Journal of Machine Learning Research 24 (2023) 1-61 Submitted 3/23; Revised 10/23; Published 12/23

A Unified Recipe for Deriving (Time-Uniform) PAC-Bayes Bounds

Ben Chugg BENCHUGG@CMU.EDU
Hongjian Wang HJNWANG@CMU.EDU
Aaditya Ramdas ARAMDAS@STAT.CMU.EDU
Departments of Statistics and Machine Learning
Carnegie Mellon University

Editor: John Shawe-Taylor

Abstract

We present a unified framework for deriving PAC-Bayesian generalization bounds. Unlike most previous literature on this topic, our bounds are anytime-valid (i.e., time-uniform), meaning that they hold at all stopping times, not only for a fixed sample size. Our approach combines four tools in the following order: (a) nonnegative supermartingales or reverse submartingales, (b) the method of mixtures, (c) the Donsker-Varadhan formula (or other convex duality principles), and (d) Ville's inequality. Our main result is a PAC-Bayes theorem which holds for a wide class of discrete stochastic processes. We show how this result implies time-uniform versions of well-known classical PAC-Bayes bounds, such as those of Seeger, McAllester, Maurer, and Catoni, in addition to many recent bounds. We also present several novel bounds. Our framework also enables us to relax traditional assumptions; in particular, we consider nonstationary loss functions and non-i.i.d. data. In sum, we unify the derivation of past bounds and ease the search for future bounds: one may simply check if our supermartingale or submartingale conditions are met and, if so, be guaranteed a (time-uniform) PAC-Bayes bound.

Keywords: PAC-Bayes, martingales, anytime-valid bounds, Ville's inequality, statistical learning theory

1. Introduction

PAC-Bayesian theory is broadly concerned with providing generalization guarantees over mixtures of predictors in statistical learning problems. It emerged in the late 1990s, catalyzed by an early paper of Shawe-Taylor and Williamson (1997) and shepherded forward by McAllester (McAllester, 1998, 1999, 2003), Catoni (Catoni, 2003, 2004, 2007), Maurer (Maurer, 2004), and Seeger (Seeger, 2002, 2003), among others. The earliest works were focused mainly on classification settings but the techniques have expanded to regression settings (Audibert, 2004; Alquier, 2008), and more recently to settings beyond supervised learning (e.g., Seldin and Tishby, 2010). We refer the reader to Alquier (2021) and Guedj (2019) for excellent surveys.

In the supervised learning setting, PAC-Bayesian (or simply "PAC-Bayes") theory seeks to bound the expected risk in terms of the expected empirical risk, where the expectation is with respect to a data-dependent distribution ρ over the hypothesis space. This is in contrast to uniform convergence guarantees, which give worst case bounds over all hypotheses. The

PAC-Bayes approach is not without limitations (Livni and Moran, 2020), but has led to non-trivial guarantees for SVMs (Ambroladze et al., 2006), sparse additive models (Guedj and Alquier, 2013), and neural networks (Dziugaite and Roy, 2017; Letarte et al., 2019). Whereas uniform convergence bounds typically rely on some notion of the complexity of the hypothesis class, PAC-Bayes bounds depend on the distance between ρ and a prior distribution ν. Depending on the choice of ν and ρ, the resulting bounds can be tighter and easier to compute.

Despite these successes, we point out two drawbacks. First, there does not seem to be a clearly established recipe to deriving PAC-Bayes bounds. Many full-length papers are dedicated to deriving one or two interesting bounds, using different techniques. Is there a common thread to tie the decades of work together? Can a unified view (achieved with the power of hindsight) yield new bounds with relative ease? Second, most existing PAC-Bayes bounds are fixed-time results. That is, the bounds hold at a fixed number of observations determined *a priori*, despite the fact that the distribution ρ can be data-dependent. In fact, this is the case for the vast majority of the learning theory literature. Undoubtedly, this is a consequence of the extensive number of fixed time concentration inequalities stemming from the statistics literature (e.g., the Chernoff bound and the Azuma-Hoeffding inequality; see Boucheron et al., 2013 for an overview). However, fixed-time bounds are not valid at stopping times; if the bound is computed at a sample size that is itself data-dependent (perhaps resulting from sequential decisions), then it is invalid. Naïve union bounds over all of time are too loose, falling short theoretically, practically, and aesthetically.

In this work, we take advantage of recent progress on unified schemes for deriving *anytime-valid* concentration inequalities (Howard et al., 2020, 2021) to give a general framework for developing anytime-valid (a.k.a. time-uniform)[1] PAC-Bayes bounds. Anytime-valid bounds hold at all stopping times. Importantly, this means they hold regardless of whether one has looked at the data or not when deciding the final sample size. They are thus inherently immune to continuous monitoring of data and adaptive stopping.

Concurrent to our own work, Haddouche and Guedj (2023) derived several anytime-valid PAC-Bayes bounds. They also employ supermartingales and Ville's inequality, two ingredients which are central to our approach. Our general framework will encompass their results, recovering their theorems as special cases of our own. More importantly however, our unified framework will cover a much broader slew of existing PAC-Bayes bounds. See Table 1 for a summary of these results.

At a high level, our approach combines four tools in the following order: (A) nonnegative supermartingales or reverse submartingales, (B) mixtures of said processes (often called the "method of mixtures"), (C) a change-of-measure inequality which provides a variational representation of some convex divergence (e.g, the Donsker-Varadhan formula in the case of KL divergence), and (D) Ville's inequality (Ville, 1939), a time-uniform extension of Markov's inequality to nonnegative supermartingales and reverse submartingales. Recent work has established that principles (A)+(D) yield a unified approach to deriving time-uniform Chernoff bounds (e.g., Howard et al., 2020), while using (A)+(B)+(D) yields a unified approach to deriving confidence sequences (e.g., Howard et al., 2021). This paper

1. In this paper, "anytime-valid" and "time-uniform" are synonymous. However, this is not always the case. See the discussion at the end of Section 1.1.

(A) Non-negative forward supermartingale or reverse submartingale	**(B)** Method of Mixtures	**(C)** Variational Representation of a Convex Divergence (Change of Measure)	**(D)** Ville's Inequality

(A) + (D) $\implies$ Unified recipe for time-uniform Chernoff bounds (Howard et al., 2020; Manole and Ramdas, 2023)

(A) + (B) + (D) $\implies$ Unified recipe for time-uniform confidence sequences (Ramdas et al., 2020; Howard et al., 2021)

(A) + (B) + (C) + (D) $\implies$ Unified recipe for time-uniform PAC-Bayes bounds (this work)

Figure 1: An overview of the tools employed in this paper, and how they relate to previous work on time-uniform bounds.

shows that adding (C) yields a unified approach to PAC-Bayes bounds. See Figure 1 for a schema of how this work relates to other unified recipes and time-uniform bounds.

1.1 Setting

We observe a sequence of data $(Z_t)_{t=1}^{\infty}$ where each Z_i lies in some domain $\mathcal{Z}$. The data have a distribution $\mathcal{D}$ over $\mathcal{Z}^{\infty}$. We emphasize that $\mathcal{D}$ is a distribution over *sequences* of observations, enabling us to consider non-i.i.d. data. We will specify the precise distributional assumptions later on. Each time step t is associated with a function $f_t : \mathcal{Z} \times \Theta \to \mathbb{R}_{\geqslant 0}$, where Θ is some (measurable) parameter space. Each $\theta \in \Theta$ gives rise to the *loss function* $f_t(\cdot, \theta)$. Thus, f_t should be seen as a family of loss functions parameterized by Θ. If $f = f_t$ does not change with time, we say it is *stationary*.

In a typical supervised learning task, the domain is taken to be the product $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ the label space. In this case, we might consider the (stationary) loss function $f(Z_t, \theta) = (Y_t - \langle \theta, X_t \rangle)^2$, where $Z_t = (X_t, Y_t)$. However, PAC-Bayesian bounds have proven useful outside of supervised learning—for instance, estimating means (Catoni and Giulini, 2017, 2018), clustering (Seldin and Tishby, 2010), and discrete density estimation (Seeger, 2003; Seldin and Tishby, 2009). Thus, we choose to adopt the more general notation. We note that allowing the loss function to change as a function of time is not the typical assumption in the PAC-Bayes literature. However, we find that our framework can handle non-stationary losses at no extra cost, so we see no harm (and some benefit) in this additional level of generality.

For a fixed $\theta \in \Theta$, the empirical risk and the (conditional) risk at time t are, respectively,

$$\widehat{R}_t(\theta) = \frac{1}{t} \sum_{i=1}^{t} f_i(Z_i, \theta), \quad \text{and} \quad R_t(\theta) = \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}[f_i(Z_i, \theta) | \mathcal{F}_{i-1}]. \tag{1}$$

Here $\mathcal{F}_{i-1}$ is the σ-algebra generated by $Z_1, \ldots, Z_{i-1}$ (formally introduced in Section 2). If the losses are stationary and the data are i.i.d. (or, more generally, $\mathbb{E}[f_t(Z_t, \theta) | \mathcal{F}_{t-1}]$ is assumed to have a common mean across all $t \geqslant 1$) then the conditional risk is constant as a function of time, and we denote it as $R(\theta) = \mathbb{E}[f(Z, \theta)]$.

Uniform convergence guarantees provide a natural and popular way to bound the risk in terms of the empirical risk. Such guarantees provide bounds simultaneously for all $\theta \in \Theta$, and typically depend on quantities such as the VC dimension or the Rademacher complexity of the family of losses (see, e.g., Wainwright, 2019). In contrast, PAC-Bayes bounds seek to give guarantees on the difference between $\mathbb{E}_{\theta\sim\rho}\widehat{R}_t(\theta)$ and $\mathbb{E}_{\theta\sim\rho}R_t(\theta)$ for all data-dependent mixture distributions $\rho \in \mathcal{M}(\Theta)$, where $\mathcal{M}(\Theta)$ is the set of probability distributions over Θ. Additionally, we typically begin with a (data-free) prior $\nu \in \mathcal{M}(\Theta)$ over the parameters.

In order to orient the reader, we state a PAC-Bayes bound due to Catoni (2003) for bounded, stationary losses in $[0, 1]$. The order of quantifiers below is particularly important to note. For all priors $\nu \in \mathcal{M}(\Theta)$, error probabilities $\delta \in (0, 1)$, sample sizes n and tuning parameters $\lambda > 0$, we have that with probability at least $1 - \delta$, for all $\rho \in \mathcal{M}(\Theta)$,

$$\mathbb{E}_{\theta\sim\rho}[R_n(\theta) - \widehat{R}_n(\theta)] \leqslant \frac{\lambda}{8n} + \frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta)}{\lambda}, \tag{2}$$

where $D_{\mathrm{KL}}(\rho\|\nu)$ is the KL divergence between ρ and ν (defined in Section 2). Said differently, *"Fixing ν, δ, n, λ, with probability $1 - \delta$, (2) holds simultaneously for all ρ."*, emphasizing the quantities that are fixed before seeing the data.

Notice that the generalization guarantee depends not on a measure of complexity of the class of functions $\{f(\cdot, \theta) : \theta \in \Theta\}$ as it would in uniform convergence bounds. Instead, it depends on the divergence between our prior ν and a data-dependent ρ. The KL divergence is the most common measure of divergence used in PAC-Bayes bounds because of the famous "change of measure" inequality by Donsker and Varadhan (1975) but Rényi divergence (Bégin et al., 2016), f divergences (Alquier and Guedj, 2018; Ohnishi and Honorio, 2021), and Integral Probability Metrics (Amit et al., 2022) have also been studied.

