. PMLR, 28–30 Mar 2022.

L. Chen, S. T. Jose, I. Nikoloska, S. Park, T. Chen, and O. Simeone. Learning with limited samples: Meta-learning and applications to communication systems. *Foundations and Trends® in Signal Processing*, 17(2):79–208, 2023.

Q. Chen, C. Shui, and M. Marchand. Generalization bounds for meta-learning: An information-theoretic analysis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25878–25890. Curran Associates, Inc., 2021.

T. Cover and J. Thomas. *Elements of Information Theory*. Wileys Series in Communication, 2006.

A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

G. Denevi, C. Ciliberto, D. Stamos, and M. Pontil. Learning to learn around a common mean. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018a.

G. Denevi, C. Ciliberto, D. Stamos, and M. Pontil. Incremental learning-to-learn with statistical guarantees. *arXiv preprint arXiv:1803.08089*, 2018b.

G. Denevi, C. Ciliberto, R. Grazzi, and M. Pontil. Learning-to-learn stochastic gradient descent with biased regularization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1566–1575. PMLR, 09–15 Jun 2019a.

G. Denevi, D. Stamos, C. Ciliberto, and M. Pontil. Online-within-online meta-learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b.

G. Denevi, M. Pontil, and C. Ciliberto. The advantage of conditional meta-learning for biased regularization and fine tuning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 964–974. Curran Associates, Inc., 2020.

G. Denevi, M. Pontil, and C. Ciliberto. Conditional meta-learning of linear representations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 253–266. Curran Associates, Inc., 2022.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 1996.

N. Ding, X. Chen, T. Levinboim, S. Goodman, and R. Soricut. Bridging the gap between practice and PAC-Bayes theory in few-shot meta-learning. *Advances in Neural Information Processing Systems*, 34:29506–29516, 2021.

M. D. Donsker and S. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. iii. *Communications on Pure and Applied Mathematics*, 28:389–461, 1976.

Z.-Y. Dou, K. Yu, and A. Anastasopoulos. Investigating meta-learning algorithms for low-resource natural language understanding tasks. *arXiv preprint arXiv:1908.10423*, 2019.

A. Doucet and A. Lee. Sequential Monte Carlo methods. In *Handbook of graphical models*, pages 165–188. CRC Press, 2018.

A. Doucet, N. De Freitas, and N. J. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

A. Fallah, A. Mokhtari, and A. Ozdaglar. Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 5469–5480. Curran Associates, Inc., 2021.

A. Farid and A. Majumdar. Generalization bounds for meta-learning via PAC-Bayes and uniform stability. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2173–2186. Curran Associates, Inc., 2021.

C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 06–11 Aug 2017.

C. Finn, A. Rajeswaran, S. Kakade, and S. Levine. Online meta-learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1920–1930. PMLR, 09–15 Jun 2019.

P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 1884–1892. Curran Associates, Inc., 2016.

S. Ghosal and A. Van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 2017.

S. Gidaris and N. Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4367–4375, 2018.

E. Grant, C. Finn, S. Levine, T. Darrell, and T. Griffiths. Recasting gradient-based meta-learning as hierarchical Bayes. *arXiv preprint arXiv:1801.08930*, 2018.

P. D. Grünwald and N. A. Mehta. Fast rates for general unbounded loss functions: from ERM to generalized Bayes. *The Journal of Machine Learning Research*, 21(1):2040–2119, 2020.

J. Gu, Y. Wang, Y. Chen, K. Cho, and V. Li. Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*, 2018.

J. Guan and Z. Lu. Fast-rate PAC-Bayesian generalization bounds for meta-learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7930–7948. PMLR, 17–23 Jul 2022a.

J. Guan and Z. Lu. Task relatedness-based generalization bounds for meta learning. In *International Conference on Learning Representations*, 2022b.

J. Guan, Y. Liu, and Z. Lu. Fine-grained analysis of stability and generalization for modern meta learning algorithms. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 18487–18500. Curran Associates, Inc., 2022.

B. Guedj. A primer on PAC-Bayesian learning. In *Proceedings of the second congress of the French Mathematical Society*, 2019.

M. Haghifam, G. K. Dziugaite, S. Moran, and D. Roy. Towards a unified information-theoretic framework for generalization. *Advances in Neural Information Processing Systems*, 34:26370–26381, 2021.

F. Hellström and G. Durisi. Evaluated CMI bounds for meta learning: Tightness and expressiveness. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 20648–20660. Curran Associates, Inc., 2022.

S. Hochreiter, A. S. Younger, and P. R. Conwell. Learning to learn using gradient descent. In *Artificial Neural Networks—ICANN 2001: International Conference Vienna, Austria, August 21–25, 2001 Proceedings 11*, pages 87–94. Springer, 2001.

P.-S. Huang, C. Wang, R. Singh, W.-T. Yih, and X. He. Natural language to structured query generation via meta-learning. *arXiv preprint arXiv:1803.02400*, 2018.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

S. M. Jose, O. Simeone, and G. Durisi. Transfer meta-learning: Information-theoretic bounds and information meta-risk minimization. *IEEE Transactions on Information Theory*, 68(1):474–501, 2021.

