Tighter PAC-Bayes Bounds Through Coin-Betting

Kyoungseok Jang[*]   KSAJKS@ARIZONA.EDU
University of Arizona

Kwang-Sung Jun   KJUN@CS.ARIZONA.EDU
University of Arizona

Ilja Kuzborskij   ILJAK@GOOGLE.COM
Google DeepMind

Francesco Orabona   FRANCESCO@ORABONA.COM
Boston University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We consider the problem of estimating the mean of a sequence of random elements $f(\theta, X_1), \ldots, f(\theta, X_n)$ where f is a fixed scalar function, $S = (X_1, \ldots, X_n)$ are independent random variables, and θ is a possibly S-dependent parameter. An example of such a problem would be to estimate the generalization error of a neural network trained on n examples where f is a loss function. Classically, this problem is approached through concentration inequalities holding uniformly over compact parameter sets of functions f, for example as in Rademacher or VC type analysis. However, in many problems, such inequalities often yield numerically vacuous estimates. Recently, the *PAC-Bayes* framework has been proposed as a better alternative for this class of problems for its ability to often give numerically non-vacuous bounds. In this paper, we show that we can do even better: we show how to refine the proof strategy of the PAC-Bayes bounds and achieve *even tighter* guarantees. Our approach is based on the *coin-betting* framework that derives the numerically tightest known time-uniform concentration inequalities from the regret guarantees of online gambling algorithms. In particular, we derive the first PAC-Bayes concentration inequality based on the coin-betting approach that holds simultaneously for all sample sizes. We demonstrate its tightness showing that by *relaxing* it we obtain a number of previous results in a closed form including Bernoulli-KL and empirical Bernstein inequalities. Finally, we propose an efficient algorithm to numerically calculate confidence sequences from our bound, which often generates nonvacuous confidence bounds even with one sample, unlike the state-of-the-art PAC-Bayes bounds.

Keywords: Concentration inequalities, PAC-Bayes, confidence sequences, coin-betting.

1. Introduction

Suppose that $S = (X_1, \ldots, X_n) \in \mathcal{X}$ are random elements distributed identically and independently from each other, on a probability space $(\mathcal{X}, \Sigma(\mathcal{X}), P_X)$. For illustration, assume that $\mathcal{X} = \mathbb{R}$: A classical problem in probability and statistics is to quantify how quickly an average $(X_1 + \cdots + X_n)/n$ converges to the mean $\mathbb{E}[X_1]$, and over the decades this problem was successfully attacked under various assumptions on the probability space through *concentration inequalities* (Boucheron et al., 2013). The key assumption which enables these concentration inequalities to exhibit fast convergence to the mean is *independence*. However, in many learning-theoretic problems we are interested in the concentration of random elements which themselves *depend on the*

---

[*] Authors are ordered alphabetically.

sample S, and therefore are not independent. In this paper, we formalize the above by assuming that we are given a *fixed* measurable function $f : \Theta \times \mathcal{X} \to [0, 1]$, where Θ is a *parameter* space, and so now we are interested in the concentration of $(f(\theta, X_1) + \cdots + f(\theta, X_n))/n$ around its mean, where θ is potentially S-dependent. For example, $f(\theta, X_i)$ could be the loss incurred by a learning algorithm on the i-th example, where the parameters θ are generated based on the sample S. In the context of this example, the mean $\mathbb{E}[f(\theta, X_1)]$ is called the statistical risk.

To this end, the classical approach to alleviating the dependence nuance is to derive *uniform* concentration inequalities that hold *simultaneously for all* parameters in a compact set Θ. For example, consider the following concentration inequality that holds with probability at least $1 - \delta$,[1]

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} f(\theta, X_i) - \mathbb{E}[f(\theta, X_1)] \right| \lesssim \sqrt{\frac{\text{capacity}(\Theta) + \ln \frac{1}{\delta}}{n}} .$$

Here the capacity term, such as VC dimension, metric entropy, or Rademacher complexity (Wainwright, 2019), scales with the "size" of the set Θ. Mentioned notions accurately capture the capacity in many learning problems, such as with linear parameterizations (Bartlett and Mendelson, 2002; Kakade et al., 2008). However, in some other problems, e.g., in learning with overparameterized neural networks, the pessimistic nature of uniform bounds makes them vacuous (Zhang et al., 2017).

On the other hand, in recent years, there has been a strong interest in the alternative to the uniform bounds, based on the *PAC-Bayes* analysis (Shawe-Taylor and Williamson, 1997; McAllester, 1998), which, remarkably, on some instances demonstrates *non-vacuous* bounds for the generalization ability of deep learning algorithms (Dziugaite and Roy, 2017; Pérez-Ortiz et al., 2021; Zhou et al., 2019). In the PAC-Bayes analysis instead of taking $\sup_{\theta \in \Theta}$ as in the uniform approach, we assume that the parameters θ are now *random* and follow a *data-dependent*, so-called, *posterior* distribution P_n. In this paper, we are interested in estimating an expected mean $\int \mu_\theta \, \mathrm{d}P_n(\theta)$ with $\mu_\theta := \int f(\theta, x) \, \mathrm{d}P_X(x)$, uniformly over all data-dependent posteriors P_n. This setting covers the one considered in the PAC-Bayes literature (Alquier, 2021) where usually the function f represents the loss of a predictor parameterized by θ from a parameter space Θ. In this view, we can think of μ_θ as the risk of a predictor parameterized by θ, while $\int \mu_\theta \, \mathrm{d}P_n(\theta)$ is the risk of a randomized predictor that uses a random θ drawn from the distribution P_n. The second important component of the PAC-Bayes model is a *prior* distribution P_0 which does not depend on S and captures our prior belief about the inductive bias in the problem instance. A basic PAC-Bayes concentration inequality (McAllester, 1998) then takes the form of

$$\left| \int \frac{1}{n} \sum_{i=1}^{n} f(\theta, X_i) \, \mathrm{d}P_n - \int \mu_\theta \, \mathrm{d}P_n(\theta) \right| \lesssim \sqrt{\frac{\mathsf{KL}(P_n \| P_0) + \ln \frac{1}{\delta}}{n}}$$

with probability at least $1 - \delta$, where instead of the capacity term we have a Kullback-Leibler (KL) divergence between the posterior and the prior. Numerically speaking, PAC-Bayes bounds tend to give much tighter bounds than their uniform counterparts, largely because the KL term is typically smaller than the capacity term (such as VC dimension), for an appropriate (user's) choice of P_n, P_0.

Our contributions: PAC-Bayes meets coin-betting   In this paper, we show that we can obtain *even tighter* PAC-Bayes bounds using recent advances in the theory of concentration inequalities through gambling algorithms (Orabona and Jun, 2021). In particular, we show that it is possible to obtain a new *coin-betting* based PAC-Bayes bound that directly *implies* a number of previous results.

---

1. The notation $\lesssim$ hides universal constants and logarithmic factors.

Moreover, numerically evaluating this new upper bound, we show that it is numerically tighter than all previous approaches. In Section 5 we present our main result, a concentration inequality of the following form, which holds with probability at least $1 - \delta$, *simultaneously for all $n \in \mathbb{N}$, all* data-dependent posteriors P_n, and all data-free priors P_0:

$$\int \max_{\lambda \in [-\frac{1}{1-\mu_\theta}, \frac{1}{\mu_\theta}]} \sum_{i=1}^{n} \ln\left(1 + \lambda(f(\theta, X_i) - \mu_\theta)\right) \mathrm{d}P_n(\theta) \leq \mathsf{KL}(P_n \| P_0) + \ln \frac{c\sqrt{n}}{\delta} \,. \qquad (1)$$

Moreover, the confidence interval for $\int \mu_\theta \, \mathrm{d}P_n(\theta)$ is obtained by solving the optimization problems

$$\left[\min_{h \in M} \int h(\theta) \, \mathrm{d}P_n(\theta), \ \max_{h \in M} \int h(\theta) \, \mathrm{d}P_n(\theta)\right], \qquad (2)$$

where M is a class of mean functions $(\mu_\theta)_{\theta \in \Theta}$ that satisfy the constraint given by Eq. (1), which is the first of its kind in the PAC-Bayes literature. We formalize these optimization problems in Section 5.1 and show that the constraint is *convex*, making them efficiently solvable in some cases. In Section 6 we also experimentally validate our approach.

In addition, we show that Eq. (1) is tighter than some well-known PAC-Bayes inequalities, such as McAllester's inequality (McAllester, 1998), Maurer's inequality for Bernoulli KL divergence (Maurer, 2004), and PAC-Bayes empirical Bernstein's inequality (Tolstikhin and Seldin, 2013). This is done by *relaxing* inequality in Eq. (1) by simple lower bounds of the logarithmic term. We show that even relaxing Eq. (1) leads to a tighter bound than Maurer's one that is known to be very tight numerically.

The observation above implies that from our result we can derive all these versions, obtain bounds, and then take an intersection of all of them *without having to split δ*. This is in stark contrast to empirical Bernstein's bounds (Tolstikhin and Seldin, 2013) that are often numerically looser than KL bounds, while being orderwise tighter than KL bounds like Maurer's one. Attempting to take an intersection with KL bounds requires splitting δ, which undesirably inflates the bound. This is not the case for our method – our result can be seen as "the right" type of concentration inequality that is superior to the rest up to constant factors inside an additive logarithmic term. Finally, Eq. (1) (and all its corollaries) holds simultaneously for all sample sizes, delivering time-uniform PAC-Bayes confidence sequences.

Organization of the paper   After a discussion of related work (Section 2) and notations (Section 3), in Section 4 we briefly present the idea behind concentration through coin-betting. In Section 5 we present our main results, discuss some implications, and include the proof of a new concentration inequality in Section 5.2. In Section 5.1, we discuss how to compute our concentration inequality numerically (without any relaxations). Finally, in Section 6 we present numerical simulations comparing our inequality to a number of baselines from the PAC-Bayes literature.

2. Related Work

Concentration from coin-betting   The coin-betting formalism considered here (see Section 4) goes back to Ville (1939) and Kelly betting system (Kelly, 1956) and has an intimate connection to the Universal Portfolio theory (Cover, 1991). Building on the ideas of Ville (1939), Shafer and Vovk (2001) introduced a general framework aiming at giving a foundation to the theory of probability rooted in gambling strategies. However, their framework is very general and it does not suggest specific methods to construct the betting strategies. The first paper to introduce the idea of using

the regret of online betting algorithms to produce new concentration inequalities was in Jun and Orabona (2019), which in turn builds on Rakhlin and Sridharan (2017) that showed the equivalence between the regret guarantees of generic online linear algorithms and martingale tail bounds.

