Better-than-KL PAC-Bayes Bounds

Ilja Kuzborskij ILJAK@GOOGLE.COM
Google DeepMind

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

Yulian Wu YULIAN.WU@KAUST.EDU.SA
King Abdullah University of Science and Technology

Kyoungseok Jang KSAJKS@GMAIL.COM
Università degli Studi di Milano

Francesco Orabona FRANCESCO@ORABONA.COM
King Abdullah University of Science and Technology

Editors: Shipra Agrawal and Aaron Roth

Abstract

Let $f(\theta, X_1), \ldots, f(\theta, X_n)$ be a sequence of random elements, where f is a fixed scalar function, $X_1, \ldots, X_n$ are independent random variables (data), and θ is a random parameter distributed according to some data-dependent *posterior* distribution P_n. In this paper, we consider the problem of proving concentration inequalities to estimate the mean of the sequence. An example of such a problem is the estimation of the generalization error of some predictor trained by a stochastic algorithm, such as a neural network, where f is a loss function. Classically, this problem is approached through a *PAC-Bayes* analysis where, in addition to the posterior, we choose a *prior* distribution which captures our belief about the inductive bias of the learning problem. Then, the key quantity in PAC-Bayes concentration bounds is a divergence that captures the *complexity* of the learning problem where the de facto standard choice is the Kullback-Leibler (KL) divergence. However, the tightness of this choice has rarely been questioned.

In this paper, we challenge the tightness of the KL-divergence-based bounds by showing that it is possible to achieve a strictly tighter bound. In particular, we demonstrate new *high-probability* PAC-Bayes bounds with a novel and *better-than-KL* divergence that is inspired by Zhang et al. (2022). Our proof is inspired by recent advances in regret analysis of gambling algorithms, and its use to derive concentration inequalities. Our result is first-of-its-kind in that existing PAC-Bayes bounds with non-KL divergences are not known to be strictly better than KL. Thus, we believe our work marks the first step towards identifying optimal rates of PAC-Bayes bounds.

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

1. Introduction

We study the standard model of statistical learning, where we are given a set of independent observations $X_1, \ldots, X_n \in \mathcal{X}$, and we have access to a measurable parametric function $f : \Theta \times \mathcal{X} \to [0, 1]$. In particular, we are interested in estimating the mean of f when its first parameter θ is a *random* parameter drawn from a distribution chosen by an algorithm based on data (typically called the *posterior* P_n). In other words, our goal is to estimate the mean[1]

$$\int \mathbb{E}[f(\theta, X_1)] \, \mathrm{d}P_n(\theta) \, .$$

In many learning scenarios, $f(\theta, X_i)$ is interpreted as a composition of a loss function and a predictor evaluated on example X_i, given the parameter $\theta \sim P_n$. Often, this problem is captured by giving bounds on the *generalization error* (sometimes called *generalization gap*)

$$\int \Delta_n(\theta) \, \mathrm{d}P_n(\theta), \qquad \text{where} \qquad \Delta_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} (f(\theta, X_i) - \mathbb{E}[f(\theta, X_1)]) \, .$$

To have a sharp understanding of the behavior of the generalization error, it is often desirable to have bounds that hold with high probability over the sample. At the same time, standard tools for this task, such as concentration inequalities (for instance, Chernoff or Hoeffding inequalities) are not applicable here, since P_n itself depends on the sample and can be potentially supported on infinite sets (which precludes the use of union bounds). In the particular setting of a randomized prediction discussed here, studying such bounds was a long topic of interest in the *PAC-Bayes* analysis (Shawe-Taylor and Williamson, 1997; McAllester, 1998). PAC-Bayes bounds typically require an additional component called the *prior* distribution over parameter space, which captures our belief about the inductive bias of the problem. Then, for any data-free prior distribution P_0, a classical result holds with a probability at least $1 - \delta$ (for a failure probability $\delta \in (0, 1)$) over the sample *simultaneously* for all choices of data-dependent posteriors P_n:

$$\int \Delta_n(\theta) \, \mathrm{d}P_n(\theta) = \mathcal{O}\left(\sqrt{\frac{D_{\mathrm{KL}}(P_n, P_0) + \ln \frac{\sqrt{n}}{\delta}}{n}} \right) \qquad \text{as} \quad n \to \infty \, . \tag{1}$$

While such a bound is uniform over all choices of P_n, it does scale with the KL divergence between them, $D_{\mathrm{KL}}(P_n, P_0) = \int \ln(\mathrm{d}P_n / \mathrm{d}P_0) \, \mathrm{d}P_n$, which can be thought of as a measure of complexity of the learning problem. Over the years, PAC-Bayes bounds were developed in many ever tighter variants, such as the ones for Bernoulli losses (Langford and Caruana, 2001; Seeger, 2002; Maurer, 2004), exhibiting fast rates given small loss variances (Tolstikhin and Seldin, 2013; Mhammedi et al., 2019), data-dependent priors (Rivasplata et al., 2020; Awasthi et al., 2020), and so on. However, one virtually invariant feature remained: high probability PAC-Bayes bounds were always stated using the KL-divergence. The reason is that virtually all of these proofs were based on the Donsker-Varadhan *change-of-measure* inequality[2] (essentially arising from a relaxation of a variational representation of KL divergence) (Donsker and Varadhan, 1975).

Recently, several works have looked into PAC-Bayes analyses arising from the use of different change-of-measure arguments (Bégin et al., 2016; Alquier and Guedj, 2018; Ohnishi and Honorio, 2021), allowing to replace KL divergence with other divergences such as χ^2 or Hellinger, however these results either did not hold with high probability or involved looser divergences (such as χ^2).

[1] Throughout, the integration $\int \equiv \int_\Theta$ is always understood w.r.t. $\theta \in \Theta$ while the expectation $\mathbb{E}$ is always w.r.t. data.

[2] For any measurable F, and any P_n, P_0, the inequality states that $\int F \, \mathrm{d}P_n \le D_{\mathrm{KL}}(P_n, P_0) + \ln \int e^F \, \mathrm{d}P_0$.

In this work, we propose an alternative change-of-measure analysis and show, to the best of our knowledge, the first high-probability PAC-Bayes bound that involves a divergence *tighter* than the KL-divergence.

Our analysis is inspired by a recent observation of Zhang et al. (2022), who pointed out an interesting phenomenon arising in regret analysis of online algorithms (Cesa-Bianchi and Orabona, 2021): They focused on a classical problem of learning with experts advice (Cesa-Bianchi and Lugosi, 2006), where the so-called parameter-free regret bounds scale with $\sqrt{D_{\mathrm{KL}}}$ between the competitor distribution over experts and the choice of the prior (Luo and Schapire, 2015; Orabona and Pál, 2016). In particular, their analysis improves over parameter-free rates replacing $\sqrt{D_{\mathrm{KL}}}$ with a divergence with the shape

$$D_{\mathrm{ZCP}}(P_n, P_0; c) = \int \left| \frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1 \right| \sqrt{\ln\left(1 + c^2 \left| \frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1 \right|^2 \right)} \,\mathrm{d}P_0(\theta), \tag{2}$$

where c is a parameter.[3] We call it Zhang-Cutkosky-Paschalidis (ZCP) divergence. Interestingly, the ZCP divergence enjoys the following upper bound (Theorem 2):

$$D_{\mathrm{ZCP}} \lesssim \sqrt{D_{\mathrm{KL}}\, D_{\mathrm{TV}}} + D_{\mathrm{TV}}, \tag{3}$$

where $D_{\mathrm{TV}}(P_n, P_0) = \frac{1}{2}\int |\,\mathrm{d}P_n - \mathrm{d}P_0|$ is the Total Variation (TV) distance. Since $D_{\mathrm{TV}} \le 1$ for any pair of distributions, D_{ZCP} is orderwise tighter[4] than $\sqrt{D_{\mathrm{KL}}}$ and we show in Section 4.1 that in some cases the gain can be substantial.

In this paper, we develop a novel and straightforward change-of-measure type analysis that leads to PAC-Bayes bounds with the ZCP divergence, avoiding the regret analysis of Zhang et al. (2022) altogether.

Our contributions Our overall contribution is to show a surprising result that the choice of the KL divergence as the complexity measure in PAC-Bayes bounds is suboptimal, which tells us that there is much room for studying optimal rates of PAC-Bayes bounds.

Specifically, we show that the KL divergence of existing PAC-Bayes bounds can be strictly improved using a different, better divergence. We achieve it through two main results. Our first result is a PAC-Bayes bound (Theorem 8)

$$\int \Delta_n(\theta)\,\mathrm{d}P_n(\theta) \le \frac{\sqrt{2}\, D_{\mathrm{ZCP}}(P_n, P_0; \sqrt{2n}/\delta) + 2 + \sqrt{\ln \frac{2\sqrt{n}}{\delta}}}{\sqrt{n}},$$

which readily improves upon Eq. (1), since by upper-bounding a ZCP-divergence we have

$$\int \Delta_n(\theta)\,\mathrm{d}P_n(\theta) = \mathcal{O}\left(\sqrt{\frac{D_{\mathrm{KL}}(P_n, P_0)\, D_{\mathrm{TV}}(P_n, P_0) + \ln \frac{\sqrt{n}}{\delta}}{n}}\right) \quad \text{as} \quad n \to \infty\,. \tag{4}$$

[3] There is one minor difference that the original divergence appeared in Zhang et al. (2022) has $\sqrt{\ln(1 + |\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1|)}$ instead of $\sqrt{\ln(1 + c^2|\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1|^2)}$.

[4] See Section 4 for a detailed explanation of orderwise tightness.

Our second contribution is a bound that extends to regimes beyond $1/\sqrt{n}$ rate, such as fast rates of order $1/n$ when the sample variance of f is small. Here, we consider variants of empirical Bernstein inequality and a bound on 'little-kl' (a bound for binary or bounded $f()$; defined in Section 3), see Section 5.1. In fact, instead of deriving individual bounds separately, we first show a generic bound that can be relaxed to obtain these styles of bounds. This technique of obtaining a generic bound is inspired by the recent advances for obtaining concentration inequalities through the regret analysis of online betting algorithms (Jun and Orabona, 2019; Orabona and Jun, 2024). In particular, we consider the expected *optimal log wealth* (denoted by $\ln \mathrm{W}_n^*$) of an algorithm that bets on the outcomes $f(\theta, X_i) - \mathbb{E}[f(\theta, X_1)]$. Then, using a regret bound of the betting algorithm (which is only required for the proof; one does not need to this algorithm), we obtain concentration inequalities from upper bounds on $\int \cdot \mathrm{d}P_n$ of

$$\ln \mathrm{W}_n^*(\theta) = \max_{\beta \in [-1,1]} \ln \prod_{i=1}^{n} \big(1 + \beta(f(\theta, X_i) - \mathbb{E}[f(\theta, X_1)])\big) \ .$$

Recently, Jang et al. (2023) have used this technique to control $\int \ln \mathrm{W}_n^*(\theta) \mathrm{d}P_n$ in terms of KL divergence, and used various lower bounds on the logarithm to recover many known PAC-Bayes inequalities, such as PAC-Bayes Bernstein inequality and others. As an illustrative example, consider the simple inequality $\ln(1+x) \geq x - x^2$ for $|x| \leq 1/2$: If one can show that $\int \ln \mathrm{W}_n^*(\theta) \mathrm{d}P_n \leq \mathrm{bound}(\delta, n, P_n, P_0)$, then the above implies that $\max_{|\beta| \leq 1/2}\{\beta \Delta_n - \beta^2 n\} \leq \mathrm{bound}(\delta, n, P_n, P_0)$ and so we can optimally tune β to obtain $\int \Delta_n(\theta) \mathrm{d}P_n \leq \sqrt{\mathrm{bound}(\delta, n, P_n, P_0)/n}$, which recovers a familiar bound in the shape of Eq. (1). A more fine-grained analysis leads to an empirical Bernstein type inequality (see Corollary 14). This suggests that the optimal log wealth $\ln \mathrm{W}_n^*$ is the fundamental quantity that unifies various existing types of concentration bounds such as Hoeffding, Bernoulli-KL, and empirical Bernstein inequalities.