Let us now introduce anytime-valid and time-uniform bounds. As stated, (2) is a fixed-time bound because, as discussed above, the universal quantifier on n is "outside" the probability statement. This is characteristic of most concentration inequalities. A time-uniform bound, on the other hand, incorporates the number of samples "inside" the probability statement. It is of the form "with probability $1 - \delta$, for all n, ...". Moving forward, we will substitute t (standing for time) in place of n to draw attention to the distinction. For instance, here is the time-uniform equivalent of (2) above. For all priors $\nu \in \mathcal{M}(\Theta)$, error probabilities $\delta \in (0, 1)$, and tuning parameters $\lambda, n > 0$, with probability at least $1 - \delta$, we have that simultaneously for all $t \geqslant 1$ and $\rho \in \mathcal{M}(\Theta)$,

$$\mathbb{E}_{\theta\sim\rho}[R_t(\theta) - \widehat{R}_t(\theta)] \leqslant \frac{\lambda}{8n} + \frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta)}{\lambda t/n}. \tag{3}$$

Here we have kept a pre-specified n in the bound to facilitate easy comparison with (2); however, this parameter could be absorbed into λ. While the distinction between time-uniform and fixed-time bounds may seem a minor notational detail, it is in fact a major mathematical difference with ramifications across science and any kind of data-driven decision-making (Howard et al., 2021; Grünwald et al., 2023; Ramdas et al., 2023). Importantly, time-uniform results are immune to "peeking" because they remain valid at stopping times.

Anytime-valid bounds, meanwhile, are (in)equalities that hold at arbitrary stopping times. A full discussion of the distinction between anytime-valid and time-uniform bounds

is beyond the scope of this work, but we refer the interested reader to Ramdas et al. (2020) for further detail (see Lemmas 2 and 3 in particular). Suffice it to say that for probability statements like above, time-uniformity is synonymous with anytime-validity. This manuscript is concerned with anytime-valid probability statements, so we use the two terms interchangeably.

1.2 Contributions and Outline

In this work, we identify a general martingale-like structure at the heart of many existing PAC-Bayes bounds. This structure takes the form of either a nonnegative supermartingale or a nonnegative reverse submartingale. Such an identification enables us to (i) give a general framework for seeking new bounds, and (ii) give time-uniform extensions of many existing PAC-Bayes bounds. Our main contribution is a general result (Theorem 4) which provides a time-uniform PAC-Bayes bound for any process which is (upper bounded by) a nonnegative supermartingale or reverse submartingale. We proceed to instantiate this bound with a variety of particular processes and relate them to existing results in the literature (Table 1). For those bounds which admit a supermartingale structure, we find that their time-uniform extensions remain as tight as their fixed-time counterparts. For those that admit a reverse submartingale structure we provide two results: (a) a time-uniform bound holding for all $t \geqslant 1$ which loses at most a constant factor plus an iterated logarithm term (i.e., $\log\log t$) over the original, and (b) a bound which holds for all times $t \geqslant n$, where n is some time of special interest chosen beforehand, which remains just as tight as the original fixed-time bounds. Finally, our framework enables us to relax many traditional assumptions (Table 2). For instance, many of our bounds do not require i.i.d. data. In fact, our supermartingale-based bounds require no explicit distributional assumptions.

As was mentioned in the introduction, the closest work to ours is the concurrent preprint of Haddouche and Guedj (2023). They apply Ville's inequality to a supermartingale identified by Bercu and Touati (2008), which gives a time-uniform PAC-Bayes bound for unbounded loss functions. In Section 4 we will demonstrate that this supermartingale was known to be a part of a much wider class of stochastic processes known as sub-ψ processes (Howard et al., 2020), and provide an anytime-valid PAC-Bayes result for this large class, recovering their result as a special case.

Stepping back from the particulars, our work is best viewed in the spirit of recent progress in time-uniform Chernoff bounds and sequential estimation (Figure 1). We draw much inspiration from the recent works by Howard et al. (2020, 2021) who study a unifying approach to time-uniform bounds via supermartingales. Howard et al. (2020) showed that many (or most, or all) Chernoff bounds can be made time-uniform at no loss (and sometimes a gain) by identifying an appropriate supermartingale and applying Ville's inequality (our Lemma 1). In other words, applying Ville's inequality to nonnegative supermartingales is a unifying strategy for generating Chernoff bounds. This insight was the inspiration for seeking to identify underlying supermartingales in PAC-Bayes bounds. Howard et al. (2021) then built upon this foundation, and developed confidence sequences (i.e., confidence intervals that hold at all stopping times) with zero asymptotic width using a variety of mixtures of supermartingales. This "method of mixtures" plays an important role in our results in two respects. For one, it is required since the PAC-Bayes framework gives bounds

	Existing result	*Our result*
Forward supermartingale	McAllester (1999), Thm. 1	Corollary 9
	Catoni (2003)	Corollary 7
	Catoni (2007)	Corollary 16
	Seldin et al. (2012), Thm. 5 & 6	Corollary 37
	Seldin et al. (2012), Thm. 7 & 8	Corollary 38
	Balsubramani (2015), Thm. 1	Corollary 38
	Alquier et al. (2016), Thm. 4.1	Corollary 7
	Kuzborskij and Szepesvári (2019), Thm. 4	Corollary 6
	Haddouche et al. (2021), Thm. 3	Corollary 17
	Haddouche and Guedj (2023), Thm. 5	Corollary 6
	Haddouche and Guedj (2023), Thm. 7	Corollary 18
	Jang et al. (2023), Thm. 1	Corollary 15
Reverse submartingale	McAllester (1999), Thm. 1	Corollary 26
	Langford and Seeger (2001), Thm. 3	Corollary 25
	Seeger (2002), Thm. 2	Corollary 42
	Maurer (2004), Thm. 5	Corollary 25
	Catoni (2007), Thm. 1.2.6	Corollary 22
	Germain et al. (2009), Thm. 2.1	Corollary 22
	Seldin et al. (2012), Thm. 4	Corollary 39
	Tolstikhin and Seldin (2013), Eqn. 3	Corollary 25
	Tolstikhin and Seldin (2013), Thm. 3 & 4	Corollary 27
	Germain et al. (2015), Thm. 18	Corollary 22
	Bégin et al. (2016), Thm. 9	Corollary 34
	Thiemann et al. (2017), Thm. 3	Corollary 22
	Alquier (2021), Eqn. (3.1)	Corollary 25
	Amit et al. (2022), Prop. 4 and 5	Corollary 30

Table 1: A summary of how various existing results are related to our framework. The first column refers to the type of underlying process used to construct the bound. For supermartingales, the time-uniform extension sacrifices no tightness compared to the original. For reverse submartingales, our anytime bound loses essentially an iterated logarithm factor over the fixed-time bound (but the fixed-time bound itself remains recoverable at no loss). The final column points to which corollary implies the existing result (either directly or as a consequence of selecting certain parameters; the precise relationship will be described in the text). The above results are mostly corollaries of Theorem 4 (a PAC-Bayes framework with the KL divergence), but several rely on Theorem 31 (a framework for general ϕ-divergences) or Theorem 33 (a framework for Rényi divergences). The PAC-Bayes literature is large and we cannot include all previous results and their relationships, but we hope this gives the reader an idea of the scope of our approach. We do not provide numbers in the second and third rows because the bounds were not explicitly written out in Catoni (2003, 2007). See Alquier (2021) for a summary.

over mixtures of hypotheses. Second, it yields novel PAC-Bayes bounds by mixing the supermartingales that underlie existing bounds with various mixing distributions. The former yields uniformity over distributions, the latter over sample size.

Interestingly, we find that not all existing PAC-Bayes bounds can be given time-uniform generalizations based on nonnegative supermartingales. For some, including those of Seeger (2003); Tolstikhin and Seldin (2013); Germain et al. (2015) which ultimately rely on applying convex functions to the risk and empirical risk, we must instead rely on *reverse submartingales*. Our inspiration for such tools comes from recent work by Manole and Ramdas (2023), who showed that convex functionals and divergences are reverse submartingales (with respect to the exchangeable filtration). Since there also exists a reverse-time Ville's inequality, backwards submartingales and (backwards) Ville's inequality provide a second unifying recipe for deriving time-uniform bounds.

In short, this paper shows how systematically combining four techniques provides a unified recipe to derive time-uniform PAC-Bayesian inequalities.

Outline. The rest of the manuscript is organized as follows. Section 2 provides relevant background on (reverse) martingales, Ville's inequalities, and the change-of-measure inequality which lies at the heart of PAC-Bayesian analysis. Section 3 provides a "master theorem" which gives an anytime-valid PAC-Bayes bound for general nonnegative stochastic processes which are upper bounded by either a supermartingale or reverse submartingale. Section 4 then explores various consequences in the supermartingale case, and Section 5 does the same for the reverse submartingale case. Section 6 then discusses a number of extensions; Sections 6.1 and 6.2 study extensions of our master Theorem to Integral Probability Metrics, ϕ-divergences, and the Rényi divergence. Section 6.3 gives some connections to recent work on time-uniform confidence sequences, Section 6.4 demonstrates that our results hold for martingale difference sequences, and Section 6.5 investigates to what extent we can employ data-dependent priors. Finally, Section 6.6 ends with an application to Gaussian process classification.

2. Background

Notation. As discussed previously, we let $\mathcal{D}$ be a distribution over sequences $(Z_t) \in \mathcal{Z}^\infty$. In order to save ourselves from an overload of notation, we will write $\mathbb{E}_{\mathcal{D}}[\cdot]$ to denote the expectation when drawing $(Z_t) \sim \mathcal{D}$, i.e., $\mathbb{E}_{\mathcal{D}}[\cdot] = \mathbb{E}_{(Z_t)\sim\mathcal{D}}[\cdot]$. Furthermore, we will use the convention that expectation over lowercase Greek letters refer to expectation over parameters $\theta \in \Theta$, e.g., $\mathbb{E}_\rho[\cdot] = \mathbb{E}_{\theta\sim\rho}[\cdot]$. We also write Z^n as shorthand for $Z_1, \dots, Z_n$. For a stochastic process $(A_t)_{t=t_0}^\infty$ (or infinite sequence more generally) we will often simply write (A_t), where t_0 will be understood from context. We write $\mathcal{M}(\Theta)$ for the set of probability distributions over Θ. We use $\mathbb{R}_{\geqslant 0}$ to be the set of nonnegative reals (similarly for $\mathbb{R}_{>0}$). When we say that $\nu \in \mathcal{M}(\Theta)$ is a prior, it should be assumed that it is data-free, i.e., independent of the data (Z_t). Writing, e.g., $\mathbb{E}_{\mathcal{D}}[g(Z_i)]$ for some function g should be taken to mean that the sequence (Z_t) was drawn from $\mathcal{D}$ but we are restricting ourselves to the i-th value. We may also write $\mathbb{E}_{Z_i}[g(Z_i)]$ in this case. Finally, we let $\mu_t(\theta) = \mathbb{E}_{\mathcal{D}}[f_t(Z_t, \theta)|\mathcal{F}_{t-1}]$.