PAC-Bayes Since the introduction of PAC-Bayes bounds by McAllester (1998), there has been significant growth and development in both theory and applications; see Alquier (2021) for a comprehensive survey. Early papers focused on tightening the bound of McAllester (1998), which can be seen as a PAC-Bayes version of Hoeffding's inequality. In particular, Langford and Caruana (2001); Seeger (2002); Maurer (2004) focused on the setting of a binary classification where the goal is to bound KL divergence between Bernoulli distributions. Such bounds are tighter than the mere difference of the risk and empirical risk due to Pinsker's inequality, and the numerically tightest known inequality within this group is Maurer's inequality (Maurer, 2004) (see for instance experiments of Mhammedi et al. (2019)). In this paper, we recover the result of Maurer (2004) by relaxing Eq. (1).

Towards data-dependent bounds, Tolstikhin and Seldin (2013) adapted an empirical Bernstein's inequality (Audibert et al., 2007; Maurer and Pontil, 2009) to the PAC-Bayes setting, making generalization bounds variance dependent. Once again, we recover the PAC-Bayes empirical Bernstein's inequality by relaxing our main result of Eq. (1) without any plug-in arguments, through a relatively straightforward proof. Several works went further in making bounds data-dependent by manipulating the KL term. Ambroladze et al. (2006) explored the idea of splitting the sample and deriving the *prior* from a held-out sample while obtaining the posterior from the remaining part. This technique proved very fruitful in making PAC-Bayes bounds much tighter. Indeed, recent non-vacuous generalization bounds for deep neural networks are largely attributed to this technique (Dziugaite and Roy, 2018; Pérez-Ortiz et al., 2021). Clearly, the results developed in this paper can be readily applied together with the splitting technique. The splitting technique was also investigated beyond the KL term. In particular, Mhammedi et al. (2019); Wu and Seldin (2022) developed intricate bounds akin to empirical Bernstein's inequalities where the splitting is done with respect to the sample variance (two variance terms) in addition to the KL term. These are among the numerically tightest known PAC-Bayes bounds. However, due to their highly problem-dependent nature, it is challenging to compare these bounds theoretically. Several other works, such as (Rivasplata et al., 2020; Awasthi et al., 2020) explored proof techniques (e.g. differential privacy and stability of resulting measures) allowing to derive PAC-Bayes bounds with data-dependent priors without splitting the sample.

The proof of our main result relies on showing that the exponential moment of the optimal log-wealth with respect to θ is a martingale. Several papers in PAC-Bayes literature have shown results exploiting (super-)martingale concentration, which allowed them to relax the independence assumption in the data sequence (Seldin et al., 2012) or to replace boundedness of $f()$ by weaker assumptions (Kuzborskij and Szepesvári, 2019; Haddouche and Guedj, 2022). To this end, Haddouche and Guedj (2022) exploited Ville's inequality (as in our proof), which allowed them to show a bound that holds uniformly over $n \in \mathbb{N}$.

Finally, it is known that solving the classical PAC-Bayes bound of McAllester (1998) for the posterior results in a *Gibbs* posterior $P_n(\theta) \propto e^{-\frac{1}{n} \sum_i f(\theta, X_i)} \, \mathrm{d}P_0(\theta)$. A large body of literature has looked at learning-theoretic properties of Gibbs predictors (Catoni, 2007; Alquier et al., 2016; Raginsky et al., 2017; Kuzborskij et al., 2019; Grünwald and Mehta, 2019). The concentration inequality we develop here (Eq. (1)) is of a very different shape compared to (McAllester, 1998),

though it can be easily relaxed to obtain it. As such, the Gibbs predictor might be a *suboptimal* solution to Eq. (1), and it is an interesting open problem to characterize such a solution.

3. Definitions

We denote by $(x)_+ = \max\{x, 0\}$. If P and Q are probability measures over Θ such that $P \ll Q$, the KL divergence between P and Q is defined as $\mathsf{KL}(P, Q) := \int P(\mathrm{d}x) \ln \frac{\mathrm{d}P}{\mathrm{d}Q}(x)$. With a slight abuse of notation, we also write $\mathsf{KL}(p\|q)$ where $p = \mathrm{d}P/\mathrm{d}\lambda$ and $q = \mathrm{d}Q/\mathrm{d}\lambda$ are densities of P and Q with respect to some common σ-finite measure λ. If a set $\mathcal{X}$ is uniquely equipped with a σ-algebra, the underlying σ-algebra will be denoted by $\Sigma(\mathcal{X})$. We formalize a "data-dependent distribution" through the notion of a *probability kernel* (see, e.g., Kallenberg, 2017) which is defined as a map $K : \mathcal{X}^n \times \Sigma(\Theta) \to [0, 1]$ such that for each $B \in \Sigma(\Theta)$ the function $s \mapsto K(s, B)$ is measurable and for each $s \in \mathcal{X}^n$ the function $B \mapsto K(s, B)$ is a probability measure over Θ. We write $\mathcal{K}(\mathcal{X}^n, \Theta)$ to denote the set of all probability kernels from $\mathcal{X}^n$ to distributions over Θ. In that light, when $P \in \mathcal{K}(\mathcal{X}^n, \Theta)$ is evaluated on $S \in \mathcal{X}^n$ we use the shorthand notation $P_n = P(S)$.

4. Warm-up: From Betting To Concentrations

In this section, we briefly explain how to obtain new concentration inequalities from betting algorithms, following Orabona and Pál (2016); Rakhlin and Sridharan (2017); Jun and Orabona (2019).

Let $c_t \in [-1, 1]$ be a sequence of "continuous coin" outcomes chosen arbitrarily. In each round, the bettor bets $|x_t|$ money on the outcome $\mathrm{SGN}(x_t)$. Then, c_t is revealed and the bettor wins/loses $x_t c_t$ money. Define the initial wealth $\mathrm{Wealth}_0 := 1$ and the wealth at the end of round t as

$$\mathrm{Wealth}_t := \mathrm{Wealth}_{t-1} + c_t x_t = 1 + \sum_{s=1}^{t} c_s x_s \ .$$

We also assume that the algorithm guarantees $\mathrm{Wealth}_t \geq 0$, hence we must have $x_t \in [-\mathrm{Wealth}_{t-1}, \mathrm{Wealth}_{t-1}]$. Given that no assumptions are made on how c_t is generated, this is essentially an online game (Cesa-Bianchi and Lugosi, 2006; Orabona, 2019). So, our aim is to achieve an amount of money close to the one of a fixed comparator. In particular, let $\mathrm{Wealth}_t(\lambda)$ be the wealth obtained by a bettor that bets $\lambda \mathrm{Wealth}_{t-1}(\lambda)$ in round t with initial wealth equal to $\mathrm{Wealth}_0(\lambda) := 1$ and

$$\mathrm{Wealth}_t(\lambda) := \mathrm{Wealth}_{t-1}(\lambda) + c_t \lambda \mathrm{Wealth}_{t-1}(\lambda) = \prod_{s=1}^{t} (1 + c_s \lambda) \ .$$

We can now formally define the regret of the betting algorithm as

$$\mathrm{Regret}_T := \frac{\max_{\lambda \in [-1,1]} \mathrm{Wealth}_T(\lambda)}{\mathrm{Wealth}_T} \ .$$

It is well-known that it is possible to design optimal online betting algorithms where the regret is polynomial in T (Cesa-Bianchi and Lugosi, 2006, Chapters 9 and 10).

Closed form concentration, following Rakhlin and Sridharan (2017) Here, we summarize the basic idea of Rakhlin and Sridharan (2017) used to obtain concentration inequalities from online learning algorithms, specializing it to online betting algorithms as in Jun and Orabona (2019).

Consider X_t to be a sequence of i.i.d. random variables supported on $[0, 1]$ such that $\mathbb{E}[X_t] = \mu$. Set $c_t = X_t - \mu \in [-1, 1]$, so that regardless of the online betting algorithm we have $\mathbb{E}[\mathrm{Wealth}_t] = 1$. Also, assume that $\mathrm{Regret}_T \leq R(T)$, where $R : \mathbb{N} \to \mathbb{R}_+$. Let's now lower bound $\mathrm{Wealth}_T(\lambda)$

to obtain a familiar quantity. Using the inequality $1 + x \geq \exp(x - x^2)$ for $x \geq -0.68$, we obtain

$$\max_{\lambda \in [-1,1]} \text{Wealth}_T(\lambda) \geq \max_{\lambda \in [-1/2,1/2]} \text{Wealth}_T(\lambda) = \max_{\lambda \in [-1/2,1/2]} \prod_{t=1}^{T}(1 + \lambda(X_t - \mu))$$

$$\geq \max_{\lambda \in [-1/2,1/2]} \exp\left(\lambda \sum_{t=1}^{T}(X_t - \mu - \lambda(X_t - \mu)^2)\right) \geq \max_{\lambda \in [-1/2,1/2]} \exp\left(\lambda \sum_{t=1}^{T}(X_t - \mu - \lambda)\right).$$

Putting it all together and using Markov's inequality, for any $\lambda \in [-1/2, 1/2]$ we get

$$\mathbb{P}\left\{\lambda \sum_{t=1}^{T}(X_t - \mu - \lambda) \geq \ln \frac{R(T)}{\delta}\right\} \leq \mathbb{P}\left\{\text{Wealth}_T(\lambda) \geq \frac{R(T)}{\delta}\right\} \leq \mathbb{P}\left\{\text{Wealth}_T \geq \frac{1}{\delta}\right\} \leq \delta.$$

Choosing λ with the proper sign and of the order of $\frac{1}{\sqrt{T}}$, we get roughly Hoeffding inequality when the Regret is $\mathcal{O}(1)$, which is possible for fixed T and for this specific lower bound to the optimal wealth. Even better concentrations can be obtained carrying around the $(X_t - \mu)^2$ terms, resulting in an empirical Bernstein-style bound.

It is important to stress that we do not need to run the betting algorithm to obtain the concentration. Instead, we only need the *existence* of a betting algorithm and its associated regret guarantee.