Using the above concept of optimal log wealth, we show (see Eq. (5)) that there exists a universal constant $c > 0$, such that almost surely for all distributions P (possibly data-dependent ones),

$$\limsup_{n \to \infty} \frac{\int \ln \mathrm{W}_n^*(\theta) \mathrm{d}P(\theta)}{\ln^{3/2}(n)} \leq c \left(1 + \sqrt{D_{\mathrm{KL}}(P, P_0)\, D_{\mathrm{TV}}(P, P_0)}\, \big(1 + \sqrt{D_{\mathrm{KL}}(P, P_0)}\big)\right) \ .$$

We state this asymptotic bound in terms of the upper bound on the ZCP divergence. However, compared to Eq. (4), this bound is more versatile as it can be used to obtain an empirical Bernstein inequality. That is, we show later that it implies

$$\limsup_{n \to \infty} \frac{\left|\int \Delta_n(\theta)\, \mathrm{d}P\right|^2}{\frac{1}{n}\left(\left|\int \Delta_n(\theta)\, \mathrm{d}P\right| + \hat{V}(P)\right) \cdot \ln^{3/2}(n)} \leq c \left(1 + \sqrt{D_{\mathrm{KL}}\, D_{\mathrm{TV}}}\, \big(1 + \sqrt{D_{\mathrm{KL}}}\big)\right),$$

where $\hat{V}(P)$ is the empirical variance averaged over P. While it comes with an additional dependency on $\sqrt{D_{\mathrm{KL}}}$, we show later in Section 4 that is still never worse than existing KL-based bounds yet enjoys orderwise better bounds in some instances.

2. Additional related work

PAC-Bayes *PAC-Bayes* has been a long-lasting topic of interest in statistical learning (Shawe-Taylor and Williamson, 1997; McAllester, 1998; Catoni, 2007), with a considerably interest in both theory and applications; see Alquier (2024) for a comprehensive survey. Over the years, most of the PAC-Bayes literature was concerned with tightening Eq. (1) by using more advanced concentration inequalities and making assumptions about data-generating distributions. A notable improvement is a bound that switches to the rate $1/n$ for a sufficiently small sample variance (Tolstikhin and Seldin, 2013), an empirical Bernstein inequality, which was further improved by Mhammedi et al. (2019). Fast rates where also noticed in other bounds, which are useful in situations when losses are sufficiently small (Catoni, 2007; Yang et al., 2019). Several results have also relaxed the independence assumption in PAC-Bayes analysis through martingale conditions (Seldin et al., 2012; Kuzborskij and Szepesvári, 2019; Haddouche and Guedj, 2023; Lugosi and Neu, 2023). Finally, a recent surge of practical interest in PAC-Bayes was stimulated by its ability to yield numerically non-vacuous generalization bounds for deep neural networks (Dziugaite and Roy, 2017; Pérez-Ortiz et al., 2021; Dziugaite et al., 2021).

All of these results, similarly as the classical ones, involve the KL divergence, which is considered as the de facto standard divergence to be used for PAC-Bayes bounds. Our paper indicates that this standard KL-based bound is in fact suboptimal by showing that there exists a new divergence that is never worse than KL orderwise while showing strict improvements in special cases.

Other divergences and connection to change-of-measure inequalities The focus of this paper is on the 'complexity' term, or divergence between the posterior and prior distributions over parameters. Several works have dedicated some attention to this topic and obtained bounds with alternative divergences. Alquier and Guedj (2018) studied the setting of unbounded losses and derived PAC-Bayes bounds with χ^2 divergence instead of KL divergence, however at the cost of a high probability guarantee: their bound scales with $1/\delta$ rather than $\ln(1/\delta)$. In another notable work, Ohnishi and Honorio (2021) proved a suite of change-of-measure inequalities distinct from the usual Donsker-Varadhan inequality. Their method is based on a tighter variational representation of f-divergences developed by Ruderman et al. (2012), which in turn improves upon Nguyen et al. (2010). The variational representation arises from the use the Fenchel-Young inequality with respect to f under the integral operator. For example, for some measurable function F, we have $\int \frac{\mathrm{d}P}{\mathrm{d}Q} \cdot F \mathrm{d}Q \leq \int \big(f(\frac{\mathrm{d}P}{\mathrm{d}Q}) + f^\star(F)\big)\mathrm{d}Q$ where $f^\star$ is a convex conjugate of f. In this paper, we focus on a particular f, whose convex conjugate is a function $f^\star(y) = \delta \, \exp(y^2/(2n))$. In fact, $\int f(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1)\mathrm{d}Q$ then gives rise to the ZCP divergence.

Interestingly, the function of a shape $y \mapsto \exp(y^2/2)$ appears in several other contexts. In the online learning literature this function, identified as the potential function or the dual of the regularizer, is used in the design and analysis of the so-called parameter-free algorithms, both for learning with expert advice (Chaudhuri et al., 2009; Luo and Schapire, 2015; Koolen and Van Erven, 2015; Orabona and Pál, 2016) and for online convex optimization (Orabona and Pál, 2016).

Chu and Raginsky (2023) also derived an interesting generalization error bound using the Fenchel-Young inequality in the L_p-Orlicz norm. In their analysis they focus on the function $f^\star(y) = \exp(y^p)$, with the majority of their results obtained with $p = 2$. Albeit their analysis commences from the Fenchel-Young inequality, later on it is simplified through the use of the inverse function f^{-1} instead of the dual $f^\star$, resulting in a looser upper bound, ultimately leading back to the KL divergence.

Concentration from coin-betting Our paper occasionally relies on the coin-betting formalism (see Section 3.1), which goes back to Ville (1939) and the Kelly betting system (Kelly, 1956). The coin-betting formalism can be thought of as a simple instance of the Universal Portfolio theory (Cover, 1991). The idea of showing concentration inequalities using regret of online betting algorithms was first investigated by Jun and Orabona (2019) who established new (time-uniform) concentration inequalities. Their work was heavily inspired by Rakhlin and Sridharan (2017) who showed the equivalence between the regret of generic online linear algorithms and martingale tail bounds. However, the idea of linking online betting to statistical testing, but without the explicit use of regret analysis, goes back at least to Cover (1974) and more recently to Shafer and Vovk (2001).

3. Definitions and preliminaries

For two nonnegative valued functions f, g, we use $f \lesssim g$ to indicate that there exists a constant $C > 0$ such that $f \leq C \operatorname{polylog}(g) g$ holds uniformly over all arguments. We denote by $(x)_+ = \max\{x, 0\}$.

Let P and Q be probability measures over Θ, such that P is absolutely continuous with respect to Q. For a convex function $f : [0, \infty) \to (-\infty, \infty]$ such that $f(x)$ is finite for all $x > 0$, $f(1) = 0$, and $f(0) = \lim_{x \to 0^+} f(x)$ (possibly infinite), the *f-divergence* between P, Q is defined as

$$D_f(P, Q) = \int_\Theta f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q \,.$$

We will encounter several f-divergences such as the KL-divergence $D_{\mathrm{KL}} = D_f$ with $f(x) = x \ln(x)$, the TV-distance $D_{\mathrm{TV}} = D_f$ with $f(x) = |x - 1|/2$, and the ZCP-divergence $D_{\mathrm{ZCP}} = D_f$ with $f(x) = |x - 1|\sqrt{\ln(1 + c^2|x - 1|^2)}$. We will call the KL divergence between Bernoulli distributions a 'little-kl' denoted by $\mathsf{kl}(p, q) = p \ln(p/q) + (1-p) \ln((1-p)/(1-q))$ for $p, q \in [0, 1]$.

We will also encounter a *Rényi divergence* of order $\alpha \in (0, 1) \cup (1, \infty)$, defined as $D_\alpha(P, Q) = \frac{1}{\alpha - 1} \ln \int \mathrm{d}P(\theta)^\alpha \mathrm{d}Q(\theta)^{1-\alpha}$, which has many interesting connections to f-divergences, as discussed by Van Erven and Harremos (2014). For instance, $\lim_{\alpha \to 1} D_\alpha = D_{\mathrm{KL}}$.

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 *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 $P_n = P(S)$.

3.1. Coin-betting game, regret, and Ville's inequality

The forthcoming analysis is intimately connected to the derivation of concentration inequalities through regret analysis of online gambling algorithms, following Jun and Orabona (2019); Orabona and Jun (2024). Here, we briefly introduce the required notions and definitions.

We consider a gambler playing a betting game repetitively. This gambler starts with initial wealth $\mathrm{W}_0 = 1$. In each round t, the gambler bets $|x_t|$ money on the outcome x_t, observes a 'continuous coin' outcome $c_t \in [-1, 1]$, which can even be adversarially chosen. At the end of the round, the gambler earns $x_t c_t$, so that if we define the wealth of the gambler at time t as W_t,

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

We also assume the gambler cannot use any loans in this game, meaning $\mathrm{W}_t \geq 0$ and $x_t \in [-\mathrm{W}_{t-1}, \mathrm{W}_{t-1}]$.

This coin-betting game is one of the representative online learning problems. Therefore, comparing the difference in wealth with a fixed strategy, as in online betting, is a natural objective. In particular, for each $\beta \in [-1, 1]$, let $\mathrm{W}_t(\beta)$ be the wealth when the gambler bets $\beta \, \mathrm{W}_{t-1}(\beta)$ in round t with the same initial wealth condition $\mathrm{W}_0(\beta) = 1$. So, we have

$$\mathrm{W}_t(\beta) = \mathrm{W}_{t-1}(\beta) + c_t \beta \, \mathrm{W}_{t-1}(\beta) = \prod_{s=1}^{t} (1 + c_s \beta) \, .$$

We denote $\mathrm{W}_n^* = \max_{\beta \in [-1,1]} \mathrm{W}_n(\beta)$ the maximum wealth a fixed betting strategy can achieve, and define the regret of the betting algorithm as

$$\mathrm{Regret}_n = \frac{\mathrm{W}_n^*}{\mathrm{W}_n} \, .$$

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

If the coin outcomes have conditional zero mean, it is intuitive that any online betting algorithm should not be able to gain any money. Indeed, the wealth remains 1 in expectation (i.e., $\mathbb{E}[\mathrm{W}_t] = 1$), so W_t forms a nonnegative martingale and thus follows a very useful time-uniform concentration bound known as Ville's inequality.

Theorem 1 (Ville's inequality (Ville, 1939, p. 84)) *Let $Z_1, \dots, Z_n$ be a sequence of non-negative random variables such that $\mathbb{E}[Z_i \mid Z_1, \dots, Z_{i-1}] = 0$. Let $M_t > 0$ be $\Sigma(Z_1, \dots, Z_{t-1})$-measurable such that $M_0 = 1$, and moreover assume that $\mathbb{E}[M_t \mid Z_1, \dots, Z_{t-1}] \leq M_{t-1}$. Then, for any $\delta \in (0, 1]$, $\mathbb{P}\left\{\max_t M_t \geq \frac{1}{\delta}\right\} \leq \delta$.*

Ville's inequality will be the main tool to leverage regret guarantees to construct our concentration inequalities.