A *forward filtration* is a sequence of σ-algebras $(\mathcal{F}_t)_{t=1}^\infty$ such that $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$ for all $t \geqslant 1$. If $\mathcal{F}_t = \sigma(Z_1, \dots, Z_t)$, we call $(\mathcal{F}_t)_{t=1}^\infty$ the canonical (forward) filtration. Intuitively, we conceive of $\mathcal{F}_t$ as all the information available at time t. Thus, if a function f is $\mathcal{F}_t$-

measurable, it may depend on data $Z_1, \ldots, Z_t$, but not on any Z_i for $i > t$. If a sequence of functions $(f_t)_{t=1}^{\infty}$ is such that f_t is $\mathcal{F}_t$ measurable for all t, then we say that $(f_t)_{t=1}^{\infty}$ is *adapted* to $\mathcal{F}_t$. If f_{t+1} is $\mathcal{F}_t$ measurable for all t, then we say the sequence is *predictable*.

A *martingale* adapted to the forward filtration $(\mathcal{F}_t)_{t=1}^{\infty}$ is a stochastic process $(S_t)_{t=1}^{\infty}$ such that S_t is $\mathcal{F}_t$ measurable and $\mathbb{E}[S_{t+1}|\mathcal{F}_t] = S_t$ for all $t \geqslant 1$. If the equality is replaced with $\leqslant$ (resp., $\geqslant$) we call (S_t) a *supermartingale* (resp., *submartingale*). Supermartingales are thus decreasing with time in expectation, whereas submartingales are increasing. Martingales stay constant in expectation. For this reason, they often represent fair games. Forward filtrations are in contrast to reverse filtrations, which we cover later in this section. Henceforth, if we discuss filtrations unencumbered by a preceding adjective, then it is a forward filtration.

It's perhaps worth remarking that a martingale is only a martingale *with respect to* a particular measure $\mathbb{P}$. For instance, the process $S_t = \frac{1}{t}\sum_i X_i - m$ for i.i.d. X_i is a martingale iff $\mathbb{P}(X_i) = m$. Formally then, one should refer to (S_t) as (possibly) being a $\mathbb{P}$-martingale. However, in our case the measure will usually be clear from context and we will simply refer to martingales. The same discussion holds for sub/supermartingales.

Supermartingales are natural tools to use when deriving anytime-valid bounds due to Ville's inequality (Ville, 1939), given in Lemma 1. Informally, Ville's inequality is a time-uniform version of Markov's inequality. It states that a nonnegative supermartingale with initial value 1 remains small (say, less than $1/\delta$) at all times with probability roughly $1 - \delta$. A digestible proof of Ville's inequality may be found in Howard et al. (2020).

Lemma 1 (Ville's Inequality for Nonnegative Supermartingales) *Let $(N_t)_{t=1}^{\infty}$ be a nonnegative supermartingale with respect to the filtration $(\mathcal{F}_t)_{t=1}^{\infty}$. For all times t_0 and $u \in \mathbb{R}_{>0}$,*

$$\mathbb{P}(\exists t \geqslant t_0 : N_t \geqslant u) \leqslant \frac{\mathbb{E}[N_{t_0}]}{u}.$$

Ville's inequality can be restated as $\mathbb{P}(\forall t \geqslant t_0 : N_t/N_{t_0} < u) \geqslant 1 - 1/u$. Written this way, its power for providing time-uniform guarantees becomes evident.

Under appropriate conditions, mixtures of martingales remain martingales. That is, if $V_t(\theta)$ is a (sub/super) martingale, then $\mathbb{E}_{\theta\sim\rho}V_t(\theta)$ for well-behaved mixtures ρ is also a (sub/super) martingale. The precise statement and corresponding proof are given in Appendix B. This is useful because if we have a family of nonnegative supermartingales (say) of the form $N_t(\lambda)$ for $\lambda \in \mathbb{R}$, we can look for appropriate mixture distributions F and conclude that $\int_{\lambda\in\mathbb{R}} N_t(\lambda)\mathrm{d}F(\lambda)$ is also a nonnegative supermartingale, and thus by Ville's inequality:

$$\mathbb{P}\left(\forall t \geqslant t_0 : \int_{\lambda\in\mathbb{R}} N_t(\lambda)\mathrm{d}F(\lambda) \leqslant 1/\delta\right) \geqslant 1 - \delta.$$

This has been called the "method of mixtures", and was noticed by Wald (1945) and Robbins (1970). Depending on the mixture distribution F, this bound can be more desirable than that based solely on $N_t(\lambda)$. Indeed, this approach has been successfully leveraged to generate time-uniform confidence intervals (i.e., confidence *sequences*) (Howard et al., 2021; Waudby-Smith et al., 2021, 2023). For our part, in Section 4.2 we give a novel PAC-Bayes bound using a Gaussian mixture distribution, as a demonstrative example.

The machinery of nonnegative supermartingales (and their mixtures) in addition to Ville's inequality is sufficient to give time-uniform PAC-Bayes bounds in a wide variety of situations. Section 4 is dedicated to this task. See the first half of Table 1 for those bounds which are recovered using this technique. However, to recover time-uniform versions of other well-known PAC-Bayes bounds, we must rely on reverse-time martingales. We introduce these next.

A *reverse filtration* $(\mathcal{R}_t)_{t=1}^{\infty}$ is a sequence of σ-algebras such that $\mathcal{R}_t \supseteq \mathcal{R}_{t+1}$ for all t. That is, a reverse filtration represents decreasing information with time. A *reverse martingale* (S_t) adapted to a reverse filtration $(\mathcal{R}_t)$ is a stochastic process such that S_t is $\mathcal{R}_t$ measurable and $\mathbb{E}[S_t|\mathcal{R}_{t+1}] = S_{t+1}$ for all $t \geqslant 1$. Again, replacing the equality with $\leqslant$ (resp., $\geqslant$) results in reverse supermartingales (resp., submartingales). Reverse processes are also called *backwards* or *reverse-time* processes. We will use such language interchangeably. An example of a reverse martingale is the empirical mean $\frac{1}{t}\sum_{i=1}^{t} Z_i$ adapted to the canonical reverse filtration $\mathcal{R}_t = \sigma(Z_t, Z_{t+1}, \dots)$. Since filtrations and stochastic processes are typically considered in the context of "increasing" time, reverse-time processes can be initially confounding. When thinking about reverse martingales, we encourage the reader to imagine time flowing backwards, i.e., information being revealed first at time t, then at time $t-1$, $t-2$ and so on. Thus, reverse submartingales are increasing in expectation in reverse-time and, if one were to plot the expected values such a process would resemble a *super*martingale in forward time. With this insight in mind, it is relieving to know that there is a variant of Ville's inequality for reverse submartingales. Proofs may be found in Lee (2019); Manole and Ramdas (2023).

Lemma 2 (Reverse Ville's Inequality) *Let (M_t) be a nonnegative reverse submartingale with respect to a reverse filtration $(\mathcal{R}_t)_{t=1}^{\infty}$. For all t_0 and $u \in \mathbb{R}_{>0}$,*

$$\mathbb{P}(\exists t \geqslant t_0 : M_t \geqslant u) \leqslant \frac{\mathbb{E}[M_{t_0}]}{u}.$$

Section 5 will employ reverse submartingales in order to give time-uniform PAC-Bayes bounds on convex functions φ of the expected and empirical risk. This will enable us to give time-uniform versions of inequalities presented by Seeger (2003); McAllester (1998); Maurer (2004); Germain et al. (2009, 2015); Tolstikhin and Seldin (2013), among others. Finally, we present the change-of-measure inequality due to Donsker and Varadhan (1975) which is central to the majority of existing PAC-Bayes bounds. Before it is stated, let us recall that the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between two distributions μ and π in $\mathcal{M}(\Theta)$ is

$$D_{\mathrm{KL}}(\mu\|\pi) = \mathbb{E}_{\theta\sim\mu}\left[\log\left(\frac{\mathrm{d}\mu(\theta)}{\mathrm{d}\pi}\right)\right] = \int_{\Theta} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(\theta)\right)\mu(\mathrm{d}\theta),$$

if μ is absolutely continuous with respect to π (i.e., $\mu(A) = 0$ whenever $\pi(A) = 0$), and $+\infty$ otherwise. Here $\frac{\mathrm{d}\mu}{\mathrm{d}\pi}$ is the Radon-Nikodym derivative. As stated in the introduction, the utility of the KL divergence in PAC-Bayes bounds comes from the following the change of measure formula. This was first stated by Kullback (1959) for finite parameter spaces, and then proved more generally by Donsker and Varadhan (1975) and Csiszár (1975).

Lemma 3 (Change of Measure) *Let $h : \Theta \to \mathbb{R}$ be a measurable function. For any $\nu \in \mathcal{M}(\Theta)$,*

$$\log \mathbb{E}_{\theta\sim\nu} \exp(h(\theta)) = \sup_{\rho\in\mathcal{M}(\Theta)} \big\{\mathbb{E}_{\theta\sim\rho}[h(\theta)] - D_{\mathrm{KL}}(\rho\|\nu)\big\}.$$

While the Donsker-Varadhan formula is the most popular change of measure formula, it is not unique in its ability to furnish PAC-Bayes bounds. In Appendix 6.2, we provide change of measure inequalities for ϕ and Rényi divergences and discuss how we can use such formulas in our bounds.

3. A General Recipe for Stochastic Processes

We now present results for nonnegative processes upper bounded by either a supermartingale or a reverse submartingale. We will consider processes $P(\theta) = (P_t(\theta))_{t\geqslant 1}$ which are functions of a parameter $\theta \in \Theta$. While the following theorem does not appear to be in the form of a traditional PAC-Bayes bound, a variety of typical bounds can be recovered by considering particular processes $P(\theta)$ (Table 1). Many such fruitful processes will be presented throughout the remainder of this manuscript.

Theorem 4 (Master anytime PAC-Bayes bound) *For each $\theta \in \Theta$, assume that a stochastic process of interest, $P(\theta) = (P_t(\theta))_{t=t_0}^{\infty}$, is upper bounded by another process $U(\theta) = (U_t(\theta))_{t=t_0}^{\infty}$, which is such that $\exp U(\theta)$ is either a supermartingale or a reverse submartingale satisfying $\mathbb{E}_{\mathcal{D}}[\exp U_{t_0}(\theta)] \leqslant 1$. Then for any $\delta \in (0,1)$ and prior $\nu \in \mathcal{M}(\Theta)$, with probability at least $1-\delta$, we have that for all $t \geqslant t_0$ and $\rho \in \mathcal{M}(\Theta)$,*

$$\mathbb{E}_\rho P_t(\theta) \leqslant D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta). \tag{4}$$

In fact, the bound also holds with ρ being replaced by ρ_t on both sides of (4) for any adapted sequence of posteriors $(\rho_t)_{t\geqslant t_0}$.

Note that the KL divergence in (4) can be replaced by a variety of other divergences, provided they have their own variational representations (which they typically do). We discuss several alternative divergences in Sections 6.1 and 6.2.