Tighter concentration inequalities From the above reasoning, it should be clear that we can obtain a tighter bound by giving up the closed-form expression by avoiding to lower bound the wealth:

$$\mathbb{P}\left\{\max_{\lambda \in [-1,1]} \sum_{t=1}^{T} \ln(1 + \lambda(X_t - \mu)) \geq \ln \frac{R(T)}{\delta}\right\} = \mathbb{P}\left\{\max_{\lambda \in [-1,1]} \text{Wealth}_T(\lambda) \geq \frac{R(T)}{\delta}\right\} \leq \delta.$$

In this case, we can numerically invert this inequality and obtain a tighter concentration.

Now, we depart from Rakhlin and Sridharan (2017) and, instead of using Markov's inequality, we follow Jun and Orabona (2019) using Ville's inequality (Theorem 10). We can do it because, by the assumptions on the betting algorithm, the wealth is a non-negative martingale. The use of Ville's inequality gives the uniformity over time for free and gives us a high-probability time-uniform concentration inequality. Namely, with probability at least $1 - \delta$, we have

$$\max_t \max_{\lambda \in [-1,1]} \sum_{s=1}^{t} \ln(1 + \lambda(X_s - \mu)) \leq \ln \frac{R(t)}{\delta}. \tag{3}$$

Note that to obtain upper and lower bounds for μ it is enough to find the set of values of μ that satisfies Eq. (3). This can be done efficiently because the argument of the max can be proved to be a quasi-convex one-dimensional function in μ (Orabona and Jun, 2021). The concentration inequality above can be seen as a tight and implicit version of the empirical Bernstein's inequality for bounded random variables, just like how the KL-divergence concentration inequality is an implicit and tight version of the Bernstein's inequality for Bernoulli random variables.

5. Main Results

The concentration inequality of Eq. (3) holds for i.i.d. random variables $S = (X_1, \ldots, X_n)$. However, in many learning-theoretic applications, we are interested in providing confidence intervals for the mean of some data-dependent function (such as the generalization error). To this end, in this sec-

tion, we explore a scenario where $X_1, \ldots, X_n$ are replaced by a sequence $f(\theta, X_1), \ldots, f(\theta, X_n)$ such that f is a fixed scalar function and θ is a *data-dependent* parameter. Clearly, elements of such a sequence are not independent, since dependence is introduced through parameter θ. Following the *PAC-Bayes* viewpoint (McAllester, 1998; Alquier, 2021), θ is now *random* and distributed according to some user-chosen data-dependent distribution P_n called *posterior*. In addition, unlike in the traditional PAC-Bayes literature, our inequalities hold uniformly *not only* in P_n, but also in the *sample size* n. Thus, we construct a high-probability *PAC-Bayes confidence sequence*.

The next theorem, proved in Section 5.2, is the main result of our paper, which generalizes the concentration analysis of Section 4 to the PAC-Bayes setting.

Theorem 1 *Let* $S = (X_1, \ldots, X_n)$ *be a tuple of i.i.d. random variables taking values in some measurable space* $\mathcal{X}$. *Let* P_n *be a data-dependent distribution over some measurable space* Θ *and let* P_0 *be any probability measure over* Θ *independent from sample* S. *Let* $f : \Theta \times \mathcal{X} \to [0, 1]$ *be any fixed measurable function, let its mean be denoted by* $\mu_\theta = \int f(\theta, x) \, \mathrm{d}P_X(x)$, *and introduce*

$$\psi_n^\star(\theta, M) := \max_{\lambda \in [-\frac{1}{1-M}, \frac{1}{M}]} \sum_{i=1}^n \ln \left(1 + \lambda(f(\theta, X_i) - M) \right), \qquad (\theta \in \Theta, M \in [0, 1]) .$$

Then, for all P_0 *and for any* $\delta \in (0, 1]$, *we have*[2]

$$\mathbb{P} \left\{ \exists n \in \mathbb{N}, \exists P_n : \int \psi_n^\star(\theta, \mu_\theta) \, \mathrm{d}P_n(\theta) - \mathsf{KL}(P_n \| P_0) - \ln \frac{\sqrt{\pi} \, \Gamma(n+1)}{\Gamma(n+\frac{1}{2})} \geq \ln \frac{1}{\delta} \right\} \leq \delta . \quad (4)$$

Now we discuss some of the implications of Eq. (4) and compare it to existing PAC-Bayes results. The important feature of Eq. (4) is that it holds simultaneously for all posterior distributions, so we can freely choose the one that depends on the data. At the same time, Eq. (4) is similar in shape to the concentration inequality of Eq. (3). In particular, $\psi_n^\star$ is an optimal log-wealth discussed in Section 4, while $\Gamma(n+1)/\Gamma(n+1/2) \sim \sqrt{n}$ is the regret bound ($R(n)$) of a certain betting algorithm. Observe that unlike Eq. (3), the left-hand side of the inequality is now integrated over $\theta \sim P_n$, and the term $\mathsf{KL}(P_n \| P_0)$ appears on the right-hand side. In particular, the term $\mathsf{KL}(P_n \| P_0)$ captures the capacity of the class of posterior distributions with respect to the prior P_0, and it is a standard component in PAC-Bayes analyses.

Obtaining known PAC-Bayes inequalities by relaxing Eq. (4) By *relaxing* Eq. (4), we demonstrate that Theorem 1 gives a tighter concentration inequality compared to some inequalities in PAC-Bayes literature (proofs are deferred to Section A). Importantly, our results extend these bounds as our relaxations hold uniformly over $n \in \mathbb{N}$, whereas previous results hold for a fixed n. Abbreviate

$$\hat{\mu}_\theta := \frac{1}{n} \sum_{i=1}^n f(\theta, X_i) \qquad \text{and} \qquad \mathcal{C}_n := \mathsf{KL}(P_n \| P_0) + \ln \frac{\sqrt{\pi} \, \Gamma(n+1)}{\Gamma(n+\frac{1}{2})} .$$

As a basic sanity-check, we first recover a classical result of McAllester (1998) through the elementary inequality $\ln(1+x) \geq x - x^2$ for $x \geq -0.68$ (similarly as in Section 4), proof in Section A.1.

Corollary 2 (McAllester's inequality) *Set* $\delta \in (0, 1]$. *Under conditions of Theorem 1, for all priors* P_0, *with probability at least* $1-\delta$ *over the sample* S, *for all* $n \in \mathbb{N}$ *and for all data-dependent*

---

2. Here $\exists n \in \mathbb{N}, \exists P_n$ is a shorthand notation for $\exists n \in \mathbb{N}, \exists P \in \mathcal{K}(\mathcal{X}^n, \Theta)$ where $\mathcal{K}()$ is a set of probability kernels as defined in Section 3.

distributions P_n simultaneously we have

$$\left| \int \mu_\theta \, \mathrm{d}P_n(\theta) - \int \hat{\mu}_\theta \, \mathrm{d}P_n(\theta) \right| \le 2\sqrt{\frac{\mathcal{C}_n + \ln \frac{1}{\delta}}{n}} \,.$$

Note that, up to constants, the above matches the result of McAllester (1998), and *extends* it — now the bound holds *simultaneously* for all $n \in \mathbb{N}$.

Now we turn our attention to a type of PAC-Bayes inequality, where we the bound is given on a KL divergence between Bernoulli distributions. Such bounds are useful in a setting of a binary classification, where the parameter of a Bernoulli distribution models a conditional probability of a positive class label. In particular, relaxing Eq. (4) gets a well-known inequality of Maurer (2004):

Corollary 3 (Maurer's inequality) *For $p, q \in [0, 1]$ let $\mathsf{kl}(p, q) := p \ln(p/q) - (1 - p) \ln((1 - p)/(1 - q))$, i.e., the KL divergence between Bernoulli distributions with parameters p and q respectively. Set $\delta \in (0, 1]$. Under the conditions of Theorem 1, for all priors P_0, with probability at least $1 - \delta$ over the sample S, for all $n \in \mathbb{N}$ and for all data-dependent distributions P_n simultaneously,*

$$\mathsf{kl}\Big(\int \hat{\mu}_\theta \, \mathrm{d}P_n(\theta), \int \mu_\theta \, \mathrm{d}P_n(\theta) \Big) \le \frac{\mathcal{C}_n + \ln \frac{1}{\delta}}{n} \,.$$

The above inequality matches Maurer's bound up to a constant inside a logarithmic factor. Furthermore, the proof of Corollary 3 in Section A.2 reveals that even relaxing Theorem 1 to have $\int \mathsf{kl}(\hat{\mu}, \mu_\theta) \, \mathrm{d}P_n(\theta)$ in place of $\psi_n^\star(\theta, \mu_\theta)$ on the LHS results in a bound that is tighter than Maurer's inequality. We confirm this numerically in Section 6.

We now consider a more sophisticated, sample variance-dependent concentration inequality, which exhibits a faster rate of order $1/n$ whenever the sample variance is sufficiently small. In the non-PAC Bayes form, such a *empirical Bernstein's inequality* was shown by Audibert et al. (2007); Maurer and Pontil (2009), whereas the PAC-Bayes version was first presented by Tolstikhin and Seldin (2013). The following result recovers their result up to constants through a much simpler proof by relaxing Theorem 1:

Corollary 4 (PAC-Bayes empirical Bernstein's inequality) *Set $\delta \in (0, 1]$. Introduce*

$$\hat{V}(P_n) := \frac{1}{n} \int \sum_{i=1}^{n} (f(\theta, X_i) - \hat{\mu}_\theta)^2 \, \mathrm{d}P_n(\theta) \,, \qquad \mathcal{C}_{n,\delta} := \mathcal{C}_n + \ln \frac{1}{\delta} \,.$$

Under the conditions of Theorem 1, for all priors P_0, with probability at least $1 - \delta$ over the sample S, for all $n \in \mathbb{N}$ and for all data-dependent distributions P_n simultaneously we have

$$\left| \int \mu_\theta \, \mathrm{d}P_n(\theta) - \int \hat{\mu}_\theta \, \mathrm{d}P_n(\theta) \right| \le \frac{\sqrt{2\mathcal{C}_{n,\delta} \hat{V}(P_n)}}{\left(\sqrt{n} - \frac{2}{\sqrt{n}} \mathcal{C}_{n,\delta} \right)_+} + \frac{2\mathcal{C}_{n,\delta}}{\left(n - 2\mathcal{C}_{n,\delta} \right)_+} \,.$$

Note that the inequality is fully empirical and non-vacuous as long as $\mathcal{C}_{n,\delta} \le n/2$ — similar (empirically verifiable) requirement is also present in (Tolstikhin and Seldin, 2013, Theorem 4). Clearly, the fact that we relaxed Theorem 1 to get Corollary 4 implies that our inequality is tighter.