4. The ZCP Divergence

Here, we look deeper into properties of the ZCP divergence defined in Eq. (2). First, note that ZCP-divergence is an f-divergence with $f(x) = |x - 1|\sqrt{\ln(1 + c^2 \, |x - 1|^2)}$ for $x \in \mathbb{R}_{\geq 0}$ and some parameter $c \geq 0$. The main interesting property of this divergence is that it is controlled simultaneously by KL-divergence and TV distance, namely:

Theorem 2 *For any pair $P, Q \in \mathcal{M}_1(\Theta)$, and any $c \geq 0$, we have*

$$D_{\mathrm{ZCP}}(P, Q; c) \leq 2\sqrt{2 \, D_{\mathrm{TV}}(P, Q) \, D_{\mathrm{KL}}(P, Q)} + 2\sqrt{\ln(2 + 2c)} \, D_{\mathrm{TV}}(P, Q) \, .$$

Note that this control only incurs an additive logarithmic cost in c (recall that $D_{\mathrm{TV}} \leq 1$). The above is a direct consequence of Lemma 3 and Lemma 4, both shown the Appendix (Section C.1 and Section C.2).

Lemma 3 *Under conditions of Theorem 8, for any $c \geq 0$,*

$$D_{\mathrm{ZCP}}(P, Q; c) \leq D_{\mathrm{ZCP}}(P, Q; 1) + 2\sqrt{\ln(2 + 2c)} \, D_{\mathrm{TV}}(P, Q) \, .$$

Lemma 4 *For any pair $P, Q \in \mathcal{M}_1(\Theta)$,*

$$D_{\mathrm{ZCP}}(P, Q; 1) \leq \sqrt{8 D_{\mathrm{TV}}(P, Q)\, D_{\mathrm{KL}}(P, Q)}\ .$$

From Theorem 2 we can see that D_{ZCP} is *orderwise tighter* than $\sqrt{D_{\mathrm{KL}}}$ in the sense that for any P, Q, we have that

$$D_{\mathrm{ZCP}}(P, Q; c) D_{\mathrm{TV}}(P, Q) \leq 2(\sqrt{2} + \sqrt{\ln(2 + 2c)})\sqrt{D_{\mathrm{KL}}(P, Q)}\ .$$

On the other hand, as we show below, there exist an infinite number of sequences of probability distributions (P_i, Q_i) such that $\lim_{i \to \infty} D_{\mathrm{KL}}(P_i, Q_i) = \infty$ and $\limsup_{i \to \infty} D_{\mathrm{ZCP}}(P_i, Q_i; c) < \infty$. Thus, D_{ZCP} is at most both a multiplicative and an additive constant away from $\sqrt{D_{\mathrm{KL}}}$ in all regimes, but the gain can be *arbitrarily large*, especially when Pinsker's inequality is not tight.

4.1. Advantage over KL Divergence in Discrete Cases

We now show that the ZCP divergence can be arbitrarily smaller than the KL divergence. As we can tell from Lemma 4, it will be enough to upper bound the product of KL divergence and TV distance to demonstrate the advantage. We first consider a basic instance of two Bernoulli random variables, with proof provided in Section C.3:

Proposition 5 *Let P and Q be Bernoulli distributions with success probabilities p and q respectively, and moreover assume that $q = p/a$ for some free parameter $a \geq 1$. Choosing:*

- *$a = e^{\frac{1}{p^2}}$, we have $D_{\mathrm{KL}}(P, Q) \cdot D_{\mathrm{TV}}(P, Q) \leq 1 - \frac{1}{e}$ while $\frac{1}{p} - \frac{1}{e} \leq D_{\mathrm{KL}}(P, Q) \leq \frac{1}{p}$;*
- *$a = e^{\frac{1}{p^{3/2}}}$, we have $D_{\mathrm{KL}}(P, Q)\sqrt{D_{\mathrm{TV}}(P, Q)} \leq \sqrt{1 - \frac{1}{e}}$ while $\frac{1}{\sqrt{p}} - \frac{1}{e} \leq D_{\mathrm{KL}}(P, Q) \leq \frac{1}{\sqrt{p}}$.*

Multivariate instances The proof of Proposition 5 can be easily extended to any pair of distributions with a finite support:

Proposition 6 *Let $P = (p_1, \ldots, p_d)$ and $Q = (q_1, \ldots, q_d)$ be probability distributions, where d is even without loss of generality, and probability weights are set as*

$$p_i = \begin{cases} p & \text{for } i \in \{1, \ldots, \frac{d}{2}\} \\ \frac{1 - \frac{pd}{2}}{d/2} & \text{for } i \in \{\frac{d}{2} + 1, \ldots, d\} \end{cases}, \quad q_i = \begin{cases} \frac{p}{a} & \text{for } i \in \{1, \ldots, \frac{d}{2}\} \\ \frac{1 - \frac{pd}{2a}}{d/2} & \text{for } i \in \{\frac{d}{2} + 1, \ldots, d\} \end{cases}$$

where $u > 0$ is a free parameter, $p = d^{-1-u}$, and $a = \exp(d^{\frac{3}{2}u})$. Then, for a sufficiently large d,

$$\begin{aligned} D_{\mathrm{KL}}(P, Q) &= \Theta\left(d\, p \ln(a)\right) = \Theta(d^{\frac{u}{2}})\ , \\ D_{\mathrm{TV}}(P, Q) &= \Theta\left(d\, p\right) = \Theta(d^{-u})\ , \\ D_{\mathrm{ZCP}}(P, Q) &= \Theta\left(d\, p \sqrt{\ln(a)}\right) = \Theta(d^{-\frac{1}{4}u})\ , \\ \sqrt{D_{\mathrm{KL}}(P, Q)} \cdot D_{\mathrm{ZCP}}(P, Q) &= \Theta(1)\ . \end{aligned}$$

Observe that $D_{\mathrm{ZCP}}(P, Q) = \Theta(d^{-\frac{1}{4}u})$ is strictly smaller than $D_{\mathrm{KL}}(P, Q) = \Theta(d^{\frac{u}{2}})$. Moreover, one could also check that $\sqrt{D_{\mathrm{KL}}} \cdot D_{\mathrm{ZCP}} = \Theta(1)$ while $D_{\mathrm{KL}}(P, Q) = \Theta(d^{\frac{u}{2}})$.

4.2. Advantage over KL Divergence in the Mixture of Gaussian Case

Now we consider a continuous case. In particular, we have the following Gaussian instance, with proof provided in Section C.4.

Proposition 7 *Let $A = \mathcal{N}(\mu, \sigma_1^2)$ and $B = \mathcal{N}(\mu, \sigma_2^2)$. Let P be a Gaussian mixture $P = pA + (1-p)B$ for $p \in [0,1]$ and let $Q = B$.*

- *Choosing $\frac{\sigma_2}{\sigma_1} = p$, we have $D_{\mathrm{KL}}(P, Q) \geq \frac{1}{2p} - 1.3$, while $D_{\mathrm{TV}}(P, Q) \cdot D_{\mathrm{KL}}(P, Q) \leq \frac{1}{2}$.*
- *Choosing $\frac{\sigma_2}{\sigma_1} = p^{\frac{3}{4}}$, we have $D_{\mathrm{KL}}(P, Q) \geq \frac{1}{2\sqrt{p}} - 1.22$, while $D_{\mathrm{KL}}(P, Q) \cdot \sqrt{D_{\mathrm{TV}}(P, Q)} \leq \frac{1}{2}$.*

5. Main results

We first present a Hoeffding-type inequality that involves D_{ZCP}, which is proved in Section 7.

Theorem 8 (Hoeffding-type ZCP inequality) *Let $\delta \in (0,1)$. Then, for any $P_0 \in \mathcal{M}_1(\Theta)$, with probability at least $1-2\delta$, simultaneously for all $n \in \mathbb{N}, P_n \in \mathcal{K}(\mathcal{X}^n, \Theta)$, we have*

$$\int \Delta_n(\theta) \, \mathrm{d}P_n(\theta) \leq \frac{\sqrt{2}\, D_{\mathrm{ZCP}}\left(P_n, P_0; \frac{\sqrt{2n}}{\delta}\right) + 2 + 2\sqrt{\ln \frac{2\sqrt{n}}{\delta}}}{\sqrt{n}} \, .$$

As shown in Eq. (4) the bound is orderwise never worse than the classical KL-based one and in Section 4 we show instances where thanks to ZCP divergence it enjoys an improved order. Moreover, Theorem 8 holds *uniformly* over $n \in \mathbb{N}$ unlike most classical bounds which only hold for a fixed n.

Remark 9 *It is possible to obtain a similar inequality by combining the regret guarantee in Zhang et al. (2022) and the recently proposed online-to-PAC framework of Lugosi and Neu (2023) that obtains PAC bounds from the regret of online learning algorithms. Both approaches are valid and we believe both proof methods have their distinct advantages. In particular, we believe that our proof method is more direct and more flexible. Indeed, we show below how to bound the integral of the log optimal wealth, a case that is not covered by the framework in Lugosi and Neu (2023) and that allows to recover various known types of inequalities such as the 'little-kl' and the empirical Bernstein's inequality.*

Next, we demonstrate a generalized inequality, which extends beyond the Hoeffding regime of $1/\sqrt{n}$ rate. In Section 3.1 we introduced a notion of max-wealth of a betting algorithm. To state the following result, we parameterize the max-wealth by θ:

$$\mathrm{W}_n^*(\theta) = \max_{\beta \in [-1,1]} \prod_{i=1}^{n} \big(1 + \beta(f(\theta, X_i) - \mathbb{E}[f(\theta, X_1)])\big) \, .$$

The central quantity in the coming result will be the expected *maximal log-wealth* $\int \ln \mathrm{W}_n^*(\theta) \, \mathrm{d}P_n(\theta)$ — it was recently shown by Jang et al. (2023) that through lower-bounding $\ln \mathrm{W}_n^*$ term we can obtain many known PAC-Bayes bounds. To this end, our second main result, shown in Section A, gives a bound on the expected maximal log-wealth:

Theorem 10 (Log-wealth ZCP inequality) *Let $\delta \in (0,1)$. Then, for any $\alpha \in (0,1)$ and for any $P_0 \in \mathcal{M}_1(\Theta)$, with probability at least $1-2\delta$, simultaneously over $n \in \mathbb{N} \setminus \{1\}, P_n \in \mathcal{K}(\mathcal{X}^n, \Theta)$,*

$$\int \ln \mathrm{W}_n^*(\theta) \mathrm{d}P_n(\theta) \le \frac{1}{\sqrt{2}} \sqrt{\ln\left(\frac{4n^2}{\delta}\right) + \frac{\alpha}{\alpha-1} \ln(n) + D_\alpha(P_n, P_0)}\, D_{\mathrm{ZCP}}\Big(P_n, P_0;\ \frac{\sqrt{2}\, n^{2.5}}{\delta}\Big) + \ln\left(2e^2\sqrt{n}\left(1+\frac{4n^2}{\delta}\right)\right) + \frac{\delta}{n(n+1)}\ .$$

Note that in addition to ZCP-divergence, now the bound now also depends on the Rényi divergence $D_\alpha()$ and its order α can be choosen freely. In particular, in the next corollary we show that asymptotically, when α is tuned based on n, the Rényi divergence can be replaced by the KL divergence.