S. T. Jose and O. Simeone. Information-theoretic generalization bounds for meta-learning and applications. *Entropy*, 23(1):126, 2021.

M. Khodak, M.-F. Balcan, and A. S. Talwalkar. Adaptive gradient-based meta-learning methods. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

J. Knoblauch, J. Jewson, and T. Damoulas. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *The Journal of Machine Learning Research*, 23(1):5789–5897, 2022.

G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR 2011*, pages 1785–1792. IEEE, 2011.

D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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

E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6): 1808–1829, 1999.

G. M. Martin, D. T. Frazier, and C. P. Robert. Computing Bayes: Bayesian computation from 1763 to the 21st century. *arXiv preprint arXiv:2004.06425*, 2020.

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

A. Maurer and T. Jaakkola. Algorithmic stability and meta-learning. *The Journal of Machine Learning Research*, 6(6), 2005.

A. Maurer, M. Pontil, and B. Romera-Paredes. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(81):1–32, 2016.

D. A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234, New York, 1998. ACM.

D. Meunier and P. Alquier. Meta-strategy for learning tuning parameters with guarantees. *Entropy*, 23(10), 2021.

T.P. Minka. A family of algorithms for approximate bayesian inference. *Ph. D. Thesis, Massachusetts Institute of Technology*, 2001.

N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel. A simple neural attentive meta-learner. *International Conference on Learning Representations*, 2018.

T. Munkhdalai and H. Yu. Meta networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2554–2563. PMLR, 06–11 Aug 2017.

C. Nguyen, T. T. Do, and G. Carneiro. Uncertainty in model-agnostic meta-learning using variational inference. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3090–3100, 2020.

A. Nichol, J. Achiam, and J. Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

A. Pentina and C. Lampert. A PAC-Bayesian bound for lifelong learning. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 991–999, Bejing, China, 22–24 Jun 2014. PMLR.

K. Qian and Z. Yu. Domain adaptive dialog generation via meta learning. *arXiv preprint arXiv:1906.03520*, 2019.

S. Qiao, C. Liu, W. Shen, and A. L. Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7229–7238, 2018.

A. Quattoni, M. Collins, and T. Darrell. Transfer learning for image classification with sparse prototype representations. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

C. Rajeswaran, A.and Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017.

A. Rezazadeh. A unified view on PAC-Bayes bounds for meta-learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18576–18595. PMLR, 17–23 Jul 2022.

A. Rezazadeh, S. T. Jose, G. Durisi, and O. Simeone. Conditional mutual information-based generalization bound for meta learning. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1176–1181. IEEE, 2021.

C. P. Robert. *The Bayesian Choice*. Springer, 2007.

C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 2004.

J. Rothfuss, V. Fortuin, M. Josifoski, and A. Krause. Pacoh: Bayes-optimal meta-learning with pac-guarantees. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9116–9126. PMLR, 18–24 Jul 2021.

J. Rothfuss, M. Josifoski, V. Fortuin, and A. Krause. PAC-Bayesian meta-learning: From theory to practice. *The Journal of Machine Learning Research*, 24(1):18474–18535, 2023.

D. Russo and J. Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. Meta-learning with memory-augmented neural networks. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1842–1850, New York, New York, USA, 20–22 Jun 2016. PMLR.

J. Shawe-Taylor and R. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 2–9, New York, 1997. ACM.

J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

S. Thrun and L. Pratt. *Learning to Learn: Introduction and Overview*. Kluwer Academic Publishers, 1998.

N. Tripuraneni, C. Jin, and M. Jordan. Provable meta-learning of linear representations. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10434–10443. PMLR, 18–24 Jul 2021.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

J. Vanschoren. *Automated Machine Learning: Methods, Systems, Challenges*. Springer International Publishing, 2019.

S. Vijaykumar. Localization, convexity, and star aggregation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4570–4581. Curran Associates, Inc., 2021.

O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

R. Wang, Y. Demiris, and C. Ciliberto. Structured prediction for conditional meta-learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2587–2598. Curran Associates, Inc., 2020.

A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 2521–2530, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

M. Yamada, W. Lian, A. Goyal, J. Chen, K. Wimalawarne, S. A. Khan, S. Kaski, H. Mamitsuka, and Y. Chang. Convex factorization machine for toxicogenomics prediction. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1215–1224, 2017.

J. Yang, S. Sun, and D. M. Roy. Fast-rate PAC-Bayes generalization bounds via shifted Rademacher processes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

J. Yoon, T. Kim, O. Dia, S. Kim, Y. Bengio, and S. Ahn. Bayesian model-agnostic meta-learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *Proceedings of the Conference on Robot Learning*, volume

100 of *Proceedings of Machine Learning Research*, pages 1094–1100. PMLR, 30 Oct–01 Nov 2020.

F. Zhang and C. Gao. Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180 – 2207, 2020.

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

Y. Zhang and Q. Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Qi. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.