Finally we show that our inequality is tighter than the 'Un-expected' Bernstein's inequality derived by Mhammedi et al. (2019). In particular, they presented a rather intricate data-dependent inequality for the purpose of numerical evaluation. However, at the core of their proof lies the following exponential moment inequality (Mhammedi et al., 2019, Lemma 13), which states that

for a random variable $Z \in [0, 1]$ and some fixed $\lambda \in [0, 1]$,

$$\mathbb{E} \exp \left(\lambda(\mathbb{E}[Z] - Z) - \alpha(\lambda)(\lambda Z)^2 \right) \leq 1 , \tag{5}$$

where $\alpha(\lambda) := (-\ln(1 - \lambda) - \lambda)/\lambda^2$. Here, the key observation is that Z^2 appears outside of the expectation compared to the standard Bernstein's inequality, which might make concentration inequalities derived from Eq. (5) numerically tighter. Next, we present essentially a corollary of Theorem 1, shown in Section A.4, which is obtained by relaxing our results:

Corollary 5 ('Un-expected' PAC-Bayes Bernstein's inequality) *Let* $\Delta_i(\theta) = f(\theta, X_i) - \mu_\theta$. *Then,*

$$\mathbb{E} \exp \left(\sup_{P_n} \int \max_{\lambda \in [0,1]} \sum_{i=1}^n \left(-\lambda \Delta_i(\theta) - \alpha(\lambda)(\lambda \Delta_i(\theta))^2 \right) \mathrm{d} P_n(\theta) - \mathcal{C}_n \right) \leq 1 .$$

Note that Corollary 5 is strictly more general compared Eq. (5) as it is a PAC-Bayes result and involves a KL term within $\mathcal{C}_n$. However, we can compare this inequality to Eq. (5) in a standard concentration setting (choose P_n, P_0 to have $\mathsf{KL}(P_n, P_0) = 0$). In such a case, $\mathcal{C}_n \leq \ln(n)/2$ (due to incurred regret) and so Corollary 5 matches Eq. (5) up to a logarithmic term.

5.1. How to compute confidence intervals from Eq. (4) numerically

So far we discussed analytically computable relaxations of our inequality. Now we turn our attention to numerical computation of Eq. (4) which does not require any relaxation. Given a concrete posterior and prior pair (P_n, P_0), we propose to obtain confidence bounds for the mean $\int \mu_\theta \, \mathrm{d} P_n(\theta)$ by solving the following optimization problem:

Proposition 6 *Set* $\delta \in (0, 1]$. *Consider the optimization problem*

$$M_U = \max_{h \in \mathcal{H}} \int h(\theta) \, \mathrm{d} P_n(\theta) \quad \text{subject to} \quad \int \psi_n^\star(\theta, h(\theta)) \, \mathrm{d} P_n(\theta) \leq \mathcal{C}_n + \ln \frac{1}{\delta} , \tag{6}$$

where $\mathcal{H}$ *is the class of mean functions over* $\mathrm{supp}(P_n)$ *and* $\mathcal{C}_n + \ln \frac{1}{\delta}$ *is the right hand side of Eq. (4). Moreover, assume that* M_L *be obtained by replacing* max *with* min. *Then, under the conditions of Theorem 1 and with probability at least* $1 - \delta$, *we have*

$$M_L \leq \int \mu_\theta \, \mathrm{d} P_n(\theta) \leq M_U .$$

In other words, the optimization in Eq. (6) is carried out over the class of means of a given distribution, and the solution gives us a valid confidence interval since Theorem 1 holds for any data-dependent posterior, and so it must hold for some posteriors with means within the class. Moreover, surprisingly enough, the optimization problem is *convex* since $\psi_n^\star(\theta, M)$ appearing in the constraint is convex in M, thanks to the following lemma proven in Section A.5.

Lemma 7 (Convexity of the constraint) *Let* $c \in [0, 1]$. *Define* $f(x) = \max_{-\frac{1}{1-x} \leq \lambda \leq \frac{1}{x}} \ln(1 + \lambda(c - x))$. *Then,* $f(x)$ *is convex for* $x \in [0, 1]$.

In Section 6 we present synthetic experiments validating the numerical tightness of the confidence intervals obtained by solving the problem in Proposition 6.

5.1.1. MONTE CARLO APPROXIMATION OF THE INTEGRAL

The confidence intervals of Proposition 6 can be obtained efficiently as long as we can efficiently compute or estimate integrals over parameters. When Θ is finite we can clearly replace integrals by summations. On the other hand, for continuous (or prohibitively large finite) Θ we can employ a Monte Carlo approximation of the integral. In particular, we can use the procedure in Algorithm 1.

Algorithm 1 Monte Carlo Approximation

1: **Input**: Failure probability δ, sample size parameters $K, m \in \mathbb{N}$, posterior P_n, prior P_0
2: Sample K tuples independently $(\theta_i)_{i \in B_k} \sim P_n^m$ for $k \in [K]$ where
   $B_k = \{(k-1)m+1, \ldots, km\}$
3: Solve the following optimization problem for every $k \in [K]$:

$$\bar{\nu}_U(k) := \max_{\{\nu_{\theta_i}\}_{i \in B_k}} \frac{1}{m} \sum_{i \in B_k} \nu_{\theta_i} \quad \text{s.t.} \quad \frac{1}{em} \sum_{i \in B_k} \psi_n^\star(\theta_i, \nu_{\theta_i}) \leq \mathcal{C}_n + \ln \frac{1}{\delta} \qquad (7)$$

4: Repeat the above while replacing $\max$ with $\min$. Call the resulting optimal objective function as $\bar{\nu}_L(k), \forall k \in [K]$
5: Let $\hat{k}_U = \arg\max_{k \in [K]} \bar{\nu}_U(k)$ and $\hat{k}_L = \arg\min_{k \in [K]} \bar{\nu}_L(k)$
6: Compute

$$M_U = \max \left\{ \mu : \mathsf{kl}\left(\bar{\nu}_U(\hat{k}_U), \mu\right) \leq \frac{\ln \frac{K}{2\delta}}{m} \right\}, \ M_L = \min \left\{ \mu : \mathsf{kl}\left(\bar{\nu}_L(\hat{k}_L), \mu\right) \leq \frac{\ln \frac{K}{2\delta}}{m} \right\} \qquad (8)$$

7: **Output:** M_L and M_U

The following corollary of Theorem 1 (proved in Section A.7) states its correctness.

Corollary 8 *Set $\delta \in (0, 1]$. Under the assumptions of Theorem 1, let $K = \lceil \ln(1/\delta) \rceil$. Then, with probability at least $1 - 3\delta$, the outputs M_L and M_U of Algorithm 1 satisfy*

$$M_L \leq \int \mu_\theta \, \mathrm{d}P_n(\theta) \leq M_U \ .$$

Algorithm 1 works by carefully controlling the Monte Carlo approximation through the deviation of the sample average over parameters from the integral. In fact, while this is straightforward for bounded random variables, $\psi_n^\star$ considered here is not bounded. One may attempt to make it bounded by clipping $\psi_n^*(\theta_i, \mu_{\theta_i})$ or reducing the range of λ in the max operator in the definition of ψ_n^*, but these both lead to nonconvex constraints in Eq. (7). Alternatively, since Eq. (6) suggests that we need to lower bound an integral, we could right away get a "low-probability" bound arising from Markov's inequality: $\frac{1}{\delta} \int \psi_n^\star(\theta, \mu_\theta) \, \mathrm{d}P_n(\theta) \geq \frac{1}{m} \sum_{i \in B_k} \psi_n^\star(\theta_i, \mu_i)$. Since this is unsatisfactory, here we resort to the "boosting-the-confidence" method (Schapire, 1990; Shalev-Shwartz et al., 2010) which allows to convert polynomial concentration bounds into exponential ones at the expense of sample partitioning and running the algorithm multiple times, as described in Corollary 8. Note that in our case this just translates into extra computation (running Monte Carlo approximation on K independent parameter tuples), because we can always sample more parameter observations from P_n. Corollary 8 is then justified through the use of the following inequality shown in Section A.6:

Proposition 9 *Under conditions of Corollary 8, with probability at least $1 - e^{-K}$,*

$$\int \psi_n^\star(\theta, \mu_\theta) \, \mathrm{d}P_n(\theta) \geq \min_{k \in [K]} \frac{1}{em} \sum_{i \in B_k} \psi_n^\star(\theta_i, \mu_{\theta_i}) \, . \tag{9}$$

How large m needs to be?   One question not discussed so far is the choice of Monte Carlo sample size m. Technically, Proposition 9 holds for any $m \in \mathbb{N}$, but we can expect that choosing small m will result in overly loose constraints in Eq. (9) and so the final confidence intervals will be wide. To gauge a good choice of m we consider a lower tail Bernstein's inequality (Maurer, 2003), which lower bounds the left-hand side of the constraint Eq. (9):

$$\frac{1}{m} \sum_{i=1}^m \psi_n(\theta_i, \mu_{\theta_i}) \geq \left(\int \psi_n^\star(\theta, \mu_\theta) \, \mathrm{d}P_n(\theta) - \sqrt{\frac{2 \ln \frac{1}{\delta}}{m} \int \psi_n^\star(\theta, \mu_\theta)^2 \, \mathrm{d}P_n(\theta)} \right)_+ \, . \tag{10}$$

Thus, having a raw second moment of $\psi_n^\star$ of order $o(m)$ guarantees asymptotic convergence of the sample average to the integral in the constraint (9). Having a finite raw second moment suggests that the constraint is tight and a reasonable choice, is, for instance, $m = n^2$. On the other hand, for "hard" problems (e.g., heavy-tailed) such moment is infinite and Monte Carlo estimation is infeasible.