Corollary 11 *Fix $P_0 \in \mathcal{M}_1(\Theta)$. Then, under the conditions of Theorem 10, with probability one for all $P \in \mathcal{M}_1(\Theta)$,*

$$\limsup_{n\to\infty} \frac{\int \ln \mathrm{W}_n^*(\theta)\mathrm{d}P(\theta)}{\sqrt{2\ln(n)\ln(en)\ln(1+\sqrt{2}\, n^{4.5})}} \le 2 + \big(2+\sqrt{D_{\mathrm{KL}}(P,P_0)}\big)\big(D_{\mathrm{ZCP}}(P,P_0;1) + D_{\mathrm{TV}}(P,P_0)\big)\ .$$

A simple consequence of the above, when combined with Lemma 4, is that there exists an absolute constant $c > 0$ such that with probability one

$$\limsup_{n\to\infty} \frac{\int \ln \mathrm{W}_n^*(\theta)\mathrm{d}P(\theta)}{\ln^{3/2}(n)} \le c\left(1+\sqrt{D_{\mathrm{KL}}(P,P_0)\, D_{\mathrm{TV}}(P,P_0)}\left(1+\sqrt{D_{\mathrm{KL}}(P,P_0)}\right)\right)\ . \tag{5}$$

In comparison, Jang et al. (2023) obtained the bound

$$\int \ln \mathrm{W}_n^*(\theta)\,\mathrm{d}P_n(\theta) = \mathcal{O}\left(D_{\mathrm{KL}}(P_n,P_0) + \ln\frac{n}{\delta}\right),$$

which is looser than the bound in Eq. (5) in terms of dependence on divergence terms, since $D_{\mathrm{KL}}(P_n,P_0)$ is orderwise at least $D_{\mathrm{KL}}(P_n,P_0)\sqrt{D_{\mathrm{TV}}(P_n,P_0)} \ge \sqrt{D_{\mathrm{KL}}(P_n,P_0)}D_{\mathrm{ZCP}}$.

Remark 12 *Comparing this result to our Hoeffding-style bound (Theorem 8) is nontrivial since the left-hand side is written in a different form. To make a clear comparison, we defer this discussion to remark 15 below, but in summary our optimal log wealth bound leads to a factor of $D_{\mathrm{TV}}^{1/4}(P_n,P_0)$ looser one compared to Theorem 8.*

Remark 13 *The proof of Theorem 10 is highly non-trivial and there are a few approaches one might think of that fail. For example, we can successfully upper bound $\int \sqrt{\ln \mathrm{W}_n^*(\theta)}\,\mathrm{d}P_n(\theta)$ but then we cannot obtain an empirical Bernstein inequality because we need a bound like $\int \sqrt{\ln \mathrm{W}_n^*(\theta)}\,\mathrm{d}P_n(\theta) \ge \sqrt{\int \ln \mathrm{W}_n^*(\theta)\,\mathrm{d}P_n(\theta)}$ yet this inequality is the opposite direction of Jensen's inequality. Alternatively, we can upper bound $\int (n\Delta_n(\theta))^2\,\mathrm{d}P_n(\theta)$ but then the empirical variance terms will appear integrated with respect to the prior instead of the posterior.*

5.1. Recovering variants of other known bounds with new divergence

As we anticipated, inequality in Theorem 10 can be relaxed to obtain various known PAC-Bayes bounds. First, this was observed by Jang et al. (2023), who derived a bound on $\int \ln \mathrm{W}_n^*(\theta)\,\mathrm{d}P_n(\theta)$ featuring KL-divergence. By applying similar lower bounding techniques as in their work, in the following we state ZCP versions of known PAC-Bayes bounds. Let the complexity term be

$$\mathrm{Comp}_n(\alpha) = (\text{r.h.s. in Theorem 10}) \lesssim \sqrt{\frac{\alpha}{\alpha-1} + D_\alpha\, D_{\mathrm{ZCP}} + 1}\ .$$

By following the same reasoning as in Corollary 11 one can state asymptotic versions of the relaxed bounds, which for a universal constant $c > 0$, will manifest

$$\frac{\mathrm{Comp}_n\left(1+\frac{1}{\ln(n)}\right)}{\ln^{3/2}(n)} \to c\left(1+\sqrt{D_{\mathrm{KL}}\,D_{\mathrm{TV}}}\left(1+\sqrt{D_{\mathrm{KL}}}\right)\right) \qquad \text{as} \qquad n\to\infty\ .$$

Empirical Bernstein inequality First, we recover an empirical Bernstein-type inequality (Tolstikhin and Seldin, 2013), where the bound switches to the 'fast rate' $1/n$ when the sample variance is sufficiently small. In particular, in Corollary 14 we show:

Corollary 14 *Under the conditions of Theorem 10, for any $\alpha \in (0,1)$, we have, with probability at least $1-2\delta$, simultaneously over every $n \in \mathbb{N}\setminus\{1\}$ and $P_n \in \mathcal{K}(\mathcal{X}^n,\Theta)$,*

$$\left|\int \Delta_n(\theta)\,\mathrm{d}P_n\right| \le \frac{\sqrt{2\,\mathrm{Comp}_n(\alpha)\,\hat{V}(P_n)}}{(\sqrt{n}-(2/\sqrt{n})\,\mathrm{Comp}_n(\alpha))_+} + \frac{2\mathrm{Comp}_n(\alpha)}{\left(n-2\,\mathrm{Comp}_n(\alpha)\right)_+},$$

where $\hat{V}(P) = \frac{1}{n(n-1)}\sum_{i<j}\int (f(\theta,X_i)-f(\theta,X_j))^2\,\mathrm{d}P(\theta)$ is the expected sample variance.

Furthermore, there exists an absolute constant $c > 0$ such that with probability one, for all $P \in \mathcal{M}_1(\Theta)$,

$$\limsup_{n\to\infty} \frac{\left|\int \Delta_n(\theta)\,\mathrm{d}P\right|^2}{\frac{1}{n}\left(\left|\int \Delta_n(\theta)\,\mathrm{d}P\right| + \hat{V}(P)\right)\cdot \ln^{3/2}(n)} \le c\left(1+\sqrt{D_{\mathrm{KL}}\,D_{\mathrm{TV}}}\,(1+\sqrt{D_{\mathrm{KL}}})\right)\ .$$

We defer the proof to Section C.5.

Remark 15 *When the sample variance is sufficiently large, that is larger than $\left|\int \Delta_n(\theta)\,\mathrm{d}P_n\right|$, the bound above provides a comparison point with our own Hoeffding style bound (Theorem 8) w.r.t. the complexity term, which scales with $D_{\mathrm{ZCP}}(P_n,P_0)$. Ignoring the fact that the bound above is asymptotic, we note that the bound in Corollary 14 scales with $\sqrt{D_{\mathrm{KL}}\,D_{\mathrm{TV}}}\cdot\sqrt{D_{\mathrm{KL}}}$, and so*

$$\sqrt{D_{\mathrm{KL}}\,D_{\mathrm{TV}}}\cdot\sqrt{D_{\mathrm{KL}}} \ge \sqrt{D_{\mathrm{KL}}\,D_{\mathrm{TV}}}\cdot\sqrt{D_{\mathrm{KL}}\,D_{\mathrm{TV}}} \ge D_{\mathrm{ZCP}}\ .$$

Note that there is a factor of $D_{\mathrm{TV}}^{1/2}$ gap. Investigating if it can be removed is left as future work.

Bernoulli KL-divergence (Langford-Caruana / Seeger / Maurer) bound Finally, we consider a tighter inequality for the specific case of Bernoulli f (e.g., Bernoulli losses). This is a well-studied setting in the PAC-Bayes literature (Langford and Caruana, 2001; Seeger, 2002; Maurer, 2004). In this case we are bounding the KL divergence between Bernoulli distributions denoted by $\mathsf{kl}()$. One can observe that the bound on $\mathsf{kl}()$ is tighter than Hoeffding-type bounds due to Pinsker's inequality. In such a case, denoting sample average and mean respectively as $\hat{p}_\theta = (f(\theta, X_1) + \dots + f(\theta, X_n))/n$ and $p_\theta = \mathbb{E}[f(\theta, X_1)]$, we have

Corollary 16 *Under the conditions of Theorem 10, for any $\alpha \in (0,1)$, we have, with probability at least $1 - 2\delta$, simultaneously over every $n \in \mathbb{N} \setminus \{1\}$ and $P_n \in \mathcal{K}(\mathcal{X}^n, \Theta)$,*

$$\mathsf{kl}\left(\int \hat{p}_\theta \, \mathrm{d}P_n(\theta), \int p_\theta \, \mathrm{d}P_n(\theta)\right) \leq \frac{\mathrm{Comp}_n(\alpha)}{n} .$$

Furthermore, there exists an absolute constant $c > 0$ such that with probability one, for all $P \in \mathcal{M}_1(\Theta)$,

$$\limsup_{n \to \infty} \frac{\mathsf{kl}\left(\int \hat{p}_\theta \, \mathrm{d}P(\theta), \int p_\theta \, \mathrm{d}P(\theta)\right)}{\frac{1}{n} \cdot \ln^{3/2}(n)} \leq c \left(1 + \sqrt{D_{\mathrm{KL}} \, D_{\mathrm{TV}}} \left(1 + \sqrt{D_{\mathrm{KL}}}\right)\right) .$$

The proof of Corollary 16 closely follows that of Jang et al. (2023, Proposition 3 in Sec. A.2).

6. Conclusions

In this paper we derived, to the best of our knowledge, a first high-probability PAC-Bayes bound with the ZCP divergence. This divergence is never worse than the KL divergence orderwise and it enjoys a strictly better scaling in some instances. In the concentration regime $1/\sqrt{n}$ for the deviation $\int \Delta_n(\theta) \, \mathrm{d}P_n(\theta)$, the new bound scales with $D_{\mathrm{ZCP}} \lesssim \sqrt{D_{\mathrm{KL}} \, D_{\mathrm{TV}}}$ instead of $\lesssim \sqrt{D_{\mathrm{KL}}}$. In other regimes, such as the Bernstein regime, the bound asymptotically scales with $\sqrt{\sqrt{D_{\mathrm{KL}}} \, D_{\mathrm{ZCP}}}$, which can be analyzed to be a factor of $D_{\mathrm{TV}}^{1/4}$ worse than our Hoeffding-style bound. Both proofs rely on a novel change-of-measure argument with respect to $x \mapsto e^{x^2/2}$ potential which might be of an independent interest.

A tantalizing open problem is whether our bounds can be further improved. It would be interesting to see if it is possible to establish some (Pareto) optimalities for PAC-Bayes bounds.

7. Proof of Theorem 8: McAllester/Hoeffding-type bound

First, we need the following lemmas.

Lemma 17 (Orabona (2019, Lemma 9.7)) *Let $F(x) = b \, e^{x^2/(2a)}$ for $a, b > 0$ and let $F^\star$ be its convex conjugate. Then, we have $F^\star(y) \leq |y| \sqrt{a \ln\left(1 + \frac{ay^2}{b^2}\right)} - b$.*

Theorem 18 (Cesa-Bianchi and Lugosi (2006)) *For any sequence of $c_i \in [-1, 1]$, there exists an online algorithm that selects in $\beta_i \in [-1, 1]$ with knowledge of $c_1, \dots, c_{t-1}$ such that for all n it uniformly guarantees*

$$\prod_{i=1}^{n} (1 + \beta_i c_i) \geq \frac{1}{\sqrt{2(n+1)}} \max_{\beta \in [-1,1]} \prod_{i=1}^{n} (1 + \beta c_i) .$$

The following lemma is shown in the appendix.