Proof For $t \geqslant t_0$, set

$$V_t^{\mathrm{mix}} := \exp \sup_\rho \big\{\mathbb{E}_{\theta\sim\rho}[U_t(\theta)] - D_{\mathrm{KL}}(\rho\|\nu)\big\}.$$

If $\exp U(\theta)$ is a supermartingale (resp., reverse submartingale), then we claim (V_t^{mix}) is a supermartingale (resp., reverse submartingale). Indeed, Lemma 3 gives $V_t^{\mathrm{mix}} = \mathbb{E}_\nu \exp U_t(\theta)$, so V_t^{mix} is a mixture of supermartingales or reverse submartingales, which is itself a supermartingale or reverse submartingale (Lemma 46). Applying Ville's inequality (either Lemma 1 or 2), we obtain

$$\begin{aligned}
&\mathbb{P}(\exists t \geqslant t_0 : \exp \sup_\rho \big\{\mathbb{E}_\rho P_t(\theta) - D_{\mathrm{KL}}(\rho\|\nu)\big\} \geqslant 1/\delta) \\
&\leqslant \mathbb{P}(\exists t \geqslant t_0 : \exp \sup_\rho \big\{\mathbb{E}_\rho U_t(\theta) - D_{\mathrm{KL}}(\rho\|\nu)\big\} \geqslant 1/\delta) \\
&= \mathbb{P}(\exists t \geqslant t_0 : V_t^{\mathrm{mix}} \geqslant 1/\delta) \leqslant \mathbb{E}_{\mathcal{D}}[V_{t_0}^{\mathrm{mix}}]\delta \leqslant \delta,
\end{aligned}$$

where the first inequality follows since $P_t(\theta) \leqslant U_t(\theta)$ by assumption. The final inequality follows since ν is data-free, enabling Fubini's theorem to be applied: $\mathbb{E}[V_{t_0}^{\text{mix}}] = \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\nu} \exp U_{t_0}(\theta) = \mathbb{E}_{\nu}\mathbb{E}_{\mathcal{D}} \exp U_{t_0}(\theta) \leqslant 1$. Thus, with probability $1-\delta$, for all $t \geqslant t_0$, $\exp \sup_{\rho \in \mathcal{M}(\Theta)}\{\mathbb{E}_{\rho} P_t(\theta) - D_{\text{KL}}(\rho\|\nu)\} \leqslant 1/\delta$. Taking logarithms gives the desired result. ■

Several remarks are in order. First, the final sentence of Theorem 4 highlights that the uniformity of time and probability measures implies that the bound holds over all *sequences* of posteriors. This is the form in which we expect the result to be most useful. A concrete example of changing posteriors is given in Section 6.6, where we apply our result to Gaussian process classification. Second, it's worth noting that Theorem 4 posits no distributional assumptions on the underlying data. Indeed, it does not even assume that the underlying filtration is the canonical data filtration. While our examples in subsequent sections will use either the canonical forward filtration $\mathcal{F}_t = \sigma(Z^t)$ or a particular backward "exchangeable" filtration $(\mathcal{E}_t)$, Theorem 4 holds for more general processes. Third, we note also that we need not specify that ρ be absolutely continuous with respect to the prior ν in inequality (4) since, if not, then $D_{\text{KL}}(\rho\|\nu) = \infty$ and the bound holds trivially. Finally, in addition to bounding $\mathbb{E}_{\mathcal{D}}[V_1^{\text{mix}}]$, the fact that the prior ν is data free is required by Lemma 46. That is, it is required to ensure that $\mathbb{E}_{\nu} \exp U_t(\theta)$ is a super/submartingale.

Condition on $(f_t)_{t\geqslant 1}$	*Condition on* $(Z_t)_{t\geqslant 1}$	*Results*
SubGaussian or subexponential	No explicit assumption	Corollaries 7, 9
Bounded	No explicit assumption	Corollaries 12, 13, 14
Bernstein	No explicit assumption	Corollary 11
Bounded MGF	No explicit assumption	Corollary 16
$\mathbb{E}[f_t^2(Z_t,\theta)\|\mathcal{F}_{t-1}] < \infty$	No explicit assumption	Corollaries 17, 18
Stn. & MGF of $\varphi_t(\theta)$ exists	Exchangeable	Corollaries 22, 24, 30, 34
Stn. & bounded in [0,1]	i.i.d.	Corollaries 25, 26, 15, 27, 42

Table 2: A summary of the conditions on the loss and the data required by several bounds. "Stn" stands for stationary. Even though for most rows there is no explicit dependence assumption required of (Z_t), the usefulness of the bounds or the establishment of conditions on (f_t) may sometimes require implicitly making distributional assumptions on the data, but these will often be (much) less restrictive than an i.i.d. assumption. See Section 4.2 after Corollary 7 for more discussion. As all results require (f_t) to be predictable, this requirement is disregarded above. We omit results from Section 6.4 (martingale difference sequences) as the setting is slightly different.

4. PAC-Bayes Bounds via Supermartingales

We first construct PAC-Bayes bounds via supermartingales in light of Theorem 4. Our general framework for doing so is based on *sub-ψ-processes* (Howard et al., 2020), which are generalizations of processes amenable to exponential concentration inequalities. Many standard concentration inequalities (e.g., Hoeffding, Bennett, Bernstein) implicitly use sub-ψ

processes which, if identified, yield time-uniform Chernoff bounds (Howard et al., 2020). For our purposes, sub-ψ processes can be used in Theorem 4 to yield a time-uniform PAC-Bayes bound (Corollary 6). Many existing PAC-Bayes bounds rely on fixed-time concentration inequalities which can be generalized to sub-ψ processes, thus yielding time-uniform extensions. We begin by defining sub-ψ processes and then proceed to give explicit bounds for light-tailed losses (Section 4.2), and then for heavier-tailed losses (Section 4.3).

4.1 The sub-ψ Condition

Roughly speaking, a sub-ψ process is a stochastic process which is upper bounded by a supermartingale but takes a particular functional form. They are at the heart of recent progress on time-uniform Chernoff bounds (Howard et al., 2020). This section presents a corollary of Theorem 4 for sub-ψ processes which, in turn, yields many time-uniform extensions of existing PAC-Bayes bounds. We find that many existing bounds are implicitly relying on sub-ψ processes without recognizing it.

Definition 5 (Sub-ψ process) *Let $(S_t)_{t=1}^{\infty} \subseteq \mathbb{R}$ and $(V_t)_{t=1}^{\infty} \subseteq \mathbb{R}_{\geqslant 0}$ be stochastic processes adapted to an underlying filtration $(\mathcal{F}_t)_{t=1}^{\infty}$. For a function $\psi : [0, \psi_{max}) \to \mathbb{R}$, we say (S_t, V_t) is a sub-ψ process if, for every $\lambda \in [0, \psi_{max})$, there exists some supermartingale $(L_t(\lambda))_{t=1}^{\infty}$ with $L_1(\lambda) \leqslant 1$ such that, for all $t \geqslant 1$,*

$$\exp\{\lambda S_t - \psi(\lambda) V_t\} \leqslant L_t(\lambda), \ \textit{a.s.} \tag{5}$$

Definition 5 may appear rather abstract at first glance. Useful intuition comes from considering what happens when (S_t) is a martingale. In this case, $(\exp(\lambda S_t))$ is a submartingale by Jensen's inequality. Thus, $\psi(\lambda)V_t$ must be a process which appropriately "dominates" S_t in order to ensure that $\exp(\lambda S_t - \psi(\lambda)V_t)$ decreases in expectation rather than increases. For instance, suppose $X_1, X_2, \dots$ are i.i.d. with mean 0. If $S_t = \sum_{i \leqslant t} X_t$, then taking $\psi(\lambda)$ to be the log-MGF $\log \mathbb{E} e^{\lambda X_1}$ and $V_t = t$ is sufficient to turn $\exp(\lambda S_t - \psi(\lambda)V_t)$ into a martingale. Indeed, $\mathbb{E}[\exp(\lambda S_t - \psi(\lambda)V_t)|\mathcal{F}_{t-1}] = \prod_{i=1}^{t} \mathbb{E}[\exp(\lambda X_i - \log \mathbb{E} e^{\lambda X_1})|\mathcal{F}_{t-1}] = \prod_{i=1}^{t-1} \exp(\lambda X_i - \log \mathbb{E} e^{\lambda X_1})$. Corollary 16 gives a PAC-Bayes bound based on this process. Another example comes from supposing the X_i are σ-subGaussian. In that case we may take $\psi(\lambda) = \lambda^2\sigma^2/2$, keeping S_t and V_t the same. Then $\exp(\lambda S_t - \psi(\lambda)V_t)$ is a supermartingale (as opposed to a martingale). This process is used (albeit in more generality) by Corollary 7. If, as in the examples above, S_t is a sum then we may let $\lambda = \lambda_t$ change as a function of time. This will be the case in the majority of our bounds. Finally, notice that in these examples, we may simply take $L_t(\lambda) = \exp(\lambda S_t - \psi(\lambda)V_t)$, meaning that the exponential process is itself a supermartingale. This is often the case. We refer the reader to Howard et al. (2020) for a more lengthy discussion and further examples.

A nonnegative process that is upper bounded by a supermartingale (but may or may not itself be a supermartingale) has recently been termed an "e-process" (Ramdas et al., 2023). Theorem 4 yields bounds for such processes. Instead of working with more general definitions, however, we prefer to base our discussion on sub-ψ processes specifically because it's helpful to consider particular functions ψ and processes (V_t) which can bound our process

(S_t) of interest. More to the point, we will often consider S_t to be the martingale

$$\sum_{i=1}^{t} \mathbb{E}_{\mathcal{D}}[f_i(Z,\theta)|\mathcal{F}_{i-1}] - f_i(Z_i,\theta). \tag{6}$$

Different assumptions on f_t (e.g., bounded, light-tailed, heavy-tailed) will then lead us to particular selections of ψ and (V_t). Moreover, our PAC-Bayes inequalities will bound S_t in terms of ψ and V_t. Consequently, if one finds themselves dealing with a sub-ψ process, then the form of the bound will be immediately apparent.

As we did for more general processes, we will consider sub-ψ processes which are indexed by parameters $\theta \in \Theta$ and we will write that $(S_t(\theta), V_t(\theta))$ is a sub-ψ process. This should be taken to mean that, for each fixed θ, $\exp\{\lambda S_t(\theta) - \psi(\lambda)V_t(\theta)\} \leqslant L_t(\lambda,\theta)$ for an appropriate supermartingale $L_t(\lambda,\theta)$. Since, by construction, sub-ψ processes are nonnegative and upper bounded by a supermartingale with unit initial value, we obtain the following corollary of Theorem 4.

Corollary 6 *Assume that for each $\theta \in \Theta$, $(S_t(\theta), V_t(\theta))$ is a sub-ψ process. Let $\nu \in \mathcal{M}(\Theta)$ be a data-free prior and let $\lambda \in [0, \psi_{max})$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have that*

$$\mathbb{E}_{\theta\sim\rho}[\lambda S_t(\theta) - \psi(\lambda)V_t(\theta)] \leqslant D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta), \tag{7}$$

for all times $t \geqslant 1$ and $\rho \in \mathcal{M}(\Theta)$.

4.2 Light-tailed losses

Here we return to our main problem setting and consider anytime bounds on the difference between the expected risk and the empirical risk. By choosing particular sub-ψ processes and applying Corollary 6, we can develop anytime bounds for light-tailed losses (this section) and more general losses (Section 4.3). It will often be useful to consider the quantity

$$\Delta_i(\theta) := \mu_i(\theta) - f_i(Z_i,\theta),$$

where $\mu_i(\theta) = \mathbb{E}_{\mathcal{D}}[f_i(Z_i,\theta)|\mathcal{F}_{i-1}]$. Note that the process $(\sum_{i\leqslant t}\Delta_i(\theta))_{t\geqslant 1}$ is a martingale, but it is not nonnegative. Throughout the remainder of this section, the underlying filtration will be the canonical data filtration $\mathcal{F}_t = \sigma(Z_1,\ldots,Z_t)$.