5.2. Proof of Theorem 1

Let $\Delta_i(\theta) := f(\theta, X_i) - \mathbb{E}[f(\theta, X_1)]$ and notably $\mathbb{E}[\Delta_i(\theta)] = 0$ for any $\theta \in \Theta$. Consider an algorithm betting a signed fraction of its wealth equal to $B_i(\theta)$ at step i and observing the outcome $\Delta_i(\theta)$. Note that $B_i(\theta)$ is $\Sigma(X_1, \ldots, X_{i-1}, \theta)$-measurable. Let the following be the cumulative loss (log-wealth) of the algorithm and the optimal cumulative loss, respectively

$$\psi_n(\theta) := \sum_{i=1}^n \ln(1 + B_i(\theta) \Delta_i(\theta)) \, , \qquad \psi_n^\star(\theta) := \max_{\lambda \in \left[-\frac{1}{1-\mu_\theta}, \frac{1}{\mu_\theta} \right]} \sum_{i=1}^n \ln \left(1 + \lambda \Delta_i(\theta) \right) \, .$$

We are interested in showing an upper bound on $\int \psi_n^\star(\theta) \, \mathrm{d}P_n(\theta)$ which holds for all data-dependent distribution P_n simultaneously, and with high probability over the data. To this end, Orabona and Jun (2021) show that, for any $\theta \in \Theta$, there exists a betting algorithm which guarantees that[3]

$$\psi_n^\star(\theta) - \psi_n(\theta) \leq \ln \frac{\sqrt{\pi} \Gamma(n+1)}{\Gamma(n + \frac{1}{2})} \, . \tag{11}$$

So it remains to give a bound on $\psi_n(P_n)$. We will need the following concentration inequality.

Theorem 10 (Ville's inequality (Ville, 1939, p. 84)) *Let $\Delta_1, \ldots, \Delta_n$ be a sequence of non-negative random variables such that $\mathbb{E}[\Delta_i \mid \Delta_1, \ldots, \Delta_{i-1}] = 0$. Let $M_t > 0$ be $\Sigma(\Delta_1, \ldots, \Delta_{t-1})$-measurable such that $M_0 = 1$, and moreover assume that $\mathbb{E}[M_t \mid \Delta_1, \ldots, \Delta_{t-1}] \leq M_{t-1}$. Then, for any $\delta \in (0, 1]$, $\mathbb{P}\left\{ \max_t M_t \geq \frac{1}{\delta} \right\} \leq \delta$.*

The proof will also require the following well-known *change-of-measure* inequality:

Lemma 11 (Donsker and Varadhan (1975); Dupuis and R. S. Ellis (1997)) *Let p and q be probability measures on Θ such that $p \ll q$. Then, for any measurable function $f : \Theta \to \mathbb{R}$, we have*

$$\int f(\theta) \, \mathrm{d}p(\theta) \leq \mathsf{KL}(p\|q) + \ln \int e^{f(\theta)} \, \mathrm{d}q(\theta) \, .$$

---

3. Data-dependent bounds on the regret were also shown by Orabona and Jun (2021).

Applying the above with $p = P_n$, $q = P_0$, $f = \psi_n$, and taking $\max_{n \in \mathbb{N}}$, we obtain

$$\max_{n \in \mathbb{N}} \sup_{P_n} \int \psi_n(\theta) \, \mathrm{d}P_n(\theta) - \mathsf{KL}(P_n \| P_0) \leq \ln \max_{n \in \mathbb{N}} \underbrace{\int \prod_{i=1}^{n} \big(1 + B_i(\theta)\Delta_i(\theta)\big) \, \mathrm{d}P_0(\theta)}_{M_n}, \qquad (12)$$

where $\sup_{P_n} = \sup_{P_n \in \mathcal{K}(\mathcal{X}^n, \Theta)}$ and where we exchanged $\ln$ and $\max_{n \in \mathbb{N}}$. Now, the plan is to apply Theorem 10 to M_n, which requires to show that M_n is a martingale. Using the notation $\mathbb{E}_i[\cdot] = \mathbb{E}[\cdot \mid \Delta_1(\theta), \dots, \Delta_i(\theta), \theta]$, we have

$$\mathbb{E}_{n-1}[M_n] = \mathbb{E}_{n-1} \int \prod_{i=1}^{n} \big(1 + B_i(\theta)\Delta_i(\theta)\big) \, \mathrm{d}P_0(\theta) \overset{(a)}{=} \int \mathbb{E}_{n-1} \prod_{i=1}^{n} \big(1 + B_i(\theta)\Delta_i(\theta)\big) \, \mathrm{d}P_0(\theta)$$

$$= \int \mathbb{E}_{n-1} \left[\big(1 + B_n(\theta)\Delta_n(\theta)\big) \right] \prod_{i=1}^{n-1} \big(1 + B_i(\theta)\Delta_i(\theta)\big) \, \mathrm{d}P_0(\theta)$$

$$= \int \prod_{i=1}^{n-1} \big(1 + B_i(\theta)\Delta_i(\theta)\big) \, \mathrm{d}P_0(\theta) = M_{n-1},$$

where (a) comes using the fact that P_0 is independent from the sample S and by Fubini's theorem. Thus, applying Theorem 10 to Eq. (12), we obtain

$$\mathbb{P}\left\{ \max_{n \in \mathbb{N}} \sup_{P_n \in \mathcal{K}(\mathcal{X}^n, \Theta)} \int \psi_n(\theta) \, \mathrm{d}P_n(\theta) - \mathsf{KL}(P_n \| P_0) \leq \ln \frac{1}{\delta} \right\} \geq 1 - \delta \qquad (\delta \in (0, 1]) \,.$$

Finally, using Eq. (11) gives the statement and completes the proof. ∎

6. Experiments

In this section, we validate the numerical tightness of Theorem 1. Additional experiments are in Section B.[4] We perform experiments on simple synthetic scenarios where the parameter space is finite, and we fix the posterior and prior distributions. We evaluated all the bounds on a sample size range $n \in \{2^c : c = 1, \dots, 15\}$, and we averaged the bound over 20 repetitions for each sample size. In particular, we compare Corollary 8 to several PAC-Bayes baselines such as McAllester (1998), London and Sandler (2019), Maurer (2004) and Tolstikhin and Seldin (2013), and one additional algorithm KL-ver under several synthetic environments.

KL-ver, the KL version of our algorithm, uses $n \cdot \mathsf{KL}(\hat{\mu}_\theta, \mu_\theta)$ for the optimization problem in Eq. (6), instead of $\psi_n^*(\theta, \mu_\theta)$. Theoretically, the log-wealth function $\psi_n^\star(\theta, \mu_\theta)$ is always greater than $n \cdot \mathsf{KL}(\hat{\mu}_\theta, \mu_\theta)$ by Proposition 3. On the other hand, Maurer's bound is looser than KL-ver, which is shown in the proof of Proposition 3 in Appendix A.2. Hence, KL-ver is an ablation study on our novel optimization problem – KL-ver is looser than our proposed method of Eq. (6) but tighter than Maurer's bound that is known to be very tight numerically.

The first experiment, reported in Figure 1, represents the case where $X_i \sim \text{Bernoulli}(1/2)$ i.i.d., $P_0 = \text{Bernoulli}(0.8)$, $P_n = \text{Bernoulli}(0.9)$, and $f(\theta, x) = x\theta$. The second experiment, in Figure 2, represents the case where $X_i \sim N(0, 1)$ i.i.d., $P_0 = (\text{Bin}(6, 0.7) - 3)/4$, $P_n = (\text{Bin}(6, 0.8) - 3)/4$, and $f(\theta, x) = (\text{erf}(x\theta) + 1)/2$. Here, erf is the Gaussian error function and $\text{Bin}(n, p)$ is a binomial distribution, with $n \in \mathbb{N}$ being the number of samples, and $p \in [0, 1]$ the probability of success.

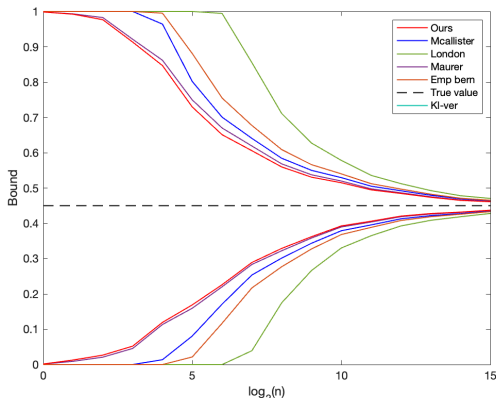4. Code can be found at: https://github.com/jajajang/coin_betting_pac_bayes
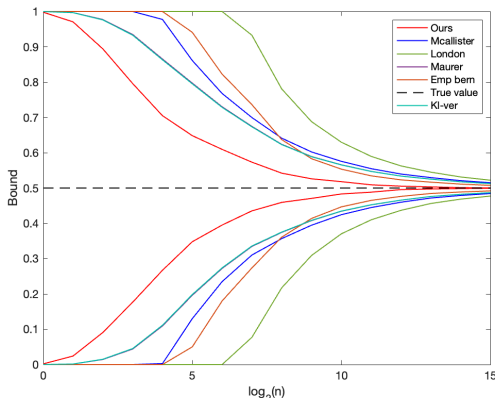
Figure 1: Bernoulli case.

Figure 2: Binomial case.

Since Θ is finite, we can explicitly calculate the means without resorting to the Monte Carlo simulation. Hence, the optimization problem of Proposition 6 reduces to (and similarly for the lower bound by replacing $\max$ with $\min$)

$$\max_{\{\mu_\theta:\theta\in\mathrm{supp}(P_n)\}} \sum_{\theta\in\Theta} \mu_\theta \cdot P_n(\theta) \quad \text{subject to} \quad \sum_{\theta\in\Theta} \psi_n^\star(\theta,\mu_\theta) \cdot P_n(\theta) \leq \mathcal{C}_n + \ln\frac{1}{\delta} . \tag{13}$$

This is a convex optimization problem as we showed before, so we can use any off-the-shelf solver.[5]

Both figures show that our confidence intervals are consistently tighter than the ones of the baselines. Moreover, our guarantee and the one of KL-ver hold uniformly over time, while it holds for a fixed number of samples for the baselines. Furthermore, for the Bernoulli case, KL-ver is the same as our bound and still better than Maurer's, and in the Binomial case, it is worse than ours and very close to Maurer's bound. This confirms our theoretical finding that our approach is "two-inequalities away" from Maurer's bound. For the case of continuous Θ, check Appendix B.

7. Conclusions, limitations, and future work

We have presented a new PAC-Bayes bound based on a concentration technique derived from the coin-betting formalism. Our new upper bound implies some previous results from PAC-Bayes literature, and at the same time, we have shown that it is tighter in numerical simulations.