Lemma 19 *For any $c_1, \dots, c_n \in [-1,1]$, we have*

$$\max_{\beta \in [-1,1]} \prod_{i=1}^n (1 + \beta c_i) \geq \exp\left(\frac{\left(\sum_{i=1}^n c_i\right)^2}{4n} \right) .$$

Finally, we will need a time-uniform version of Hoeffding's inequality (Orabona and Jun, 2024), (Jang et al., 2023, Sec. A.1), (Orabona, 2019, Sec. 12.7):

Proposition 20 *Let $Y_1, \dots, Y_n \in [0,1]$ be independent random variables. Then, for any $\delta \in (0,1]$, with probability at least $1-\delta$, for all $n \in \mathbb{N}$ simultaneously,*

$$\left| \sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) \right| \leq 2\sqrt{n \ln \frac{2\sqrt{n}}{\delta}} .$$

Now we proceed with the proof of Theorem 8. Throughout the proof it will be convenient to work with unnormalized gap (instead of normalized one, Δ):

$$\bar{\Delta}(\theta) = \sum_{i=1}^n \bar{\Delta}(\theta, i) , \qquad \bar{\Delta}(\theta, i) = f(\theta, X_i) - \mathbb{E}[f(\theta, X_1)] .$$

Consider a change-of-measure decomposition

$$\int \bar{\Delta}(\theta) \, \mathrm{d}P_n(\theta) \leq \underbrace{\int \bar{\Delta}(\theta) \cdot \left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1 \right) \mathrm{d}P_0(\theta)}_{(i)} + \underbrace{\int \bar{\Delta}(\theta) \, \mathrm{d}P_0(\theta)}_{(ii)}$$

and note right away that by the fact that term (ii) can be controlled by Proposition 20 since $\int (f(\theta, X_i) \, \mathrm{d}P_0 \in [0,1]$. Namely, with probability at least $1-\delta$, simultaneously for all $n \in \mathbb{N}$,

$$(ii) = \int \bar{\Delta}(\theta) \, \mathrm{d}P_0(\theta) \leq 2\sqrt{n \ln \frac{2\sqrt{n}}{\delta}} .$$

We turn our attention to term (i). By Fenchel-Young inequality, for a convex conjugate $F^\star : \mathbb{R} \to \mathbb{R}$,

$$(i) \leq \int F(\bar{\Delta}(\theta)) \, \mathrm{d}P_0(\theta) + \int F^\star \left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1 \right) \mathrm{d}P_0(\theta) . \tag{6}$$

Now, we will make a particular choice of $F()$ by using Lemma 17, choosing $a = 2n$, and leaving b to be tuned later: $F(\bar{\Delta}(\theta)) = b \exp\left(\frac{\bar{\Delta}^2(\theta)}{4n} \right)$. Throughout the rest of the proof we will control the above. In particular, we make a key observation that the above term is controlled by the *maximal wealth* achievable by some online algorithm and using its regret bound we can argue that $\int F(\bar{\Delta}^2(\theta)) \, \mathrm{d}P_0(\theta)$ concentrates well. In particular, we will need Lemma 19 (shown in the appendix), which is then combined with the regret bound of Theorem 18. Let $B_{i-1}(\theta)$ be a prediction of a betting algorithm after observing $\bar{\Delta}(\theta, 1), \dots \bar{\Delta}(\theta, i-1)$, and define its wealth at step n as

$$\mathrm{W}_n(\theta) = \prod_{i=1}^n (1 + B_{i-1}(\theta) \, \bar{\Delta}(\theta, i)) . \tag{7}$$

Then, wealth is related to the max-wealth through Theorem 18, $\mathrm{W}^*(\theta) \leq \sqrt{2(n+1)}\ \mathrm{W}_n(\theta)$. The final bit to note that $\int W_n(\theta)\,\mathrm{d}P_0(\theta)$ is a martingale, that is[5]

$$\mathbb{E}_{n-1} \int W_n(\theta)\,\mathrm{d}P_0(\theta) = \int \mathbb{E}_{n-1} W_n(\theta)\,\mathrm{d}P_0(\theta) = \int W_{n-1}(\theta)\,\mathrm{d}P_0(\theta), \tag{8}$$

where we used Fubini's theorem to exchange $\mathbb{E}$ and $\int$. This fact allows us to use Ville's inequality (Theorem 1). So, we obtain

$$\begin{aligned} \int F(\bar{\Delta}(\theta))\,\mathrm{d}P_0(\theta) &\leq b \int \exp\left(\sum_{i=1}^n \ln(1 + B_{i-1}(\theta)\,\bar{\Delta}(\theta, i)) + \ln\sqrt{2(n+1)} \right) \mathrm{d}P_0(\theta) \\ &= b\sqrt{2(n+1)} \int W_n(\theta)\,\mathrm{d}P_0(\theta) \leq b\,\frac{\sqrt{2(n+1)}}{\delta} \qquad \text{(By Ville's inequality)} \\ &= \sqrt{2(n+1)} \leq 2\sqrt{n}\,. \qquad \text{(Tuning } b = \delta) \end{aligned}$$

That said, using Lemma 17 and the above provide a bound on Eq. (6) that is

$$\begin{aligned} (i) &\leq 2\sqrt{n} + \sqrt{2n} \int \left| \frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1 \right| \sqrt{\ln\left(1 + \frac{2n}{\delta^2}\left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1\right)^2\right)}\,\mathrm{d}P_0(\theta) - \delta \\ &= 2\sqrt{n} + \sqrt{2n}\, D_{\mathrm{ZCP}}\left(P_n, P_0; \frac{\sqrt{2n}}{\delta}\right) - \delta\,. \end{aligned}$$

Completing the bound and dividing it though by n completes the proof. ■

Acknowledgements

Kwang-Sung Jun was supported in part by National Science Foundation under grant CCF-2327013. The authors would like to thank András György for providing valuable feedback on an early draft of this paper.

References

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

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

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.

[5] We use notation $\mathbb{E}_{n-1}[\cdot] = \mathbb{E}[\cdot \mid X_1, \ldots, X_{n-1}]$.

L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy. PAC-Bayesian bounds based on the rényi divergence. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 435–444. PMLR, 2016.

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.

N. Cesa-Bianchi and F. Orabona. Online learning algorithms. *Annual review of statistics and its application*, 8:165–190, 2021.

K. Chaudhuri, Y. Freund, and D. J. Hsu. A parameter-free hedging algorithm. *Advances in Neural Information Processing Systems*, 22, 2009.

Y. Chu and M. Raginsky. A unified framework for information-theoretic generalization bounds. In *Advances in Neural Information Processing Systems*, 2023.

T. M Cover. Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin. Technical Report 12, Department of Statistics, Stanford University, 1974. URL https://purl.stanford.edu/js411qm9805.

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.

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.

G. K. Dziugaite, K. Hsu, W. Gharbieh, G. Arpino, and D. M. Roy. On the role of data in PAC-Bayes bounds. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

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

M. Haddouche and B. Guedj. PAC-Bayes with unbounded losses through supermartingales. *Transactions on Machine Learning Research (TMLR)*, 2023.

K. Jang, K.-S. Jun, I. Kuzborskij, and F. Orabona. Tighter PAC-Bayes bounds through coin-betting. In *Conference on Computational Learning Theory (COLT)*, 2023.

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.

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.

W. M. Koolen and T. Van Erven. Second-order quantile methods for experts and combinatorial games. In *Conference on Computational Learning Theory (COLT)*, pages 1155–1175. PMLR, 2015.

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

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

G. Lugosi and G. Neu. Online-to-PAC conversions: Generalization bounds via regret analysis. arXiv:2305.19674, 2023.

H. Luo and R. E. Schapire. Achieving all with no parameters: AdaNormalHedge. In *Proc. of the Conference on Learning Theory (COLT)*, pages 1286–1304, 2015.

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

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.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11): 5847–5861, 2010.

F. Nielsen and K. Sun. Guaranteed deterministic bounds on the total variation distance between univariate mixtures. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2018.

Y. Ohnishi and J. Honorio. Novel change of measure inequalities with applications to PAC-Bayesian bounds and Monte Carlo estimation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1711–1719. PMLR, 2021.

F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019. URL https://arxiv.org/abs/1912.13213. Version 6.

F. Orabona and K.-S. Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Transactions on Information Theory*, 70(1):436–455, 2024.

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.

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.

A. Ruderman, M. Reid, D. García-García, and J. Petterson. Tighter variational representations of f-divergences via restriction to probability measures. *International Conference on Machine Learing (ICML)*, 2012.

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.

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.

T. Van Erven and P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

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

J. Yang, S. Sun, and D. M. Roy. Fast-rate PAC-Bayes generalization bounds via shifted Rademacher processes. *Advances in Neural Information Processing Systems*, 32, 2019.

Z. Zhang, A. Cutkosky, and Y. Paschalidis. Optimal comparator adaptive online learning with switching cost. *Advances in Neural Information Processing Systems*, 35:23936–23950, 2022.

Appendix A. Proof of Theorem 10: Bound on log-wealth

Recall that $\mathrm{W}_n(\theta)$ is wealth of a betting algorithm as defined in Eq. (7). In the following we will work with its truncated counterpart, that is $\mathrm{W}_n^1(\theta) = 1 \vee \mathrm{W}_n(\theta)$ which satisfies $\mathrm{W}_n^1(\theta) \leq 1 + \mathrm{W}_n(\theta)$.

The proof is similar to that of Theorem 8, however it has a different structure than the usual PAC-Bayes bounds and we believe it might be interesting on its own. Here, we are working with $\ln \mathrm{W}_n^1(\theta)$ instead of $\Delta(\theta)$. As in the proof of Theorem 8 we apply Fenchel-Young, but with a family of functions, $F_\theta(x) = \exp(\frac{x^2}{2a})$, and we will choose a different a for each θ, that is, $a = \ln \mathrm{W}_n^1(\theta)/2$, which will lead to the term $\ln \mathrm{W}_n^1(\theta)$ as a part of the divergence. Then, the main idea of this proof is to control the $\ln \mathrm{W}_n^1(\theta)$ term in the divergence using the fact that $\ln \mathrm{W}_n^1(\theta)$ is often small. Therefore we condition on the event when it is small and the remaining part we control pessimistically, i.e., $\ln \mathrm{W}^2 \leq n$. Lemma 21 tells us when $\ln \mathrm{W}$ is small.

Lemma 21 *Let $Q \in \mathcal{M}_1(\Theta)$ be independent from data. Then, for any $\delta \in (0,1)$ and any $t > 0$,*

$$\mathbb{P}\left\{\mathbb{P}_Q\left(\mathrm{W}_n^1(\theta) \geq t\right) < \frac{2}{t\delta}\right\} \geq 1 - \delta\,.$$

Proof For every $\theta \in \Theta$, Markov's inequality implies that

$$\mathbb{P}_Q\{\mathrm{W}_n^1(\theta) \geq t\} \leq \frac{\mathbb{E}_Q \mathrm{W}_n^1(\theta)}{t}\,.$$

Furthermore, another application of Markov's inequality (with respect to data) implies that

$$\mathbb{P}\left\{\mathbb{E}_Q \mathrm{W}_n^1(\theta) \geq \frac{2}{\delta}\right\} \leq \frac{\delta}{2} \mathbb{E}\,\mathbb{E}_Q \mathrm{W}_n^1(\theta) \leq \frac{\delta}{2} \mathbb{E}\,\mathbb{E}_Q[1 + \mathrm{W}_n(\theta)] = \delta,$$

where we have used the fact that $\mathbb{E}\,\mathbb{E}_Q \mathrm{W}_n(\theta) = \mathbb{E}_Q \mathbb{E} \mathrm{W}_n(\theta)$ by Fubini's theorem and $\mathbb{E} \mathrm{W}_n(\theta) = 1$ (see Proposition 22). We conclude the proof by chaining the two displays above:

$$1 - \delta \leq \mathbb{P}\left\{\mathbb{E}_Q \mathrm{W}_n^1(\theta) < \frac{2}{\delta}\right\} \leq \mathbb{P}\left\{t\,\mathbb{P}_Q(\mathrm{W}_n^1(\theta) \geq t) < \frac{2}{\delta}\right\}\,.$$

■

Proposition 22 *For a fixed θ, $\mathbb{E} \mathrm{W}_n(\theta) = 1$.*

Proof Denote $\mathrm{W}_n \equiv \mathrm{W}_n(\theta)$ and note that $\mathrm{W}_n = \prod_{i=1}^n (1 + B_{i-1} Z_i) = (1 + B_{n-1} Z_n) \mathrm{W}_{n-1}$ where $(Z_i)_{i=1}^n$ are zero-mean independent random variables and a random variable B_{i-1} depends only on $(Z_j)_{j=1}^{i-1}$. Hence, $\mathbb{E}[\mathrm{W}_n] = \mathbb{E}[1 + B_{n-1} Z_n \mid Z_1, \dots, Z_{n-1}] \mathbb{E}[\mathrm{W}_{n-1}] = \mathbb{E}[\mathrm{W}_{n-1}] = \cdots = 1$.