4.2.1 SubGaussian losses

We begin by giving an anytime-valid PAC-Bayes bound for subGaussian losses. Recall that a random variable Y is *σ-subGaussian conditional on $\mathcal{F}$* if $\mathbb{E}[\exp(s(Y-\mathbb{E}[Y]))|\mathcal{F}] \leqslant \exp(s^2\sigma^2/2)$ for all $s \in \mathbb{R}$. We will say the loss f_t is σ-subGaussian if $f_t(Z_t,\theta)$ is σ-subGaussian for all $\theta \in \Theta$.

Corollary 7 *Let (Z_t) be a stream of (not necessarily i.i.d.) data. Let $(f_t)_{t=1}^{\infty}$ be a predictable sequence of loss functions such that f_i is σ_i-subGaussian conditional on $\mathcal{F}_{i-1}$. Let (λ_t) be a nonnegative predictable sequence and consider any data-free prior $\nu \in \mathcal{M}(\Theta)$.*

Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all t and measures $\rho \in \mathcal{M}(\Theta)$,

$$\sum_{i=1}^{t} \lambda_i \mathbb{E}_\rho \Delta_i(\theta) \leqslant \sum_{i=1}^{t} \frac{\lambda_i^2 \sigma_i^2}{2} + D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta).$$

The proof is in Appendix A.1. Suppose the loss is stationary and bounded in $[0,H]$, implying that it is $H/2$-subGaussian. If $\lambda_i = \lambda$ is constant, then Corollary 7 implies that with probability at least $1-\delta$,

$$\mathbb{E}_\rho \mathbb{E}_{\mathcal{D}} f(Z,\theta) \leqslant \mathbb{E}_\rho \widehat{R}_t(\theta) + \frac{\lambda H^2}{8} + \frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta)}{\lambda t}, \tag{8}$$

for all times t and measures $\rho \in \mathcal{M}(\Theta)$. For any fixed time n of special interest, setting $\lambda_i = \lambda/n$ for all λ recovers (2) (Catoni's bound) exactly at time n, but still makes a nontrivial claim for all $t \neq n$ at no extra cost. This time-uniform bound for bounded losses was recently also given by Haddouche and Guedj (2023). As noted previously, it generalizes well-known fixed-time bounds of the same flavour (Catoni, 2003, 2004, 2007; Alquier et al., 2016). This phenomenon of exactly recovering a fixed-time Chernoff-style bound by a more general time-uniform bound was a central contribution of the unified "supermartingale + Ville" framework of Howard et al. (2020).

Remark 8 *As in Corollary 7, the remainder of Section 4 is concerned with bounding the conditional risk $\frac{1}{t}\sum_{i=1}^{t} \mathbb{E}_{\theta\sim\rho}\mu_i(\theta)$ where $\mu_i(\theta) = \mathbb{E}[f_i(Z_i,\theta)|\mathcal{F}_{i-1}]$. However, we will often state bounds on $\sum_{i=1}^{t} \lambda_i \mathbb{E}_{\theta\sim\rho}\mu_i(\theta)$, where $(\lambda_t)_{t\geqslant 1}$ is a predictable sequence of positive scalars. Considering such sequences is useful if the conditional risk is constant as a function of t, i.e., $R(\theta) = \mu_t(\theta)$ for all t, as we can then remove $R(\theta)$ from the sum and divide by $\sum_{i=1}^{t}\lambda_i$. Values of λ_t can be chosen such that difference between $\mathbb{E}_\rho R(\theta)$ and $t^{-1}\sum_{i\leqslant t}\mathbb{E}_\rho f_i(Z_i,\theta)$ — the* width *of the bounds — goes asymptotically to zero with t. This has been called the method of "predictable plug-ins" (see, e.g., Waudby-Smith and Ramdas (2023)).*

On the other hand, if $\mu_t(\theta)$ is changing with time, then we must select $\lambda_i = \lambda$ to be constant in order to isolate the mean. In this case, we can still achieve bounds with widths that go to zero, but via a different (and more complicated) method of applying different bounds over geometrically spaced epochs. We provide details in Section 6.3. Otherwise, if such a method is not used, one may still select λ as a function of some fixed-time n, in which case the bound will be tight at that point but progressively looser as the number of samples t moves away from n (the bound will remain valid at all times, however).

The lack of independence assumptions in Corollary 7 may seem surprising at first, but it is another consequence of the supermartingale approach. The proof of the Corollary is based on the process

$$N_t(\theta) := \prod_{i=1}^{t} \exp\left\{\lambda_i(\mu_i(\theta) - f(Z_i,\theta)) - \frac{\lambda_i\sigma_i^2}{2}\right\}. \tag{9}$$

Since $N_{t-1}(\theta)$ is $\mathcal{F}_{t-1}$ measurable,

$$\mathbb{E}[N_t(\theta)|\mathcal{F}_{t-1}] = N_{t-1}(\theta)\cdot \mathbb{E}\exp\left\{\lambda_t(\mu_t(\theta) - f_t(Z_t,\theta)) - \frac{\lambda_t\sigma_t^2}{2}\middle|\mathcal{F}_{t-1}\right\}.$$

By definition of (conditional) subGaussianity, the expected value term in the above display is at most 1. This demonstrates that $(N_t(\theta))$ is a nonnegative supermartingale, meaning that Theorem 4 applies. The same reasoning holds for other bounds we will present: if $\exp\{\lambda_t \Delta_t(\theta) - g_t | \mathcal{F}_{t-1}\}$ has expectation at most 1, then $(\exp\{\sum_i \lambda_i \Delta_i(\theta) - \sum_i g_i\})_{t \geqslant 1}$ is a supermartingale, yielding a time-uniform PAC-Bayes bound with no independence assumptions on the data. However, we feel it important to emphasize that there is no free lunch. Despite there being no such assumptions, the fact that we must have $\mathbb{E}[f_t(Z_t, \theta)|\mathcal{F}_{t-1}] < \infty$ is implicitly relying on a type of dependence between f_t and the past. In some sense, the lack of distributional assumptions places the burden on (f_t) as opposed to (Z_t). Thus, while the mathematics holds with no conditions on (Z_t), the bounds may be meaningless for very "ill-behaved" data and/or losses.

We now present a novel bound for subGaussian losses based on the method of mixtures. Fix $\lambda_i = \lambda$ above to consider the supermartingale $M_t(\lambda, \theta) := \prod_{i=1}^t \exp\left\{\lambda \Delta_i(\theta) - \frac{\lambda^2}{2}\sigma_i^2\right\}$. As discussed in Section 2 and proven in Appendix B, the mixture

$$M_t(\theta) := \int_{\lambda \in \mathbb{R}} M_t(\lambda, \theta) \mathrm{d}F(\lambda), \tag{10}$$

is also a nonnegative supermartingale for an appropriate distribution F. By choosing F to be Gaussian with mean 0 and some fixed variance, we can generate the following bound. The proof is in Appendix A.2.

Corollary 9 (Gaussian-mixture bound for subGaussian losses) *Let $Z_1, Z_2, \ldots$ be a stream of (not necessarily i.i.d.) data. Let $(f_t)_{t=1}^\infty$ be a predictable sequence of loss functions such that f_i is σ_i-subGaussian. Let $\nu \in \mathcal{M}(\Theta)$ be a data-free prior. Then, for all $\delta \in (0,1)$ and $\beta > 0$, with probability at least $1-\delta$ over the random draw of (Z_t), for all times t and measures $\rho \in \mathcal{M}(\Theta)$,*

$$\sum_{i=1}^t \mathbb{E}_\rho \Delta_i(\theta) \leqslant \left(\frac{s_t(\beta)}{\beta}\left(D_{\mathrm{KL}}(\rho\|\nu) + \log\frac{s_t(\beta)}{\delta}\right)\right)^{1/2}, \tag{11}$$

where $s_t(\beta) = 1 + \beta \sum_{i=1}^t \sigma_i^2$.

The parameter β comes from the variance of the Gaussian mixture in (10). It is worth comparing the above bound to the one from McAllester (1999). Considering stationary loss functions bounded in $[0,1]$, McAllester's fixed time bound reads

$$\mathbb{E}_\rho R(\theta) \leqslant \mathbb{E}_\rho \widehat{R}_n(\theta) + \left(\frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(n/\delta)}{2(n-1)}\right)^{1/2}. \tag{12}$$

In our case, f being bounded implies that $\sigma_i^2 = 1/4$ for all i since f is 1/2-subGaussian. Fix a time n of interest and take β such that $s_n(\beta) = n$, i.e., $\beta = 4(n-1)/n$. The Gaussian mixture bound (11) then yields McAllester's bound, but tighter by a factor of $\sqrt{2}$. Meanwhile, we can achieve a time-uniform version of McAllester's bound by considering $\beta = 1$, in which case $s_t(\beta) = 1 + t/4 \leqslant t$ for all $t \geqslant 2$ and the bound becomes

$$\mathbb{E}_\rho R(\theta) \leqslant \mathbb{E}_\rho \widehat{R}_t(\theta) + \left(\frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(t/\delta)}{t}\right)^{1/2}, \tag{13}$$

which is looser than McAllester's by a factor of $\sqrt{2}$. We might thus consider (13) to be a time-uniform generalization of McAllester's bound. However, this was for a particular choice of β. In general, our bound contains the parameter β over which we can optimize. Performing this optimization gives an implicit equation for β:

$$\log(s_t(\beta)) + \frac{1}{\beta} = \log(\delta) - D_{\mathrm{KL}}(\rho\|\nu).$$

(Though note that the result should not depend on t unless it is fixed in advance.) This is difficult to solve in closed-form, but after choosing ν and ρ and computing the KL divergence, we might generate an approximate solution computationally. Section 5 will explore another generalization of McAllester's bound using a separate (reverse submartingale based) technique and Section 4.2.4 will discuss yet another generalization using betting martingales.