One limitation of our result is that it lacks a closed-form minimizer of the upper bound, such as the Gibbs measure in the standard PAC-Bayes analysis. While this is not surprising, it introduces a trade-off between computational complexity and tightness of the bound that was absent in previous approaches. In the future, we aim at precisely characterizing this trade-off, possibly delineating its Pareto frontier. Another interesting venue is to investigate the numerical minimization of our upper bound over data-dependent distributions for risk minimization problems.

Acknowledgments

---

5. We use the `fmincon` function in Matlab.

References

P. Alquier. User-friendly introduction to PAC-Bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. *Advances in Neural Information Processing Systems*, 19, 2006.

J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory (ALT)*, pages 150–165. Springer, 2007.

P. Awasthi, S. Kale, S. Karp, and M. Mohri. Pac-bayes learning bounds for sample-dependent priors. In *Advances in Neural Information Processing Systems*, volume 33, pages 4403–4414, 2020.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

O. Catoni. PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. IMS Lecture Notes-Monograph Series, 56, 2007.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

T. M. Cover. Universal portfolios. *Mathematical Finance*, pages 1–29, 1991.

M. D. Donsker and S. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28, 1975.

P. Dupuis and R. S. R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley-Interscience, 1997.

G. K. Dziugaite and D. Roy. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors. In *International Conference on Machine Learing (ICML)*, pages 1376–1385, 2018.

G. K. Dziugaite and D. M. Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In *Uncertainty in Artificial Intelligence (UAI)*, 2017.

X. Fan, I. Grama, and Q. Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20:1–22, 2015.

P. D. Grünwald and N. A. Mehta. A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity. In *Algorithmic Learning Theory (ALT)*, volume 98, pages 433–465. PMLR, 2019.

M. Haddouche and B. Guedj. PAC-Bayes with unbounded losses through supermartingales. *arXiv preprint arXiv:2210.00928*, 2022.

K.-S. Jun and F. Orabona. Parameter-free online convex optimization with sub-exponential noise. In *Proc. of the Conference on Learning Theory (COLT)*, 2019.

S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in Neural Information Processing Systems*, 21, 2008.

O. Kallenberg. *Random Measures, Theory and Applications*. Springer, 2017.

J. L. Kelly, jr. A new interpretation of information rate. *IRE Transactions on Information Theory*, 2 (3):185–189, 1956.

I. Kuzborskij and Cs. Szepesvári. Efron-Stein PAC-Bayesian Inequalities. arXiv:1909.01931, 2019.

I. Kuzborskij, N. Cesa-Bianchi, and Cs. Szepesvári. Distribution-Dependent Analysis of Gibbs-ERM Principle. In *Conference on Computational Learning Theory (COLT)*, volume 99, pages 2028–2054. PMLR, 2019.

J. Langford and R. Caruana. (Not) bounding the true error. In *Advances in Neural Information Processing Systems*, pages 809–816, 2001.

T. Lattimore and Cs. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. URL https://tor-lattimore.com/downloads/book/book.pdf.

B. London and T. Sandler. Bayesian counterfactual risk minimization. In *International Conference on Machine Learing (ICML)*, 2019.

A. Maurer. A bound on the deviation probability for sums of non-negative random variables. *J. Inequalities in Pure and Applied Mathematics*, 4(1):15, 2003.

A. Maurer. A note on the PAC Bayesian theorem. *arXiv preprint arXiv:0411099*, 2004.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. In *Conference on Computational Learning Theory (COLT)*, 2009.

D. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234, 1998.

Z. Mhammedi, P. Grünwald, and B. Guedj. PAC-Bayes un-expected Bernstein inequality. 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.

F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019. URL https://arxiv.org/abs/1912.13213.

F. Orabona and K.-S. Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *arXiv preprint arXiv:2110.14099*, 2021.

F. Orabona and D. Pál. Coin betting and parameter-free online learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 577–585. Curran Associates, Inc., 2016.

M. Pérez-Ortiz, O. Rivasplata, J. Shawe-Taylor, and Cs. Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 2021.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Computational Learning Theory (COLT)*, pages 1674–1703. PMLR, 2017.

A. Rakhlin and K. Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. In *Proc. of the Conference On Learning Theory (COLT)*, pages 1704–1722, 2017.

O. Rivasplata, I. Kuzborskij, Cs. Szepesvári, and J. Shawe-Taylor. Pac-bayes analysis beyond the usual bounds. In *Advances in Neural Information Processing Systems*, volume 33, pages 16833–16845, 2020.

R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

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

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

G. Shafer and V. Vovk. *Probability and finance: it's only a game!* John Wiley & Sons, 2001.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

J. Shawe-Taylor and R. C. Williamson. A pac analysis of a bayesian estimator. In *Conference on Computational Learning Theory (COLT)*, 1997.

I. O. Tolstikhin and Y. Seldin. PAC-Bayes-empirical-Bernstein inequality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

J. Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939. URL http://archive.numdam.org/item/THESE_1939__218__1_0/.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Y.-S. Wu and Y. Seldin. Split-kl and PAC-Bayes-split-kl inequalities for ternary random variables. In *Advances in Neural Information Processing Systems*, 2022.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

W. Zhou, V. Veitch, M. Austern, R. P. Adams, and P. Orbanz. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach. In *International Conference on Learning Representations (ICLR)*, 2019.

Appendix A. Remaining proofs

A.1. Proof of Corollary 2

Consider the left hand side of Eq. (3) without $\max_{n \in \mathbb{N}}$. Then, we have the following inequalities:

$$\int \max_{\lambda \in [-\frac{1}{1-\mu_\theta}, \frac{1}{\mu_\theta}]} \sum_{i=1}^{n} \ln\left(1 + \lambda(f(\theta, X_i) - \mu_\theta)\right) dP_n(\theta)$$

$$\overset{(a)}{\geq} \int \max_{\lambda \in [-1,1]} \sum_{i=1}^{n} \ln\left(1 + \lambda(f(\theta, X_i) - \mu_\theta)\right) dP_n(\theta)$$

$$\geq \int \max_{\lambda \in [-1,1]} \left\{ \lambda \sum_{i=1}^{n} (f(\theta, X_i) - \mu_\theta) - \lambda^2 \sum_{i=1}^{n} (f(\theta, X_i) - \mu_\theta)^2 \right\} dP_n(\theta)$$

$$\geq \int \max_{\lambda \in [-1,1]} \left\{ \lambda \sum_{i=1}^{n} (f(\theta, X_i) - \mu_\theta) - \lambda^2 n \right\} dP_n(\theta)$$

$$= \frac{1}{4n} \int \left(\sum_{i=1}^{n} (f(\theta, X_i) - \mu_\theta) \right)^2 dP_n(\theta) \quad \text{(Maximizing in } \lambda \text{; note that optimal } \lambda \in [-1, 1])$$

$$\geq \frac{1}{4n} \left(\int \sum_{i=1}^{n} (f(\theta, X_i) - \mu_\theta) dP_n(\theta) \right)^2, \qquad \text{(Jensen's inequality)}$$

where step (a) comes since $[-1, 1] \subset [-\frac{1}{1-\mu_\theta}, \frac{1}{\mu_\theta}]$ since $\mu_\theta \in [0, 1]$ almost surely. Now, applying Theorem 1 gives

$$\mathbb{P}\left\{ \exists n \in \mathbb{N}, \exists P_n : \frac{1}{4n} \left(\int \sum_{i=1}^{n} (f(\theta, X_i) - \mu_\theta) dP_n(\theta) \right)^2 - \mathcal{C}_n \geq \ln \frac{1}{\delta} \right\} \leq \delta$$

and the statement follows. ∎

A.2. Proof of Corollary 3

The proof is based on the following proposition:

Proposition 12 ((Orabona and Jun, 2021, Proposition 1)) *Let* $X_1, \ldots, X_n \in [0, 1]$, *let* $\hat{\mu} := (X_1 + \cdots + X_n)/n$, *and moreover let* $\mu := \mathbb{E}[X_1] \in [0, 1]$. *Then,*

$$\max_{\lambda \in [-\frac{1}{1-\mu}, \frac{1}{\mu}]} \sum_{i=1}^{n} \ln \left(1 + \lambda(X_i - \mu) \right) \geq n \, \mathsf{kl}(\hat{\mu}, \mu) \, .$$

Moreover, if $X_1, \ldots, X_n \in \{0, 1\}$, *we achieve equality in the above.*

Then, Proposition 12 combined with Theorem 1 gives that with probability at least $1 - \delta$, simultaneously for all $n \in \mathbb{N}$ and all P_n, we have

$$\frac{\ln \frac{1}{\delta} + \mathcal{C}_n}{n} \geq \int \mathsf{kl}(\hat{\mu}_\theta, \mu_\theta) \, \mathrm{d}P_n(\theta)$$

$$= \int \sum_{x \in \{0,1\}} \ln \left(\frac{\mathrm{Bern}(x \mid \hat{\mu}_\theta)}{\mathrm{Bern}(x \mid \mu_\theta)} \right) \mathrm{Bern}(x \mid \hat{\mu}_\theta) \, \mathrm{d}P_n(\theta)$$

$$\overset{(a)}{\geq} \sum_{x \in \{0,1\}} \ln \left(\frac{\int \mathrm{Bern}(x \mid \hat{\mu}_\theta) \, \mathrm{d}P_n(\theta)}{\int \mathrm{Bern}(x \mid \mu_\theta) \, \mathrm{d}P_n(\theta)} \right) \int \mathrm{Bern}(x \mid \hat{\mu}_\theta) \, \mathrm{d}P_n(\theta)$$

$$= \mathsf{kl} \left(\int \hat{\mu}_\theta \, \mathrm{d}P_n(\theta), \int \mu_\theta \, \mathrm{d}P_n(\theta) \right),$$

where (a) comes by exchanging summation and integration and applying the log-sum inequality. ∎

A.3. Proof of Corollary 4

The proof largely follows that of Orabona and Jun (2021, Theorem 6). Set $\epsilon_\theta = \mu_\theta - \hat{\mu}_\theta$, and so $\epsilon_\theta + \hat{\mu}_\theta \in [0, 1]$. Then, we have

$$\psi^\star(\theta, \mu_\theta) \geq \max_{\lambda \in [-1,1]} \sum_{i=1}^{n} \ln \left(1 + \lambda(f(\theta, X_i) - (\hat{\mu}_\theta + \epsilon_\theta)) \right),$$

and applying Jensen's inequality

$$\int \psi^\star(\theta, \mu_\theta) \, \mathrm{d}P_n \geq \max_{\lambda \in [-1,1]} \int \sum_{i=1}^{n} \ln \left(1 + \lambda(f(\theta, X_i) - (\hat{\mu}_\theta + \epsilon_\theta)) \right) \mathrm{d}P_n \, . \tag{14}$$

Now, we further relax the above by taking a lower bound. In particular, (Fan et al., 2015, Eq. 4.12) shows that for any $|x| \leq 1$ and $|\lambda| \leq 1$,

$$\ln(1 + \lambda x) \geq \lambda x + \left(\ln(1 - |\lambda|) + |\lambda| \right) x^2 \, . \tag{15}$$

The above is combined with the following lemma:

Lemma 13 ((Orabona and Jun, 2021, Lemma 5)) *Let* $f(\lambda) = a\lambda + b \left(\ln(1 - |\lambda|) + |\lambda| \right)$ *for some* $a \in \mathbb{R}, b \geq 0$. *Then,* $\max_{\lambda \in [-1,1]} f(\lambda) \geq \frac{a^2}{(4/3)|a| + 2b}$.