■

Consider the following decomposition w.r.t. a free parameter $t > 0$ to be tuned later:

$$\begin{aligned}
\int \ln \mathrm{W}_n(\theta) \,\mathrm{d}P_n(\theta) \quad &\leq \int \ln \mathrm{W}_n^1(\theta) \,\mathrm{d}P_n(\theta) \\
&= \int \ln(\mathrm{W}_n^1(\theta)) \cdot \left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta)\right) \mathrm{d}P_0(\theta) \\
&= \int \ln(\mathrm{W}_n^1(\theta)) \cdot \left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta)\right) \mathbb{1}\left\{\mathrm{W}_n^1(\theta) > t\right\} \mathrm{d}P_0(\theta) \\
&\quad + \int \ln(\mathrm{W}_n^1(\theta)) \cdot \left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta)\right) \mathbb{1}\left\{\mathrm{W}_n^1(\theta) \leq t\right\} \mathrm{d}P_0(\theta) \\
&\leq \underbrace{n \int \left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta)\right) \mathbb{1}\left\{\mathrm{W}_n^1(\theta) > t\right\} \mathrm{d}P_0(\theta)}_{(i)} \\
&\quad + \underbrace{\int \ln(\mathrm{W}_n^1(\theta)) \cdot \left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1\right) \mathbb{1}\left\{\mathrm{W}_n^1(\theta) \leq t\right\} \mathrm{d}P_0(\theta)}_{(ii)} \\
&\quad + \underbrace{\int \ln(\mathrm{W}_n^1(\theta))\, \mathbb{1}\left\{\mathrm{W}_n^1(\theta) \leq t\right\} \mathrm{d}P_0(\theta),}_{(iii)}
\end{aligned}$$

where to get (i) we have upper bounded $\ln \mathrm{W}_n^1(\theta)$ with the pessimistic upper bound $\ln(1+2^n) \leq n$.

Bounding (iii). We note right away that by the fact that $\int \mathrm{W}_n(\theta) \,\mathrm{d}P_0$ is a martingale (see Eq. (8)), we can use Ville's inequality (Theorem 1) to have

$$\begin{aligned}
(iii) \leq \int \ln \mathrm{W}_n^1(\theta) \,\mathrm{d}P_0(\theta) &\leq \ln\left(\int \mathrm{W}_n^1(\theta) \,\mathrm{d}P_0(\theta)\right) && \text{(Jensen's inequality)} \\
&\leq \ln\left(1 + \int \mathrm{W}_n(\theta) \,\mathrm{d}P_0(\theta)\right) \\
&\leq \ln\left(1 + \frac{1}{\delta}\right) . && \text{(Ville's inequality; w.p. } \geq 1-\delta)
\end{aligned}$$

Now we handle remaining terms (i) and (ii).

Bounding (i). For the term (i), we have

$$\begin{aligned}
(i) &= n \int \left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta)\right) \mathbb{1}\left\{\mathrm{W}_n^1(\theta) > t\right\} \mathrm{d}P_0(\theta) \\
&\overset{(a)}{\leq} n \left(\mathbb{P}_{P_0}\left(\mathrm{W}_n^1(\theta) > t\right)\right)^{1-\frac{1}{\alpha}} \underbrace{\left(\int \left|\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta)\right|^{\alpha} \mathrm{d}P_0(\theta)\right)^{\frac{1}{\alpha}}}_{I_\alpha} && (\alpha > 1) \\
&\leq \frac{n I_\alpha}{(t\delta/2)^{1-\frac{1}{\alpha}}}, && \text{(By Lemma 21; w.p. } \geq 1-\delta)
\end{aligned}$$

where step (a) comes by Hölder's inequality.

Bounding (ii). By Fenchel-Young inequality, for a family of convex $F_\theta^\star : \mathbb{R} \to \mathbb{R}$,

$$\begin{aligned}(ii) &= \int \ln(\mathrm{W}_n^1(\theta)) \cdot \left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1\right) \mathbb{1}\Big\{\mathrm{W}_n^1(\theta) \le t\Big\} \,\mathrm{d}P_0(\theta) \\ &\le \int F_\theta(\ln(\mathrm{W}_n^1(\theta))) \,\mathrm{d}P_0(\theta) + \int F_\theta^\star \left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1\right) \mathbb{1}\Big\{\mathrm{W}_n^1(\theta) \le t\Big\} \,\mathrm{d}P_0(\theta) \,.\end{aligned}$$

We use Lemma 17, choosing $a = \ln(\mathrm{W}_n^1(\theta))/2$ and $b = \delta$ to have

$$F_\theta(\ln(\mathrm{W}_n^1(\theta))) = \delta \, \exp\left(\frac{\ln^2(\mathrm{W}_n^1(\theta))}{\ln(\mathrm{W}_n^1(\theta))}\right) = \delta \; \mathrm{W}_n^1(\theta)$$

and so

$$\int F_\theta(\ln(\mathrm{W}_n^1(\theta))) \,\mathrm{d}P_0(\theta) = \delta \int \mathrm{W}_n^1(\theta) \,\mathrm{d}P_0(\theta) \le \delta + \delta \int \mathrm{W}_n(\theta) \,\mathrm{d}P_0(\theta) \le 1 + \delta,$$

where the last inequality is a consequence of Ville's inequality. Finally, we complete the bound on (ii) by having a chain of inequalities on

$$\begin{aligned}&\int F^\star \left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1\right) \mathbb{1}\Big\{\mathrm{W}_n^1(\theta) \le t\Big\} \,\mathrm{d}P_0(\theta) \\ &\overset{(b)}{\le} \int \left|\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1\right| \sqrt{\frac{1}{2} \ln(\mathrm{W}_n^1(\theta)) \ln\left(1 + \frac{\ln(\mathrm{W}_n^1(\theta))}{2\delta^2} \, \left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1\right)^2\right)} \, \mathbb{1}\Big\{\mathrm{W}_n^1(\theta) \le t\Big\} \,\mathrm{d}P_0(\theta) \\ &\le \sqrt{\frac{\ln(t)}{2}} \int \left|\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1\right| \sqrt{\ln\left(1 + \frac{\ln(\mathrm{W}_n^1(\theta))}{2\delta^2} \, \left(\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(\theta) - 1\right)^2\right)} \,\mathrm{d}P_0(\theta) \\ &\le \sqrt{\frac{\ln(t)}{2}} \, D_{\mathrm{ZCP}}\Big(P_n, P_0; \sqrt{\frac{n}{2\delta^2}}\Big),\end{aligned}$$

where (b) comes by Lemma 17 and where we once more used a bound $\ln \mathrm{W}_n^1(\theta) \le n$.

Tuning t and completing the proof. Putting all together, we have

$$\int \ln \mathrm{W}_n(\theta) \,\mathrm{d}P_n(\theta) \le \sqrt{\frac{\ln(t)}{2}} \, D_{\mathrm{ZCP}}\Big(P_n, P_0; \sqrt{\frac{n}{2\delta^2}}\Big) + \frac{nI_\alpha}{(t\delta/2)^{1-\frac{1}{\alpha}}} + \ln\left(1 + \frac{1}{\delta}\right) + 1 + \delta \,.$$

and setting $t = \frac{2}{\delta}\,(nI_\alpha)^{\frac{1}{1-\frac{1}{\alpha}}}$ we obtain

$$\begin{aligned}&\int \ln \mathrm{W}_n(\theta) \,\mathrm{d}P_n(\theta) \\ &\le \frac{1}{\sqrt{2}} \sqrt{\ln\left(\frac{2}{\delta}\right) + \frac{\alpha}{\alpha - 1} \, \ln(nI_\alpha)} \, D_{\mathrm{ZCP}}\Big(P_n, P_0; \sqrt{\frac{n}{2\delta^2}}\Big) + \ln\left(1 + \frac{1}{\delta}\right) + 2 + \delta \\ &\le \frac{1}{\sqrt{2}} \sqrt{\ln\left(\frac{2}{\delta}\right) + \frac{\alpha}{\alpha - 1} \, \ln(n) + D_\alpha(P_n, P_0)} \, D_{\mathrm{ZCP}}\Big(P_n, P_0; \sqrt{\frac{n}{2\delta^2}}\Big) + \ln\left(1 + \frac{1}{\delta}\right) + 2 + \delta \,.\end{aligned}$$

Now we make this bound to hold uniformly over $n \in \mathbb{N} \setminus \{1\}$. In particular, denoting by $\mathrm{bound}(\delta)$ the r.h.s. of the inequality in the above, we have that $\mathbb{P}\{\int \ln \mathrm{W}_n(\theta) \,\mathrm{d}P_n(\theta) > \mathrm{bound}(\delta)\} \le 2\delta$ (note that 2δ comes by applying a union bound since we used Lemma 21 and Ville's inequality). Now, to make this bound uniform over n we apply a union bound over a set $n \in \bigcup_{i=2}^\infty [i] = \mathbb{N} \setminus \{1\}$, that is

$$\begin{aligned}
&\mathbb{P}\left\{\exists n \in \mathbb{N} \setminus \{1\} \ : \ \int \ln \mathrm{W}_n(\theta)\, \mathrm{d}P_n(\theta) > \mathrm{bound}\left(\frac{\delta}{n(n+1)}\right)\right\} \\
&\qquad \le \sum_{n \ge 2} \mathbb{P}\left\{\int \ln \mathrm{W}_n(\theta)\, \mathrm{d}P_n(\theta) > \mathrm{bound}\left(\frac{\delta}{n(n+1)}\right)\right\} \le \sum_{n \in \mathbb{N}\setminus\{1\}} \frac{2\delta}{n(n+1)} \le \delta\,.
\end{aligned}$$

Finally we apply a regret bound (Theorem 18) to lower bound $\int \ln \mathrm{W}_n(\theta)\, \mathrm{d}P_n(\theta)$ with $\int \ln \mathrm{W}^*_n(\theta)\, \mathrm{d}P_n(\theta)$ (the regret $\ln(\sqrt{2(n+1)})$ appears at the r.h.s.). Eventually we get

$$\begin{aligned}
&\int \ln \mathrm{W}^*(\theta) \mathrm{d}P_n(\theta) \\
&\quad \le \frac{1}{\sqrt{2}} \sqrt{\ln\left(\frac{2n(n+1)}{\delta}\right) + \frac{\alpha}{\alpha-1}\, \ln(n) + D_\alpha(P_n, P_0)}\, D_{\text{ZCP}}\Big(P_n, P_0;\, \sqrt{\frac{n(n(n+1))^2}{2\delta^2}}\Big) \\
&\qquad + \ln\left(1 + \frac{n(n+1)}{\delta}\right) + \ln(\sqrt{2(n+1)}) + 2 + \frac{\delta}{n(n+1)} \\
&\quad \le \frac{1}{\sqrt{2}} \sqrt{\ln\left(\frac{4n^2}{\delta}\right) + \frac{\alpha}{\alpha-1}\, \ln(n) + D_\alpha(P_n, P_0)}\, D_{\text{ZCP}}\Big(P_n, P_0;\, \frac{\sqrt{2}\, n^{2.5}}{\delta}\Big) \\
&\qquad + \ln\left(2e^2\sqrt{n}\left(1 + \frac{4n^2}{\delta}\right)\right) + \frac{\delta}{n(n+1)}\,.
\end{aligned}$$