Remark 10 *Corollaries 7 and 9 may be strengthened to handle* sub-exponential *losses, where we say that Y is subexponential with parameters (σ, c) if $\mathbb{E}\exp(s(Y-\mathbb{E}Y)) \leqslant \exp(s^2\sigma^2/2)$ for all $|s| \leqslant 1/c$. SubGaussian random variables are subexponential random variables with $c = 0$. To extend Corollary 7 to subexponential variables, we take $\lambda_i \leqslant 1/c_i$ if f_i is subexponential with parameter (σ_i, c_i).*

It is worth noting here that the method of mixtures has been previously employed in the PAC-Bayesian literature. For instance, it was used by Kuzborskij and Szepesvári (2019), who were interested in providing bounds on $h(Z^n, \theta) - \mathbb{E}_{\mathcal{D}}[h(Z^n, \theta)]$, where $h : \mathcal{Z}^n \times \Theta \to \mathbb{R}$ is a measurable function and $n \in \mathbb{N}$. Here $Z^n = (Z_1, \ldots, Z_n)$, where we assume all elements Z_i are drawn i.i.d. from $\mathcal{D}$. Kuzborskij and Szepesvári (2019) give bounds based on an Efron-Stein variance proxy:

$$V(\theta) = \sum_{i=1}^{n} V_i(\theta), \quad \text{where} \quad V_i(\theta) = \mathbb{E}[(h(Z^n, \theta) - h(Z^{(i)}, \theta))^2 | \mathcal{F}_i], \tag{14}$$

where $Z^{(i)}$ is the same as Z^n but contains an independent copy of Z_i. Equations (17) and (18) in Kuzborskij and Szepesvári (2019) show that the Doob-decomposition of $h(Z^n, \theta) - \mathbb{E}[h(Z^n, \theta)]$ for a fixed θ obeys a sub-ψ condition with respect to $\{V_t(\theta)\}$. Denoting $D_i(\theta) = \mathbb{E}[h(Z^n, \theta)|\mathcal{F}_i] - \mathbb{E}[h(Z^n, \theta)|\mathcal{F}_{i-1}]$, they showed that

$$\mathbb{E}\left[\exp\left(\lambda D_i(\theta) - \frac{\lambda^2}{2} V_i(\theta)\right) \middle| \mathcal{F}_{i-1}\right] \leqslant 1; \quad 1 \leqslant i \leqslant n.$$

The processes $\{(P_t(\theta))_{t\geqslant 1} : \theta \in \Theta\}$ where $P_t(\theta) = \prod_{i\leqslant t} \exp\{\lambda D_i(\theta) - \frac{\lambda^2}{2} V_i(\theta)\}$ thus constitute a family of supermartingales, and mixing over the family yields another supermartingale. Theorems 3 and 4 of Kuzborskij and Szepesvári (2019) are based on such a mixture (again using a Gaussian mixture distribution), and provide bounds on $h(Z^n, \theta) - \mathbb{E}_{\mathcal{D}}[h(Z^n, \theta)]$ in terms of $V(\theta)$. Their result thus follows as a consequence of Corollary 6 and the method of mixtures. Note, however, that time-uniformity apparently does not gain us much in this case because the empirical risk (in this case $h(Z^n, \theta)$) is not computable until time $t = n$.

4.2.2 Losses obeying a Bernstein condition.

The consideration of subexponential random variables in Remark 10 naturally leads us to consider a Bernstein condition on the losses, which implies that they're subexponential. In particular, we say that a random variable Y satisfies *Bernstein's condition* with parameter c if

$$|\mathbb{E}[(Y - \mathbb{E}(Y))^k]| \leqslant \frac{1}{2}\mathsf{Var}(Y)k!c^{k-2}, \quad \forall k \in \mathbb{N},\ k \geqslant 2.$$

It is well known that if Y is Bernstein with parameter c then it is subexponential with parameters $(\sqrt{2\mathsf{Var}(Y)}, 1/2c)$ (see, e.g., Boucheron et al., 2013, Theorem 2.10 or Wainwright, 2019, Corollary 2.10). For bounded random variables, the resulting concentration inequality can be much tighter than Hoeffding's (which is not variance adaptive), especially when the variance of Y is much smaller than its range. It is therefore worth stating the following PAC-Bayes result for Bernstein-type losses, the proof of which is in Appendix A.6.

Corollary 11 (Bernstein condition anytime bound) *Let (Z_t) be a stream of (not necessarily i.i.d.) data. Let (f_t) be a predictable sequence of functions with $\mathsf{Var}(f_t(Z_t, \theta)|\mathcal{F}_{t-1}) \leqslant \sigma_t^2$ and, for all t and integers $k \geqslant 2$, $|\mathbb{E}_{\mathcal{D}}[(f_t(Z_t, \theta) - \mu_t(\theta))^k|\mathcal{F}_{t-1}]| \leqslant \frac{1}{2}\sigma_t^2 k! c_t^{k-2}$. Let (λ_t) be a predictable sequence such that $\lambda_t \in (0, 1/c_t)$ for all t. Fix a prior $\nu \in \mathcal{M}(\Theta)$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all times $t \geqslant 1$ and measures $\rho \in \mathcal{M}(\Theta)$, we have*

$$\sum_{i=1}^{t} \lambda_i \mathbb{E}_\rho \Delta_i(\theta) \leqslant \sum_{i=1}^{t} \frac{\lambda_i^2 \sigma_i^2}{2(1 - c_i\lambda_i)} + D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta).$$

To our knowledge, this result (even the implied fixed-time counterpart) is new to the PAC-Bayes literature.

4.2.3 Bounded losses

The next few results consider bounded loss functions. The first relies on a Bernstein-type process

$$B_t(\theta) := \prod_{i=1}^{t} \exp\left\{\lambda_i \Delta_i(\theta) - \lambda_i^2 (e-2)\mathbb{E}[\Delta_i^2(\theta)|\mathcal{F}_{i-1}]\right\}. \tag{15}$$

It is so termed because $(B_t(\theta))$ can be seen to be a supermartingale via the application of Bernstein's inequality. The details are in the proof of the following Proposition, which can be found in Appendix A.3. The resulting bound has been applied to martingale difference sequences (MDSs) (Seldin et al., 2012, Theorem 7). Corollary 38 gives the precise time-uniform extension of the MDS result.

Corollary 12 (Bernstein-like anytime bound for bounded losses) *Let (Z_t) be a stream of (not necessarily i.i.d.) data. Let (f_t) be a predictable sequence of loss functions such that $\|f_t\|_\infty \leqslant H_t$ for all t and constants $H_t > 0$. Let (λ_t) be a predictable sequence such that $\lambda_t \in [0, 1/H_t]$ for all t. Fix a prior $\nu \in \mathcal{M}(\Theta)$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all times t and measures $\rho \in \mathcal{M}(\Theta)$, we have*

$$\sum_{i=1}^{t} \lambda_i \mathbb{E}_\rho \Delta_i(\theta) \leqslant (e-2)\sum_{i=1}^{t} \lambda_i^2 \mathbb{E}_\rho \mathbb{E}_{\mathcal{D}}[\Delta_i^2(\theta)|\mathcal{F}_{i-1}] + D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta).$$

A second result for bounded losses can be obtained via a supermartingale based on a Bennett-like inequality (Boucheron et al., 2013, Theorem 2.9). It is the first example in this paper of a result where the empirical risk is bounded by the expected risk. That is, it is a bound on $-\Delta_i(\theta)$. The reader can be forgiven for wondering whether such bounds are useful. However, Catoni's MGF-based PAC-Bayes bound (Catoni, 2007) is also an example of such a bound and has found various uses, such as in estimating means of random vectors and matrices (Catoni and Giulini, 2017). We therefore opt to include the next result. More discussion can be found in the next section when we present a time-uniform extension of Catoni's bound (Corollary 16).

Corollary 13 (Bennet-like anytime bound for bounded losses) *Let (Z_t) be a stream of (not necessarily i.i.d.) data. Let (f_t) be a predictable sequence of loss functions such that $\|f_t\|_\infty \leqslant H_t$ for all t and constants $H_t > 0$. Let (λ_t) be a predictable sequence of positive values with $\lambda_t < \inf_\theta\{1/\mathbb{E}_{\mathcal{D}}[f_t(Z_t,\theta)|\mathcal{F}_{t-1}]\}$. Fix a prior $\nu \in \mathcal{M}(\Theta)$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all times t and measures $\rho \in \mathcal{M}(\Theta)$, we have*

$$\sum_{i=1}^{t} \lambda_i \mathbb{E}_\rho(f_i(Z_i,\theta) - \mu_i(\theta)) \leqslant \sum_{i=1}^{t} \frac{\mathbb{E}_{\rho,\mathcal{D}}[f_i^2(Z_i,\theta)|\mathcal{F}_{t-1}]}{H_i^2} \psi_P(\lambda_i H_i) + D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta),$$

where $\psi_P(x) = (e^x - x - 1)$.

The proof is in Appendix A.4. Both Corollary 12 and 13 are based on one-sided concentration inequalities and thus hold in the more general setting when losses are not non-negative. The subscript in ψ_P references sub-Poisson processes (Howard et al., 2020).

Our final result for bounded losses comes via an "unexpected Bernstein inequality" provided by Mhammedi et al. (2019, Lemma 13) and based on an inequality in Fan et al. (2015). More specifically, Fan et al. (2015, Equation (4.11)) demonstrate that for random variables $X \geqslant -1$ and $\lambda \in [0,1)$,

$$\mathbb{E}\exp\{\lambda X + (\lambda + \log(1-\lambda))X^2\} \leqslant 1.$$

This bound was extended to derive empirical Bernstein concentration inequalities and confidence sequences in Howard et al. (2021). Following this lead, Mhammedi et al. (2019) use Fan's inequality to show that for a random variable $X \in (-\infty, b)$, for all $0 \leqslant \lambda < 1/b$,

$$\exp\{\lambda(\mathbb{E}[X] - X - cX^2)\} \leqslant 1, \quad \forall c \geqslant \lambda\vartheta(\lambda b). \tag{16}$$

where $\vartheta(\alpha) = \frac{-\log(1-\alpha)-\alpha}{\alpha^2}$. We can use such an inequality to construct a supermartingale which furnishes the following result.

Corollary 14 (Unexpected Bernstein anytime bound for bounded losses) *Let (Z_t) be a stream of (not necessarily i.i.d.) data. Let (f_t) be predictable sequence of loss functions such that $\|f_t\|_\infty \leqslant H_t$ for all t and constants $H_t > 0$. Let (λ_t) be a predictable sequence of positive values such that $0 \leqslant \lambda_t \leqslant 1/H_t$. Let (c_t) be a predictable sequence with*

$c_t \geqslant \lambda_t \vartheta(\lambda_t H_t)$. Fix a prior $\nu \in \mathcal{M}(\Theta)$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all times t and measures $\rho \in \mathcal{M}(\Theta)$, we have

$$\sum_{i=1}^{t} \lambda_i \mathbb{E}_\rho \Delta_i(\theta) \leqslant \sum_{i=1}^{t} \lambda_i c_i \mathbb{E}_\rho f_i^2(Z_i, \theta) + D_{\mathrm{KL}}(\rho \| \nu) + \log(1/\delta).$$

A big distinction between Corollary 14 and Corollaries 12 and 13 is the lack of an expectation over $\mathcal{D}$ on the right hand side. Instead, we work directly with the random variables $f_t(Z_t, \theta)$. While the process used in the proof of Corollary 14 is at the core of the result of Mhammedi et al. (2019), our result is not a time-uniform version of theirs. Indeed, they employ several tools which make an anytime-valid extension challenging, such as the use of data-dependent priors. We discuss such priors more in Section 6.5.

4.2.4 Interlude: Implicit Bounds via Wealth Processes

There has been a recent surge of interest in so-called game-theoretic probability and statistics (Shafer and Vovk, 2019; Ramdas et al., 2023) owing to its fresh perspective on sequential, anytime-valid inference. Here we demonstrate how some of the ideas may be employed to generate PAC-Bayes bounds, and how this perspective recovers some concurrent work by Jang et al. (2023).

Central to game-theoretic statistics is the idea of a fictitious bettor playing an iterated game against nature. The game is structured as follows. The bettor begins with an initial wealth of $\mathcal{K}_0 = 1$. At time t, the bettor chooses a $\mathcal{F}_{t-1}$-measurable *payoff function* $S_t : \mathcal{V} \to [0, \infty]$ obeying $\mathbb{E}[S_t(V)|\mathcal{F}_{t-1}] \leqslant 1$, where the expectation is taken with respect to some strategically chosen distribution(s) P. For instance, in hypothesis testing problems, P is chosen to be the set of distributions comprising the null. See Ramdas et al. (2023) for more details. Nature then reveals a value $V_t \in \mathcal{V}$ and the bettor updates his wealth as $\mathcal{K}_t = \mathcal{K}_{t-1} \cdot S_t(V_t)$. The total wealth of the bettor at time t is therefore $\mathcal{K}_t = \prod_{i=1}^{t} S_i(V_i)$, and the process $(\mathcal{K}_t)_{t \geqslant 0}$ is guaranteed to be a supermartingale (on P) due to the assumption on the payoff function.