Thus,

$$\int \psi^\star(\theta, \mu_\theta)\, \mathrm{d}P_n \overset{(a)}{\geq} \lambda \int \sum_{i=1}^{n} \big(f(\theta, X_i) - (\hat{\mu}_\theta + \epsilon_\theta)\big)\, \mathrm{d}P_n$$

$$+ \big(\ln(1 - |\lambda|) + |\lambda|\big) \int \sum_{i=1}^{n} \big(f(\theta, X_i) - (\hat{\mu}_\theta + \epsilon_\theta)\big)^2 \mathrm{d}P_n$$

$$= -n\lambda \int \epsilon_\theta\, \mathrm{d}P_n$$

$$+ \big(\ln(1 - |\lambda|) + |\lambda|\big) \left(\int \sum_{i=1}^{n}(f(\theta, X_i) - \hat{\mu}_\theta)^2\, \mathrm{d}P_n + n \int \epsilon_\theta^2\, \mathrm{d}P_n\right)$$

$$\overset{(b)}{\geq} -n\lambda \int \epsilon_\theta\, \mathrm{d}P_n$$

$$+ \big(\ln(1 - |\lambda|) + |\lambda|\big) \left(\int \sum_{i=1}^{n}(f(\theta, X_i) - \hat{\mu}_\theta)^2\, \mathrm{d}P_n + n \left(\int \epsilon_\theta\, \mathrm{d}P_n\right)^2\right)$$

$$\overset{(c)}{\geq} \frac{n^2 \left(\int \epsilon_\theta\, \mathrm{d}P_n\right)^2}{(4/3)n \left|\int \epsilon_\theta\, \mathrm{d}P_n\right| + 2\int \sum_{i=1}^{n}(f(\theta, X_i) - \hat{\mu}_\theta)^2\, \mathrm{d}P_n + 2n \left(\int \epsilon_\theta\, \mathrm{d}P_n\right)^2},$$

where step (a) comes by application of Eqs. (14) and (15), step (b) comes by Jensen's inequality, and step (c) comes by application of Lemma 13. Now, the above combined with Theorem 1 gives

$$n^2 \left(\int \epsilon_\theta\, \mathrm{d}P_n\right)^2 \leq n \left(\mathcal{C}_n + \ln\frac{1}{\delta}\right) \left(\frac{4}{3}\left|\int \epsilon_\theta\, \mathrm{d}P_n\right| + 2\hat{V}(P_n) + 2\left(\int \epsilon_\theta\, \mathrm{d}P_n\right)^2\right).$$

Finally, solving the above for $\int \epsilon_\theta\, \mathrm{d}P_n$, using subadditivity of square root, and relaxing some numerical constants we get

$$\left|\int \epsilon_\theta\, \mathrm{d}P_n\right| \leq \frac{\sqrt{2\left(\mathcal{C}_n + \ln\frac{1}{\delta}\right)\hat{V}(P_n)}}{\left(\sqrt{n} - \frac{2}{\sqrt{n}}\left(\mathcal{C}_n + \ln\frac{1}{\delta}\right)\right)_+} + \frac{2\left(\mathcal{C}_n + \ln\frac{1}{\delta}\right)}{\left(n - 2\left(\mathcal{C}_n + \ln\frac{1}{\delta}\right)\right)_+}.$$

■

A.4. Proof of Corollary 5

Recall that $\mathcal{C}_n = \mathsf{KL}(P_n \| P_0) + \ln \frac{\sqrt{\pi}\Gamma(n+1)}{\Gamma(n+\frac{1}{2})}$. We will use several facts from the proof of Theorem 1 appearing in Section 5.2. Consider Eq. (12) (without taking $\max_{n \in \mathbb{N}} \cdot$) combined with Eq. (11), which gives us

$$\sup_{P_n} \int \psi_n^\star(\theta)\, \mathrm{d}P_n(\theta) - \mathcal{C}_n \leq \ln \underbrace{\int \prod_{i=1}^{n} \big(1 + B_i(\theta)\Delta_i(\theta)\big)\, \mathrm{d}P_0(\theta)}_{M_n},$$

and taking exponential and expectation on both sides,

$$\mathbb{E}\exp\left(\sup_{P_n}\int\psi_n^\star(\theta)\,\mathrm{d}P_n(\theta)-\mathcal{C}_n\right)\leq\mathbb{E}M_n=1$$

since $M_0=1$.

Now we will lower bound the left hand side in the above by using (Fan et al., 2015, Eq. 4.11) (which also appears in (Mhammedi et al., 2019, Eq. 15)): namely for $x\in[0,1]$,

$$\exp(-\lambda x-\lambda^2\alpha(\lambda)x^2)\leq 1-\lambda x \qquad (\lambda\in(0,1))$$

where $\alpha(\lambda)=(-\ln(1-\lambda)-\lambda)/\lambda^2$. Recalling definition of $\psi_n^\star$ and using the above inequality,

$$
\begin{aligned}
\psi_n^\star(\theta) &= \max_{\lambda\in\left[-\frac{1}{1-\mu_\theta},\frac{1}{\mu_\theta}\right]}\sum_{i=1}^n\ln\left(1+\lambda\Delta_i(\theta)\right)\\
&\geq \max_{\lambda\in[0,1]}\sum_{i=1}^n\ln\left(1+\lambda\Delta_i(\theta)\right)\\
&\geq \max_{\lambda\in[0,1]}\sum_{i=1}^n\left(-\lambda\Delta_i(\theta)-\alpha(\lambda)\lambda^2\Delta_i(\theta)^2\right)
\end{aligned}
$$

which completes the proof. ∎

A.5. Proof of Lemma 7

We can rewrite f as

$$f(x)=\max_{0\leq b\leq 1}\ln\left[1+(c-x)\left(-\frac{1}{1-x}+\left(\frac{1}{1-x}+\frac{1}{x}\right)b\right)\right].$$

Now, consider the argument of the max. We claim that it is convex in x for any $b\in[0,1]$. In fact, the second derivative is

$$\frac{1}{x^2}+\frac{1}{(1-x)^2}-\frac{(b+c-1)^2}{(-x(b+c)+bc+x)^2}=\frac{1}{x^2}+\frac{1}{(1-x)^2}-\frac{1}{(x+\frac{bc}{1-b-c})^2} \qquad (16)$$

We claim that for $b,c\in[0,1]$, we have $g(b,c):=\frac{bc}{1-b-c}\in(-\infty,-1]\cup[0,\infty)$. To see this, fix b and consider $c\leq 1-b$. In this regime, $g(b,0)=0$ and $g(b,c)$ is increasing to infinity as c goes from 0 to $1-b$. In the other regime of $c>1-b$, we have $g(b,1)=-1$ and $g(b,c)$ is decreasing as c decreases from 1 to $1-b$. This proves the claim.

Therefore, to lower bound (16) we need to lower bound $(x+z)^2$ where $z\in(-\infty,-1]\cup[0,\infty)$. Since $(x+z)^2$ is minimized at $-x$ but the range of z never includes $-x$ (except for the boundary case), we have that $\min_{z\in(-\infty,-1]\cup[0,\infty)}(x+z)^2=\min\{(x-1)^2,x^2\}$. Therefore,

$$-\frac{(b+c-1)^2}{(-x(b+c)+bc+x)^2}\geq-\max\left\{\frac{1}{(x-1)^2},\frac{1}{x^2}\right\}\geq-\frac{1}{(x-1)^2}-\frac{1}{x^2}.$$

Hence, f is a maximum of convex functions, that concludes the proof. ∎

A.6. Proof of Proposition 9

It is clear that the optimal log-wealth $\psi_n^\star(\theta, \mu_\theta)$ is non-negative over all arguments, because the maximization range includes 0. Then, for a fixed block B_k, by the non-negativity of $\psi_n^\star$, Markov's inequality gives

$$\mathbb{P}\left\{ e \int \psi_n^\star(\theta, \mu_\theta)\, \mathrm{d}P_n(\theta) \leq \frac{1}{m} \sum_{i \in B_k} \psi_n^\star(\theta_i, \mu_{\theta_i}) \right\} \leq 1/e .$$

Since all blocks are independent,

$$\mathbb{P}\left\{ \bigcap_{k \in [K]} \left\{ e \int \psi_n^\star(\theta, \mu_\theta)\, \mathrm{d}P_n(\theta) \leq \frac{1}{m} \sum_{i \in B_k} \psi_n^\star(\theta_i, \mu_{\theta_i}) \right\} \right\} \leq e^{-K}$$

and so there exists at least one B_k such that with probability at least $1 - e^{-K}$,

$$\int \psi_n^\star(\theta, \mu_\theta)\, \mathrm{d}P_n(\theta) \geq \frac{1}{em} \sum_{i \in B_k} \psi_n^\star(\theta_i, \mu_{\theta_i}) .$$

Thus, with probability at least $1 - e^{-K}$,

$$\int \psi_n^\star(\theta, \mu_\theta)\, \mathrm{d}P_n(\theta) \geq \min_{k \in [K]} \frac{1}{em} \sum_{i \in B_k} \psi_n^\star(\theta_i, \mu_{\theta_i}) .$$