The proof is now complete. ■

Appendix B. Proof of Corollary 11: Asymptotic behaviour of expected log-wealth bound

Fix $P_0 \in \mathcal{M}_1(\Theta)$. Let $\mathcal{K} = \bigcup_{n=1}^\infty \mathcal{K}(\mathcal{X}^n, \Theta)$. Consider Theorem 10 with $\delta = 1/n^2$ and $\alpha = \alpha_n := 1 + \frac{1}{\ln(n)}$:

$$\begin{aligned}
&\int \ln \mathrm{W}^*(\theta) \mathrm{d} P_n(\theta) \\
&\quad \leq \frac{1}{\sqrt{2}} \sqrt{\ln(2n^4) + \ln(n)\ln(en) + D_{\alpha_n}(P, P_0)}\, D_{\text{ZCP}}\Big(P, P_0;\ \sqrt{2}\, n^{4.5}\Big) \\
&\qquad + \ln\left(\sqrt{n}\left(1 + 4n^4\right)\right) + \ln(2e^2) + \frac{1}{n^3(n+1)} \\
&\quad \leq \sqrt{\ln(2n^4) + \ln(n)\ln(en) + D_{\alpha_n}(P, P_0)} \left(\frac{D_{\text{ZCP}}(P, P_0; 1)}{\sqrt{2}} + \sqrt{2\ln(2 + 2\sqrt{2}n^{4.5})} D_{\text{TV}}(P, P_0)\right) \\
&\qquad + \ln\left(\sqrt{n}\left(1 + 4n^4\right)\right) + \ln(2e^3)
\end{aligned}$$

where we applied Lemma 3 to get the second inequality.

Thus, abbreviating $F_n(P) = \int \ln W_n^*(\theta)\, \mathrm{d}P(\theta)$ and a right-hand side in the above by $B_n(P, P_0)$ we have that

$$\mathbb{P}\left(\exists P \in \mathcal{K}\ :\ F_n(P) > B_n(P, P_0)\right) \leq \frac{1}{n^2}$$

Introduce

$$L(n) = \sqrt{2\ln(n)\ln(en)\ln(2 + 2\sqrt{2}\, n^{4.5})}$$

and note that $L(n)$ is a dominating log-term in $B_n(P, P_0)$ for $n \geq 25$, and so using subadditivity of $\sqrt{\cdot}$ and some basic calculus gives

$$\frac{B_n(P, P_0)}{L(n)} \leq \underbrace{2 + \left(2 + \sqrt{D_{\alpha_n}(P, P_0)}\right)\left(D_{\text{ZCP}}(P, P_0; 1) + D_{\text{TV}}(P, P_0)\right)}_{=:A_n(P, P_0)} .$$

So,

$$\begin{aligned}
&\mathbb{P}\left(\exists P \in \mathcal{K}\ :\ \frac{F_n(P)}{L(n) A_n(P, P_0)} > 1\right) \\
&\quad \leq \mathbb{P}\left(\exists P \in \mathcal{K}\ :\ \frac{F_n(P)}{L(n)\, A_n(P, P_0)} > \frac{B_n(P, P_0)}{L(n)\, A_n(P, P_0)}\right) \leq \frac{1}{n^2} .
\end{aligned}$$

Now verify that

$$\sum_{n=25}^\infty \mathbb{P}\left(\exists P \in \mathcal{K}\ :\ \frac{F_n(P)}{L(n) A_n(P, P_0)} > 1\right) < \infty$$

and by the Borel-Cantelli lemma

$$\mathbb{P}\left(\cap_{n=1}^\infty \cup_{m=n}^\infty \cup_{P \in \mathcal{K}} \left\{\frac{F_m(P)}{L(m)\, A_m(P, P_0)} > 1\right\}\right) = 0 .$$

Note that we have

$$\cup_{P\in\mathcal{K}} \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} \left\{ \frac{F_m(P)}{L(m)\, A_m(P, P_0)} > 1 \right\}$$
$$\implies \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} \cup_{P\in\mathcal{K}} \left\{ \frac{F_m(P)}{L(m)\, A_m(P, P_0)} > 1 \right\} \ .$$

Thus,

$$\mathbb{P}\left(\exists P \in \mathcal{K} \ : \ \frac{F_n(P)}{L(n)\, A_n(P, P_0)} > 1 \quad i.o. \right) = 0$$

which implies that

$$\mathbb{P}\left(\forall P \in \mathcal{K} \ : \ \limsup_{n\to\infty} \ \frac{F_n(P)}{L(n)\, A_n(P, P_0)} \leq 1 \right) = 1$$

Now consider the property of $\limsup$ which states that for bounded real sequences $(a_n)_{n\geq 1}, (b_n)_{n\geq 1}$, $\limsup \ a_n b_n = b \limsup \ a_n$ whenever $\lim_{n\to\infty} b_n = b$. Assuming for now that $(A_n(P, P_0))_n$ is bounded we have

$$\limsup_{n\to\infty} \ \frac{F_n(P)}{L(n)\, A_n(P, P_0)} = \left(\frac{1}{\lim_{n\to\infty} \ A_n(P, P_0)} \right) \limsup_{n\to\infty} \ \frac{F_n(P)}{L(n)}$$

which means that

$$\mathbb{P}\left(\forall P \in \mathcal{K} \ : \ \limsup_{n\to\infty} \ \frac{F_n(P)}{L(n)} \leq \lim_{n\to\infty} \ A_n(P, P_0) \right) = 1 \ .$$

Finally, we look at the limit of A_n. Since P is absolutely continuous with respect to P_0, $D_{\text{KL}}(P, P_0) < \infty$ and the same hold for D_{ZCP}. Using the fact that $\lim_{n\to\infty} \ D_{\alpha_n}(P, P_0) = \lim_{\alpha\to 1} \ D_\alpha(P, P_0) = D_{\text{KL}}(P, P_0)$, we have

$$\lim_{n\to\infty} A_n = 2 + \left(2 + \sqrt{D_{\text{KL}}(P, P_0)}\right)\left(D_{\text{ZCP}}(P, P_0; 1) + D_{\text{TV}}(P, P_0)\right) \ .$$

The proof is now complete. ■

Appendix C. Other omitted proofs

C.1. Proof of Lemma 3

The proof relies on the following lemma:

Lemma 23 *For every $x \in \mathbb{R}$ and $c \geq 0$, we have*

$$\ln(1 + cx^2) \leq \ln(1 + x^2) + \ln(2 + 2c)$$

Proof If $cx^2 \geq 1$, then we have

$$\ln(1 + cx^2) \leq \ln(2cx^2) = \ln(x^2) + \ln(2c) \leq \ln(1 + x^2) + \ln(2c) \; .$$

Otherwise, we have

$$\ln(1 + cx^2) \leq \ln(2) \leq \ln(1 + x^2) + \ln(2) \; .$$

In either case, we have

$$\ln(1 + cx^2) \leq \ln(1 + x^2) + \ln(2 \vee 2c) \leq \ln(1 + x^2) + \ln(2 + 2c) \; .$$

■

Using the lemma above, we obtain

$$\begin{aligned}
&D_{\text{ZCP}}(P, Q; c) \\
&\quad = \int \left| \frac{\mathrm{d}P}{\mathrm{d}Q}(\theta) - 1 \right| \sqrt{\ln\left(1 + c^2 \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\theta) - 1\right)^2\right)} \, \mathrm{d}Q(\theta) \\
&\quad \leq \int \left| \frac{\mathrm{d}P}{\mathrm{d}Q}(\theta) - 1 \right| \sqrt{\ln\left(1 + \left|\frac{\mathrm{d}P}{\mathrm{d}Q}(\theta) - 1\right|^2\right)} \, \mathrm{d}Q(\theta) + \sqrt{\ln(2 + 2c)} \int \left| \frac{\mathrm{d}P}{\mathrm{d}Q}(\theta) - 1 \right| \mathrm{d}Q(\theta) \\
&\quad = D_{\text{ZCP}}(P, Q; 1) + 2\sqrt{\ln(2 + 2c)}\, D_{\text{TV}}(P, Q) \; .
\end{aligned}$$

■

C.2. Proof of Lemma 4

One can see that $\forall x \in \mathbb{R}, \ln(1+x^2) \le \ln(1+2|x|+x^2) \le \ln((1+|x|)^2) = 2\ln(1+|x|)$. Using

$$\begin{aligned}
D_{\mathrm{ZCP}}(P,Q;1) &= \int_\Theta \left|\frac{\mathrm{d}P(\theta)}{\mathrm{d}Q(\theta)} - 1\right| \sqrt{\ln\left(1 + \left|\frac{\mathrm{d}P(\theta)}{\mathrm{d}Q(\theta)} - 1\right|^2\right)} \,\mathrm{d}Q(\theta) \\
&\le \sqrt{2}\int_\Theta \left|\frac{\mathrm{d}P(\theta)}{\mathrm{d}Q(\theta)} - 1\right| \sqrt{\ln\left(1 + \left|\frac{\mathrm{d}P(\theta)}{\mathrm{d}Q(\theta)} - 1\right|\right)} \,\mathrm{d}Q(\theta) \\
&\overset{(a)}{\le} \sqrt{2\int_\Theta \left|\frac{\mathrm{d}P(\theta)}{\mathrm{d}Q(\theta)} - 1\right| \mathrm{d}Q(\theta) \cdot \int_\Theta \left|\frac{\mathrm{d}P(\theta)}{\mathrm{d}Q(\theta)} - 1\right| \ln\left(1 + \left|\frac{\mathrm{d}P(\theta)}{\mathrm{d}Q(\theta)} - 1\right|\right) \mathrm{d}Q(\theta)} \\
&= \sqrt{2\int_\Theta \left|\frac{\mathrm{d}P(\theta)}{\mathrm{d}Q(\theta)} - 1\right| \mathrm{d}Q(\theta) \cdot D_{f_1}(P,Q)} \qquad (f_1(x) = |x-1|\ln(1+|x-1|)) \\
&\overset{(b)}{\le} \sqrt{2\int_\Theta \left|\frac{\mathrm{d}P(\theta)}{\mathrm{d}Q(\theta)} - 1\right| \mathrm{d}Q(\theta) \cdot 2D_{f_2}(P,Q)} \qquad (f_2(x) = 1 - x + x\ln(x)) \\
&= \sqrt{4D_{\mathrm{TV}}(P,Q) \cdot 2D_{\mathrm{KL}}(P,Q)},
\end{aligned}$$

where (a) follows by Cauchy-Schwartz inequality and (b) is by Zhang et al. (2022, Lemma C.1). ■

C.3. Proof of Proposition 5

First, consider setting $a = \exp(1/p^2)$. In this way, we have that the total variation is

$$1/2(|p-q| + |1-p-(1-q)|) = |p-q| = p(1-1/a)\ .$$

On the other hand, the KL divergence is

$$p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q} = p\ln a + (1-p)\ln\frac{1-p}{1-p/a}\ .$$

Observe that the second term is non-positive and we have

$$(1-p)\ln\frac{1-p}{1-p/a} = (1-p)\ln(1-p) - (1-p)\ln(1-p/a) \ge (1-p)\ln(1-p) \ge -\exp(-1)\ .$$

Hence, we have that $p\ln a - \exp(-1) \le D_{\mathrm{KL}}(P,Q) \le p\ln a$. We now consider the two cases:

- Setting $a = \exp(1/p^2)$, we have $D_{\mathrm{KL}}(P,Q)\cdot D_{\mathrm{TV}}(P,Q) \le 1 - \frac{1}{a} = 1 - \exp(-1/p^2) \le 1 - \exp(-1)$ while $1/p - \exp(-1) \le D_{\mathrm{KL}}(P,Q) \le 1/p$.
- Setting $a = \exp(1/p^{3/2})$, we have

  $$\begin{aligned}
  D_{\mathrm{KL}}(P,Q)\cdot\sqrt{D_{\mathrm{TV}}(P,Q)} &\le \sqrt{1 - \frac{1}{a}} \\
  &= \sqrt{1 - \exp(-1/p^{3/2})} \\
  &\le \sqrt{1 - \exp(-1)}
  \end{aligned}$$

  while $1/\sqrt{p} - \exp(-1) \le D_{\mathrm{KL}}(P,Q) \le 1/\sqrt{p}$ Note that we could also use this setting for the case above as well.