We apply this to the PAC-Bayes setting as follows. Assume that the data are i.i.d. and that the losses are stationary and bounded in $[0,1]$. We will consider playing a game for each parameter $\theta \in \Theta$, and will thus have a family of wealth processes $\{(\mathcal{K}_t(\theta))_{t \geqslant 0} : \theta \in \Theta\}$. We take the values V_t to be the losses $f(Z_t, \theta)$. Following recent work in this area, suppose we use the following payoff function for each θ: $S_t(f(Z_t, \theta)) = 1 + \lambda_t(\theta)(f(Z_t, \theta) - \mu(\theta))$ where $(\lambda_t(\theta))_{t \geqslant 0}$ is a predictable sequence (often called a *betting strategy*) and we enforce that $\lambda_t(\theta) \in [-1/(1-\mu(\theta)), 1/\mu(\theta)]$ to ensure that S_t is nonnegative. Recalling that $\mu(\theta) = \mathbb{E}[f(Z_t, \theta)|\mathcal{F}_{t-1}]$, it's easy to verify that the resulting wealth process defined by

$$\mathcal{K}_t(\theta) = \prod_{i=1}^{t} \big\{1 + \lambda_i(\theta)(f(Z_i, \theta) - \mu(\theta))\big\}, \tag{17}$$

is a nonnegative martingale. Consequently, it may be employed in Theorem 4 where we take $\exp U_t(\theta)$ to be $\mathcal{K}_t(\theta)$. However, the results of Jang et al. (2023) concern not only the

wealth, but the *optimal* wealth, which is defined as

$$\mathcal{K}_t^*(\theta) := \max_{\lambda \in C(\mu(\theta))} \prod_{i=1}^{t} \big\{1 + \lambda(f(Z_i, \theta) - \mu(\theta))\big\}, \quad C(x) := \left[\frac{-1}{1-x}, \frac{1}{x}\right]. \tag{18}$$

Orabona and Jun (2023) show that there exists a betting strategy such that the wealth and the optimal wealth are related as

$$\log \mathcal{K}_t^*(\theta) - \log \mathcal{K}_t(\theta) \leqslant \log\left(\frac{\pi \Gamma(t+1)}{\Gamma(t+1/2)}\right). \tag{19}$$

This yields the following result, which is Theorem 1 of Jang et al. (2023).

Corollary 15 (Betting-based anytime bound) *Let (Z_t) be i.i.d. and f a stationary loss function bounded in [0,1]. Fix a data-free prior $\nu \in \mathcal{M}(\Theta)$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all times t and measures $\rho \in \mathcal{M}(\Theta)$,*

$$\mathbb{E}_\rho \log \mathcal{K}_t^*(\theta) \leqslant D_{\mathrm{KL}}(\rho \| \nu) + \log(1/\delta) + \log\left(\frac{\pi \Gamma(t+1)}{\Gamma(t+1/2)}\right). \tag{20}$$

Proof Since $(\mathcal{K}_t(\theta))$ is a nonnegative martingale, Theorem 4 applied to the wealth process immediately yields that with probability $1-\delta$ over (Z_t), $\mathbb{E}_\rho \log \mathcal{K}_t(\theta) \leqslant D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta)$. Applying (19) then finishes the proof. ■

By applying various inequalities to the term $\log(1+\lambda(f(Z_i,\theta)-\mu(\theta)))$, Jang et al. (2023) are able to recover (up to constants), Theorems of McAllester (1999), Maurer (2004), and Tolstikhin and Seldin (2013). We omit the details here and refer the reader to Propositions 2, 3, and 4 in Jang et al. (2023).

4.2.5 Losses with bounded MGF

Finally, we consider losses which may not be bounded or subGaussian but which have bounded moment generating functions (MGFs). The following bound is an anytime-valid version of Catoni's bound based on the log-MGF of the loss (Catoni, 2007). Like Corollary 13, it is somewhat of an unusual bound seeing as the empirical risk is "on the wrong side", i.e., we bound the empirical risk in terms of the log-MGF of the expected risk. However, as discussed above, the bound has proven useful in various estimation problems (Catoni and Giulini, 2017, 2018). It reads as follows. Suppose f is stationary and the data are i.i.d. Fix $n \in \mathbb{N}$ and a prior ν. Then, with probability at least $1-\delta$, for all ρ,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\rho f(Z_i, \theta) \leqslant \log \mathbb{E}_\rho \mathbb{E}_{\mathcal{D}}[\exp(f(Z,\theta))] + \frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta)}{n}. \tag{21}$$

Our time-uniform extension is given by Corollary 16. It recovers (21) exactly by taking $t = n$, $\lambda_i = 1$ for all i, and then dividing both sides by n (and assuming that the losses are stationary and the data i.i.d.).

Corollary 16 (Losses with bounded MGF) *Let $Z_1, Z_2, \dots$ be a stream of (not necessarily i.i.d.) data. Let (f_t) be a predictable sequence of loss functions. Let (λ_t) be a nonnegative predictable sequence and consider any data-free prior $\nu \in \mathcal{M}(\Theta)$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all times t and measures $\rho \in \mathcal{M}(\Theta)$,*

$$\sum_{i=1}^{t} \lambda_i \mathbb{E}_\rho f_i(Z_i, \theta) \leqslant \sum_{i=1}^{t} \log \mathbb{E}_\rho \mathbb{E}_{\mathcal{D}}[\exp(\lambda_i f_i(Z, \theta)) | \mathcal{F}_{i-1}] + D_{\mathrm{KL}}(\rho \| \nu) + \log(1/\delta). \tag{22}$$

The proof may be found in Appendix A.7.

4.3 More General Losses

Now we consider less well-behaved losses.

4.3.1 Losses with Bounded Second Moment.

Our second bound in this section assumes only that the conditional second moment of the loss is finite, i.e., $\mathbb{E}_{\mathcal{D}}[f_t^2(Z, \theta) | \mathcal{F}_{t-1}] < \infty$ for all $\theta \in \Theta$, and relies on the nonnegative process

$$M_t(\theta) := \prod_{i=1}^{t} \exp\left\{ \lambda_i \Delta_i(\theta) - \frac{\lambda_i^2}{2} \mathbb{E}_{\mathcal{D}}[f_i(Z_i, \theta)^2 | \mathcal{F}_{i-1}] \right\},$$

which can be seen to be a supermartingale via an application of a one-sided Bernstein inequality. Lemma 44 gives the relevant statement and proof of this result. As far as we are aware, the resulting PAC-Bayes bound is novel.

Corollary 17 (Losses with bounded conditional second moment) *Let $Z_1, Z_2, \dots$ be a stream of (not necessarily i.i.d.) data. Let (λ_t) be a nonnegative predictable sequence and consider any data-free prior $\nu \in \mathcal{M}(\Theta)$. Let (f_t) be a sequence of predictable loss functions such that $\sigma_t^2(\theta) = \mathbb{E}_{\mathcal{D}}[f_t^2(Z, \theta) | \mathcal{F}_{t-1}] < \infty$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all t and $\rho \in \mathcal{M}(\Theta)$,*

$$\sum_{i=1}^{t} \lambda_i \mathbb{E}_\rho \Delta_i(\theta) \leqslant \sum_{i=1}^{t} \frac{\lambda_i^2}{2} \mathbb{E}_\rho \sigma_i^2(\theta) + D_{\mathrm{KL}}(\rho \| \nu) + \log(1/\delta). \tag{23}$$

We now give another bound assuming only the second moment is finite. It is based on a supermartingale discovered by Bercu and Touati (2008) and the resulting bound (for stationary losses f_t and constant $\lambda = \lambda_i$) was given by Haddouche and Guedj (2023). Let

$$M_t(\theta) = \sum_{i=1}^{t} \Delta_i(\theta) = \sum_{i=1}^{t} (\mu_i(\theta) - f_i(Z_i, \theta)).$$

The *quadratic variation* of $M_t(\theta)$ is $[M(\theta)]_t := \sum_{i=1}^{t} \Delta_i^2(\theta)$ and its conditional quadratic variation is $\langle M(\theta) \rangle_t := \sum_{i=1}^{t} \mathbb{E}[\Delta_i^2(\theta) | \mathcal{F}_{i-1}]$. Bercu and Touati (2008) (see also Howard et al., 2020, Table 3) demonstrate that the process

$$L_t(\theta) = \exp\left\{ \lambda M_t(\theta) - \frac{\lambda^2}{6} \Big([M(\theta)]_t + 2\langle M(\theta) \rangle_t \Big) \right\},$$

is a supermartingale for all $\lambda \in \mathbb{R}$. Our unified proof technique leads us immediately to the following result.

Corollary 18 (Anytime bound finite second moment) *Let $Z_1, Z_2, \dots$ be a stream of (not necessarily i.i.d.) data and (f_t) be a sequence of predictable loss functions such that $\mathbb{E}_{\mathcal{D}}[f_t^2(Z,\theta)|\mathcal{F}_{t-1}] < \infty$. Let (λ_t) be a nonnegative predictable sequence and consider any data-free prior $\nu \in \mathcal{M}(\Theta)$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all times t and measures $\rho \in \mathcal{M}(\Theta)$,*

$$\sum_{i \leqslant t} \lambda_i \mathbb{E}_\rho \Delta_i(\theta) \leqslant \frac{1}{6} \sum_{i \leqslant t} \lambda_i^2 \mathbb{E}_\rho \Big(\Delta_i^2(\theta) + 2\mathbb{E}_{\mathcal{D}}[\Delta_i^2(\theta)|\mathcal{F}_{i-1}] \Big) + \log(1/\delta) + D_{\mathrm{KL}}(\rho\|\nu). \tag{24}$$

The proof of this result (including that $L_t(\theta)$ above forms a supermartingale) can be found in Appendix A.9. The right hand side of (24) can be upper bounded to give a more interpretable result. In particular, if we consider stationary losses and i.i.d. data, then we can replace (24) with the following:

$$\mathbb{E}_\rho R(\theta) \leqslant \mathbb{E}_\rho \widehat{R}_t(\theta) + \frac{\lambda}{6t} \sum_{i \leqslant t} f^2(Z_i,\theta) + \frac{\lambda}{3} \mathbb{E}_{\rho,\mathcal{D}}[f^2(Z,\theta)] + \frac{\log(1/\delta) + D_{\mathrm{KL}}(\rho\|\nu)}{\lambda t}. \tag{25}$$

This recovers (with slightly tighter constants), Theorem 2.3 of Haddouche and Guedj (2023). The relationship between (25) and Corollary 17 is worth investigating. For i.i.d. data and fixed $\lambda_i = \lambda > 0$, (23) can be rearranged to read

$$\mathbb{E}_\rho R(\theta) \leqslant \mathbb{E}_\rho \widehat{R}_t(\theta) + \frac{\lambda}{2} \mathbb{E}_{\rho,\mathcal{D}}[f^2(Z,\theta)] + \frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta)}{\lambda t}. \tag{26}$$

Subtracting the right hand side of (25) from (26) gives

$$D := \frac{\lambda}{6t} \sum_{i \leqslant t} f^2(Z_i,\theta) - \frac{\lambda}{6} \mathbb{E}_{\rho,\mathcal{D}}[f^2(Z,\theta)],$$

which converges to zero almost surely via the LLN. Because (25) is looser than (24), this implies that Corollary 17 is looser than Corollary 18. Corollary 17 is, however, a cleaner result, and one we felt was worth stating. Let also note that using (25), Haddouche and Guedj (2023) are able to generalize previous work of Haddouche et al. (2021) on unbounded losses under the Hypothesis Dependent Range Condition (HYPE). The same discussion and generalization thus applies here.