∎

Note that we can change e to any other constant $C > 1$. Then, with probability $1 - C^{-K}$,

$$\int \psi_n^\star(\theta, \mu_\theta)\, \mathrm{d}P_n(\theta) \geq \min_{k \in [K]} \frac{1}{Cm} \sum_{i \in B_k} \psi_n^\star(\theta_i, \mu_{\theta_i}) .$$

A.7. Proof of Corollary 8

Recall that $(\theta_i)_{i \in B_k} \sim P_n^m$ for any $k \in [K]$ and $\mu_{\theta_i} = \mathbb{E}[f(\theta_i, X_1)], \forall i \in [Km]$. Our goal is to find an upper and lower bound for $\mu^* := \int \mu_\theta\, \mathrm{d}P_n(\theta)$. Given $P_n(\theta)$, Theorem 1 gives us

$$\mathbb{P}\left\{ \exists n \in \mathbb{N}, \exists P_n : \int \psi_n^\star(\theta, \mu_\theta)\, \mathrm{d}P_n(\theta) - \mathcal{C}_n \leq \ln \frac{1}{\delta} \right\} \geq 1 - \delta . \tag{17}$$

Moreover, using the union bound and the standard KL-divergence concentration inequality for $[0, 1]$-bounded random variable, we have

$$\mathbb{P}\left\{ \max_{k \in [K]} \mathsf{kl}\left(\frac{1}{m} \sum_{i \in B_k} \mu_{\theta_i},\ \mu^* \right) \leq \ln \frac{K}{2\delta} \right\} \geq 1 - \delta . \tag{18}$$

Furthermore, Proposition 9 states that

$$\mathbb{P}\left\{ \min_{k \in [K]} \frac{1}{m} \sum_{i \in B_k} \psi_n^\star(\theta_i, \mu_{\theta_i}) \leq \int \psi_n^\star(\theta, \mu_\theta)\, \mathrm{d}P_n(\theta) \ \middle|\ X_1, \ldots, X_n \right\} \geq 1 - e^{-K} . \tag{19}$$

By the union bound, and assuming setting $K = \lceil \ln(1/\delta) \rceil$ one can see that with probability at least $1 - 3\delta$, the concentration events in the three inequalities are all true. Assume that these events are true. It suffices to show $\mu^* \leq M_U$ since the proof of $\mu^* \geq M_L$ is symmetric.

Denote by $\nu_{\theta_i}^U$ the solutions of the optimization problem in (7).

Let $k^* = \arg\min_{k\in[K]} \frac{1}{m}\sum_{i\in B_k} \psi_n^\star(\theta_i, \mu_{\theta_i})$. By the events in (17) and (19), we have that $(\mu_{\theta_i})_{i\in B_{k^*}}$ is a feasible solution of (7) with $k = k^*$. This implies that

$$\bar{\mu}(k^*) := \frac{1}{m}\sum_{i\in B_{k^*}} \mu_{\theta_i} \le \frac{1}{m}\sum_{i\in B_{k^*}} \nu_{\theta_i}^U \le \frac{1}{m}\sum_{i\in B_{\hat{k}_U}} \nu_{\theta_i}^U = \bar{\nu}_U(\hat{k}_U) =: \nu^*. \tag{20}$$

Consider the following two cases:

- Case 1: $\nu^* \ge \mu^*$.
  We just need to verify that $\nu^* \le M_U$. This holds by the definition of M_U.

- Case 2: $\nu^* < \mu^*$.
  We have

  $$\mathsf{kl}(\nu^*, \mu^*) \le \mathsf{kl}(\bar{\mu}(k^*), \mu^*) \le \ln\frac{K}{2\delta} \overset{(8)}{=} \mathsf{kl}(\nu^*, M_U).$$

  where the first inequality is due to $\bar{\mu}(k^*) \le \nu^* < \mu^*$. Applying the monotonicity of kl in the second argument to $\mathsf{kl}(\nu^*, \mu^*) \le \mathsf{kl}(\nu^*, M_U)$, we have $\mu^* \le M_U$.

This concludes the proof. ∎

Appendix B. Monte Carlo experiment

In this section, we investigate numerically the Monte Carlo approximation discussed in Section 5.1.1. In particular, we want to validate the claim that the confidence intervals calculated with Algorithm 1 are better than the ones of Maurer (2003) when enough Monte Carlo sample are used.

Here is the list of parameters in our experiment:

- The number of samples: $n = 32$ (Figure 3), $n \in \{2^c : c \in \{1, \ldots, 4\}\}$ (Figure 4).

- The number of groups: $K = 4$, the corresponding multiplier: $C = 2.1147$ (check the end of Appendix A.6).

- The number of Monte Carlo samples in each group: $m \in \{2^c : c = 1, \ldots, 10\}$ (Figure 3), $m = 256$ (Figure 4).

- Parameter space $\Theta = \mathbb{R}$.

- Prior distribution P_0 is $N(0, 1)$.

- Posterior distribution P_n is $N(0, 0.25)$.

- Samples: $X_1, \ldots, X_n$ drawn from $N(0, 1)$, i.i.d.

- $f(\theta, x) = (\mathrm{erf}(x\theta) + 1)/2$.

- Failure probability $\delta = 0.05$.

For the fair use of parameters, we ensure that the total number of MC samples used in Maurer's bound matches that of ours. That is, when ours use m Monte Carlo samples for each group $k \in [K]$, Maurer's bound use $M = m \cdot K$ Monte Carlo samples. We describe how we compute Maurer's bound numerically in Algorithm 2.

Algorithm 2 Monte Carlo Approximation for Maurer's Bound

1: **Input**: Failure probability δ, sample size parameters $M \in \mathbb{N}$, posterior P_n, prior P_0
2: Sample $(\theta_i)_{i \in [M]} \sim P_n^M$
3: $\hat{\mu}_M = \frac{1}{M} \sum_{j=1}^M \hat{\mu}_{\theta_i} = \frac{1}{Mn} \sum_{j=1}^M \sum_{i=1}^n f(\theta_i, X_i)$
4: Let $w = \sqrt{\frac{\ln(\frac{2}{\delta})}{M}}$
5: Compute the monte-carlo error bound.

$$\hat{k}_U = \max \left\{ \mu \colon \mathsf{kl}\left(\hat{\mu}_M, \mu\right) \leq w \right\} \ , \ \hat{k}_L = \min \left\{ \mu \colon \mathsf{kl}\left(\hat{\mu}_M, \mu\right) \leq w \right\} \tag{21}$$

6: Compute

$$M_U = \max \left\{ \mu \colon \mathsf{kl}\left(\hat{k}_U, \mu\right) \leq \frac{\mathcal{C}_n + \ln \frac{2}{\delta}}{n} \right\} \ , \ M_L = \min \left\{ \mu \colon \mathsf{kl}\left(\hat{k}_L, \mu\right) \leq \frac{\mathcal{C}_n + \ln \frac{2}{\delta}}{n} \right\} \tag{22}$$

7: **Output:** M_L and M_U

The value w in Algorithm 2 is the confidence width of the Monte Carlo error between $\hat{\mu}_m$ and $\int \hat{\mu}_\theta \, \mathrm{d}P_n(\theta)$ based on the following Proposition 14 that can be shown using the standard Chernoff technique (e.g., Corollary 10.4 of Lattimore and Szepesvári (2020)).

Proposition 14 (Binomial Chernoff bound) *Let $\delta \in (0, 1]$. Let $A_1, \ldots, A_M \in [0, 1]$ be a sequence of i.i.d. random variables with mean $\mu = \mathbb{E}[A]$. Let $\bar{A} = \frac{1}{M} \sum_{i=1}^M A_i$. With probability at least $1 - \delta$, we have*

$$\mathsf{kl}\left(\bar{A}, \mu\right) \leq \sqrt{\frac{\ln(\frac{2}{\delta})}{M}} \ .$$

When the parameter space Θ is discrete, we can precisely compute $\int \hat{\mu}_\theta \, \mathrm{d}P_n(\theta)$, and from this value we can use Maurer's bound to obtain the confidence bound. However, in our current setting we only have $[\hat{k}_L, \hat{k}_U]$ that is a confidence interval for $\int \hat{\mu}_\theta \, \mathrm{d}P_n(\theta)$. In this case, the confidence bound of actual $\int \mu_\theta \, \mathrm{d}P_n(\theta)$ is the union of all possible confidence intervals, or formally,

$$\cup_{\mu' \in [\hat{k}_L, \hat{k}_U]} \left\{ \mu : \mathsf{kl}(\mu', \mu) \leq \frac{\mathcal{C}_n + \ln \frac{1}{\delta}}{n} \right\} \ .$$

Thanks to the property of the kl, we only need to check two endpoints $\hat{k}_L$ and $\hat{k}_U$ to get the final Maurer's bound (M_L, M_U).

Figure 3 shows that our Algorithm 1 outperforms the Maurer's bound with sufficient amount of Monte Carlo samples. Even with the increasing number of data samples (Figure 4), our algorithm steadily outperforms the Maurer's bound.

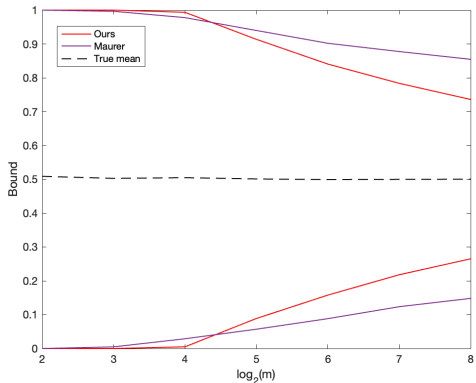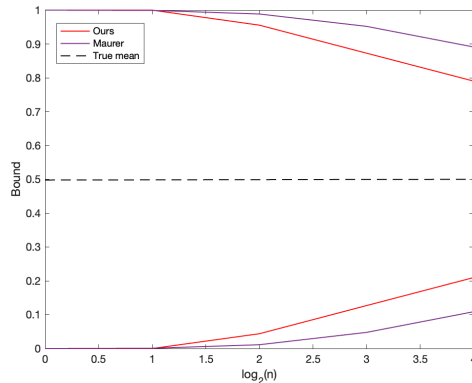Figure 3: Confidence width vs Monte Carlo samples (m).

Figure 4: Confidence width vs the number of samples (n).