■

C.4. Proof of Proposition 7

Case 1: From (Nielsen and Sun, 2018) we have that the TV distance is bounded as $D_{\mathrm{TV}}(P,Q) \le \frac{1}{2} \cdot 2p = p$. On the other hand, from the convexity of KL divergence, we have that

$$D_{\mathrm{KL}}(P,Q) \le pD_{\mathrm{KL}}(A,B) + (1-p)D_{\mathrm{KL}}(B,B) = pD_{\mathrm{KL}}(A,B) \ .$$

Now, for Gaussians we have that

$$D_{\mathrm{KL}}(A,B) = \ln \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2}{2\sigma_2^2} - \frac{1}{2}$$

and in particular choosing $\sigma_2 = p\sigma_1$, $D_{\mathrm{KL}}(A,B) = \ln(p) + \frac{1}{2p^2} - \frac{1}{2}$, which in turn implies that $D_{\mathrm{KL}}(P,Q) \le pD_{\mathrm{KL}}(A,B) = p(\ln p + \frac{1}{2p^2} - \frac{1}{2}) \le \frac{1}{2p}$. Thus $D_{\mathrm{TV}}(P,Q) \cdot D_{\mathrm{KL}}(P,Q) \le \frac{1}{2}$. On the other hand, we have that

$$\begin{aligned}
D_{\mathrm{KL}}(P,Q) &= \int \mathrm{d}P \ln \frac{p \,\mathrm{d}A + (1-p)\,\mathrm{d}B}{\mathrm{d}B} \\
&= \int \mathrm{d}P \ln \left(\frac{p\,\mathrm{d}A}{\mathrm{d}B} + (1-p) \right) \\
&= \ln(1-p) + \int \mathrm{d}P \ln \left(\frac{p\,\mathrm{d}A}{(1-p)\,\mathrm{d}B} + 1 \right) \\
&= \ln(1-p) + \int (p\,\mathrm{d}A + (1-p)\,\mathrm{d}B) \ln \left(\frac{p\,\mathrm{d}A}{(1-p)\,\mathrm{d}B} + 1 \right) \\
&\ge \ln(1-p) + \int p\,\mathrm{d}A \ln \left(\frac{p\,\mathrm{d}A}{(1-p)\,\mathrm{d}B} + 1 \right) \\
&\ge \ln(1-p) + \int p\,\mathrm{d}A \ln \frac{p\,\mathrm{d}A}{(1-p)\,\mathrm{d}B} \\
&= \ln(1-p) + pD_{\mathrm{KL}}(A,B) + p \ln \frac{p}{1-p} \\
&= \ln(1-p) + p\ln(p) + \frac{1}{2p} - \frac{p}{2} + p \ln \frac{p}{1-p} \\
&\ge \frac{1}{2p} - 1.3,
\end{aligned}$$

where -1.3 comes from the minimization of $p \mapsto \ln(1-p) + p\ln(p) - \frac{p}{2} + p\ln\frac{p}{1-p}$ over $p \in [0,1]$.

Case 2: Choosing $\sigma_2 = p^{3/4}\sigma_1$, $D_{\mathrm{TV}}(P,Q) \le p$ and $D_{\mathrm{KL}}(P,Q) \le pD_{\mathrm{KL}}(A,B) = p(\ln \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2}{2\sigma_2^2} - \frac{1}{2}) = p(\frac{3}{4}\ln p + \frac{1}{2p^{3/2}} - \frac{1}{2}) \le \frac{1}{2\sqrt{p}}$. Thus, $D_{\mathrm{KL}}(P,Q) \cdot \sqrt{D_{\mathrm{TV}}(P,Q)} \le \frac{1}{2}$. On the other hand, reasoning as above, we have

$$\begin{aligned}
D_{\mathrm{KL}}(P,Q) &\ge \ln(1-p) + pD_{\mathrm{KL}}(A,B) + p \ln \frac{p}{1-p} \\
&= \ln(1-p) + p \left(\frac{3}{4} \ln p + \frac{1}{2p^{3/2}} - \frac{1}{2} \right) + p \ln \frac{p}{1-p}, \\
&\ge \frac{1}{2\sqrt{p}} - 1.22,
\end{aligned}$$

where -1.22 comes from the minimization of $\ln(1-p) + p(\frac{3}{4}\ln p - \frac{1}{2}) + p\ln\frac{p}{1-p}$ over $p \in [0,1]$.
■

C.5. Proof of empirical Bernstein inequalities (Corollary 14)

To show this corollary we follow Jang et al. (2023, Proposition 4). Abbreviate $\hat{\mu}_\theta = \frac{1}{n}\sum_{i=1}^n f(\theta, X_i)$, and so $\Delta_n(\theta) + \hat{\mu}_\theta \in [0,1]$. Then, we have

$$\ln \mathrm{W}_n^*(\theta) \geq \max_{\beta \in [-1,1]} \sum_{i=1}^n \ln\Big(1 + \beta\left(f(\theta, X_i) - (\hat{\mu}_\theta + \Delta(\theta))\right)\Big) ,$$

and applying Jensen's inequality

$$\int \ln \mathrm{W}_n^*(\theta) \,\mathrm{d}P_n \geq \max_{\beta \in [-1,1]} \int \sum_{i=1}^n \ln\Big(1 + \beta\left(f(\theta, X_i) - (\hat{\mu}_\theta + \Delta(\theta))\right)\Big) \,\mathrm{d}P_n \,. \tag{9}$$

We relax the above by taking a lower bound of (Fan et al., 2015, Eq. 4.12) which shows that for any $|x| \leq 1$ and $|\beta| \leq 1$,

$$\ln(1 + \beta x) \geq \beta x + \big(\ln(1 - |\beta|) + |\beta|\big) x^2 \,. \tag{10}$$

Then, combined with the following lemma:

Lemma 24 (Orabona and Jun (2024, Lemma 5)) *Let* $f(\beta) = a\beta + b\big(\ln(1 - |\beta|) + |\beta|\big)$ *for some* $a \in \mathbb{R}, b \geq 0$*. Then,* $\max_{\beta \in [-1,1]} f(\beta) \geq \frac{a^2}{(4/3)|a| + 2b}$.

we get a chain of inequalities:

$$\begin{aligned}
\int \ln \mathrm{W}_n^*(\theta) \,\mathrm{d}P_n &\overset{(a)}{\geq} \beta \int \sum_{i=1}^n \left(f(\theta, X_i) - (\hat{\mu}_\theta + \Delta_n(\theta))\right) \mathrm{d}P_n \\
&\quad + \big(\ln(1 - |\beta|) + |\beta|\big) \int \sum_{i=1}^n \left(f(\theta, X_i) - (\hat{\mu}_\theta + \Delta_n(\theta))\right)^2 \mathrm{d}P_n \\
&= -n\beta \int \Delta_n(\theta) \,\mathrm{d}P_n \\
&\quad + \big(\ln(1 - |\beta|) + |\beta|\big) \left(\int \sum_{i=1}^n (f(\theta, X_i) - \hat{\mu}_\theta)^2 \,\mathrm{d}P_n + n \int \Delta_n(\theta)^2 \,\mathrm{d}P_n \right) \\
&\overset{(b)}{\geq} -n\beta \int \Delta_n(\theta) \,\mathrm{d}P_n \\
&\quad + \big(\ln(1 - |\beta|) + |\beta|\big) \left(\int \sum_{i=1}^n (f(\theta, X_i) - \hat{\mu}_\theta)^2 \,\mathrm{d}P_n + n \left(\int \Delta_n(\theta) \,\mathrm{d}P_n \right)^2 \right) \\
&\overset{(c)}{\geq} \frac{n^2 \left(\int \Delta_n(\theta) \,\mathrm{d}P_n\right)^2}{(4/3) n \left|\int \Delta_n(\theta) \,\mathrm{d}P_n\right| + 2 \int \sum_{i=1}^n (f(\theta, X_i) - \hat{\mu}_\theta)^2 \,\mathrm{d}P_n + 2n \left(\int \Delta(\theta) \,\mathrm{d}P_n\right)^2} \,.
\end{aligned}$$

Here, step (a) comes from Eqs. (9) and (10), whereas (b) comes from Jensen's inequality, and step (c) is due to application of Lemma 24. At this point, the first result of Corollary 14 comes from the above combined with the fact that $|\Delta_n| \leq 1$ and by using Eq. (5).

Now, to state the second result of Corollary 14 we use a PAC-Bayes bound of Theorem 10 to have

$$n^2 \left(\int \Delta_n(\theta) \,\mathrm{d}P_n \right)^2 \leq n \,\mathrm{Comp}_n(\alpha) \left(\frac{4}{3} \left| \int \Delta_n(\theta) \,\mathrm{d}P_n \right| + 2\hat{V}(P_n) + 2 \left(\int \Delta_n(\theta) \,\mathrm{d}P_n \right)^2 \right) \,.$$

Solving the above for $\int \Delta_n(\theta)\,\mathrm{d}P_n$, using subadditivity of square root, and relaxing some numerical constants we get

$$\left|\int \Delta_n(\theta)\,\mathrm{d}P_n\right| \le \frac{\sqrt{2\,\mathrm{Comp}_n(\alpha)\,\hat{V}(P_n)}}{(\sqrt{n} - (2/\sqrt{n})\,\mathrm{Comp}_n(\alpha))_+} + \frac{2\mathrm{Comp}_n(\alpha)}{\big(n - 2\,\mathrm{Comp}_n(\alpha)\big)_+}\,.$$

The proof of the asymptotic version is immediate from the proof of Corollary 11. ■

C.6. Proof of Lemma 19

We have that

$$\begin{aligned}
\exp\left(\frac{\left(\sum_{i=1}^n c_i\right)^2}{4n}\right) &= \max_{\beta\in[-1/2,1/2]} \exp\left(\beta\sum_{i=1}^n c_i - \beta^2 n\right)\\
&\le \max_{\beta\in[-1/2,1/2]} \exp\left(\beta\sum_{i=1}^n c_i - \beta^2 \sum_i c_i^2\right)\\
&\le \max_{\beta\in[-1/2,1/2]} \exp\left(\sum_{i=1}^n \ln(1+\beta c_i)\right),
\end{aligned}$$

where we use the elementary inequality $\ln(1+x) \ge x - x^2$ for $|x| \le 1/2$. ■