We end this section with an open problem: Can we obtain a time-uniform PAC-Bayes bound for losses under the sole assumption of a bounded p-th moment, $1 < p \leqslant 2$? Wang and Ramdas (2023a), based on previous work of Chen et al. (2021) and Catoni (2012), have provided nonnegative supermartingales under such conditions. They do not, however, result in closed form expressions of the risk, making the resulting PAC-Bayes bound difficult to use.

5. PAC-Bayes Bounds via Submartingales

While Section 4 was able to generalize several fixed-time PAC-Bayes bounds, the sub-ψ approach explored therein does not cover all existing PAC-Bayes bounds. Here we explore the other half of Theorem 4, giving bounds based on reverse-time submartingales.

Throughout this section, for reasons that will become clear later, we will require that the loss is stationary ($f_t = f$) and that the data (Z_t) are *exchangeable*. In particular, for all $t \geqslant 1$, and permutations $g : [t] \to [t]$, $(Z_1, \dots, Z_t) \overset{d}{=} (Z_{g(1)}, \dots, Z_{g(t)})$. Exchangeability is slightly weaker than the i.i.d. assumption. For instance, sampling without replacement gives rise to exchangeable sequences which are not i.i.d. Another example comes from considering $X_1 + Y, \dots, X_n + Y$ for some random variable Y and i.i.d. $X_1, \dots, X_n$. Observe that exchangeability implies a common mean, so throughout this section we set $R(\theta) = R_t(\theta) = \mathbb{E}_{\mathcal{D}}[f(Z, \theta)]$ for all t.

The bounds in the previous section were based on the process $S_t = \sum_{i=1}^t (\mu_i(\theta) - f_i(Z_i, \theta))$, while those in this section will be based on the process (S_t/t). This is because, while the partial sums (S_t) form a martingale, only the *partial means* (S_t/t) form a reverse submartingale. We'll see that while PAC-bounds based on reverse submartingales can capture a larger variety of relationships between $R_t(\theta)$ and $\widehat{R}_t(\theta)$, this comes at the expense of slightly looser bounds in addition to stronger distributional assumptions.

A formidable example of a bound which is not recovered by appealing to supermartingales is that of Germain et al. (2015) (a similar bound was stated by Lever et al. (2010); Theorem 1). This generalizes a class of bounds which consider convex functions acting on the risk and empirical risk. In particular, this recovers earlier bounds of Seeger (2002, 2003); Germain et al. (2009); McAllester (1998, 2003). A similar bound was given recently by Rivasplata et al. (2020) when considering PAC-Bayes bounds for stochastic kernels.

Proposition 19 (Germain et al., 2015) *Let $Z_1, \dots, Z_n$ be i.i.d., $\varphi : [0,1]^2 \to \mathbb{R}$ be convex and $f = f_t$ be stationary and bounded in $[0,1]$. Let ν be a data-free prior. For all n and $\lambda > 0$, with probability at least $1 - \delta$ over the random draw of (Z_t), for all $\rho \in \mathcal{M}(\Theta)$,*

$$\varphi(\mathbb{E}_\rho \widehat{R}_n(\theta), \mathbb{E}_\rho R(\theta)) \leqslant \frac{1}{\lambda} \log \mathbb{E}_\nu \mathbb{E}_{\mathcal{D}} \exp(\lambda \varphi(\widehat{R}_n(\theta), R(\theta)) + \frac{D_{\mathrm{KL}}(\rho \| \nu) + \log(1/\delta)}{\lambda}.$$

Let us consider for a moment attempting to give an anytime-valid version of the above result using the machinery from Section 4. One would need to guarantee that the nonnegative process $P_t(\theta) = \exp\left\{\lambda \varphi(\mathbb{E}_\rho \widehat{R}_n(\theta), \mathbb{E}_\rho R(\theta)) - \log \mathbb{E}_\nu \mathbb{E}_{\mathcal{D}} \exp(\lambda \varphi(\widehat{R}_t(\theta), R(\theta))\right\}$ is upper bounded by a supermartingale. Since φ may not be linear, however, one cannot write this as a product of exponential terms, thereby making it difficult to write $\mathbb{E}[P_t(\theta) | \mathcal{F}_{t-1}]$ in terms of $P_{t-1}(\theta)$. We thus require a different approach. Interestingly, one can show that convex functions acting on the empirical risk are reverse submartingales with respect to an appropriate filtration, which we define below. From here, Ville's inequality for reverse submartingales (Lemma 2) will provide us with an anytime version of Proposition 19.

Given a sequence of data $Z_1, Z_2, \dots$, the *exchangeable reverse filtration* $(\mathcal{E}_t)_{t=1}^\infty$ is the reverse filtration where $\mathcal{E}_t$ is the σ-algebra generated by all (Borel) measurable functions of the data which are permutation symmetric in their first t arguments. We say a function s is permutation symmetric if $s(Z_1, \dots, Z_t) = s(Z_{g(1)}, \dots, Z_{g(t)})$ for all permutations $g : [t] \to$

$[t]$. Formally, $\mathcal{E}_t$ is written

$$\mathcal{E}_t = \sigma\Big(\big\{s(Z_1,\dots,Z_t) : s \text{ is permutation symmetric } \big\} \cup \{Z_j\}_{j>t}\Big). \tag{27}$$

We find the following intuition from Manole and Ramdas (2023) helpful when thinking about $\mathcal{E}_t$. $\mathcal{E}_1$ might be viewed as an omniscient oracle with access to all information over the whole future. As time goes on, her memory of the past decays but she retains perfect knowledge of the future. Importantly, she does not forget what happened in the past, only the *order* in which events occurred. That is, the oracle $\mathcal{E}_t$ is omniscient with respect to $Z_{t+1}, Z_{t+2}, \dots$, but forgets the order of $Z_1, \dots, Z_t$. Manole and Ramdas (2023) also give a sufficient condition for a process to be a reverse submartingale with respect to $(\mathcal{E}_t)$.

Lemma 20 (Leave-one-out, Manole and Ramdas, 2023, Corollary 5) *If a sequence of permutation invariant functions $\{h_t : \mathcal{Z}^t \to \mathbb{R}\}$ satisfies the "leave-one-out" property, namely, $h_t(Z^t) \leqslant \frac{1}{t}\sum_{i=1}^t h_{t-1}(Z^t_{-i})$ (where Z^t_{-i} omits Z_i), then $(h_t(Z^t))_{t=0}^\infty$ is a reverse submartingale with respect to $(\mathcal{E}_t)$. Moreover, if the expression above holds with equality then $(h_t(Z^t))$ is a reverse martingale with respect to $(\mathcal{E}_t)$.*

To reduce notational clutter, given a convex function $\varphi : \mathbb{R}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$, define

$$\varphi_t(\theta) := \varphi(\widehat{R}_t(\theta), R(\theta)). \tag{28}$$

That is, φ_t simply fixes the second argument of φ as $R(\theta)$ and sets the first as the empirical risk at time t. Considering φ_t is useful because the stochastic process $(\varphi_t(\theta))_{t=1}^\infty$ for fixed θ is a reverse submartingale with respect to $(\mathcal{E}_t)$. This holds by Lemma 20, since the empirical risk $\widehat{R}_t(\theta)$ is permutation invariant and the convexity of φ ensures that the leave-one-out property holds, as proven below.

Lemma 21 *For an exchangeable sequence (Z_t), $(\varphi_t(\theta))$ is a reverse submartingale with respect to $(\mathcal{E}_t)$.*

Proof First note that $\varphi_t(\theta)$ is permutation invariant by construction. Thus, by Lemma 20, we need only show that it satisfies the leave-one-out property. For each $i \in [t]$, define

$$\widehat{R}_t^{(-i)}(\theta) := \frac{1}{t-1}\sum_{j\neq i} f(Z_j, \theta),$$

and observe that

$$\sum_{i=1}^t \widehat{R}_t^{(-i)}(\theta) = \frac{1}{t-1}\sum_{i=1}^t\sum_{j\neq i} f(Z_j,\theta) = \sum_{i=1}^t f(Z_j,\theta) = t\widehat{R}_t(\theta).$$

Consequently, by the convexity of φ and Jensen's inequality,

$$\begin{aligned}\varphi_t(\theta) &= \varphi(\widehat{R}_t(\theta), R(\theta)) = \varphi\bigg(\frac{1}{t}\sum_{i=1}^t \widehat{R}_t^{(-i)}(\theta), R(\theta)\bigg)\\ &\leqslant \frac{1}{t}\sum_{i=1}^t \varphi(\widehat{R}_t^{(-i)}(\theta), R(\theta)) = \frac{1}{t}\sum_{i=1}^t \varphi_0((Z_1,\dots,Z_{i-1},Z_{i+1},\dots,Z_t),\theta),\end{aligned}$$

which is precisely the leave-one-out property. ■

Our reliance on Lemma 20 is the reason that this section considers only stationary loss functions (but so do the bounds we generalize). More specifically, stationary losses are required for $\varphi_t(\theta)$ to be permutation invariant. We cannot in general swap Z_i and Z_k if f_i and f_k are different.

5.1 A Time-Uniform Bound for Convex Functions

As we alluded to in the introduction, while the supermartingale approach of Section 4 was able to generalize fixed-time bounds at no cost, this is not true for the bounds presented in this section. Roughly speaking, this is because even though the process $(\varphi_t(\theta))_{t\geqslant 1}$ is a reverse submartingale with respect to $(\mathcal{E}_t)$ (and therefore so is $(\exp(\lambda\varphi_t(\theta)))_{t\geqslant 1}$), the process $(\exp\{\lambda\varphi_t(\theta) - \log \mathbb{E}_{\nu,\mathcal{D}} \exp(\lambda\varphi_t(\theta))\})_{t\geqslant 1}$ may not be. Thus, we cannot use such a process in Theorem 4 to recover (a time-uniform version of) Proposition 19 exactly.

Instead, we rely on a "stitching" argument in a similar vein to Howard et al. (2021) and Manole and Ramdas (2023). This entails considering a series of submartingales over geometrically spaced epochs $[2^{t-1}, 2^t)$, $t \geqslant 0$, each holding with a precise probability such that we may take the union bound over all such intervals to obtain our result. As we'll see, the resulting bounds will suffer at most a small constant factor plus an iterated logarithm factor over the originals.

Formally, we consider a "stitching function" function $\ell : \mathbb{N}_{>0} \to (1, \infty)$ such that $\sum_{k=1}^{\infty} \frac{1}{\ell(k)} \leqslant 1$. Different choices will leads to different shapes of the resulting bounds. For clarity and concreteness, however, we will consider the following stitching function for the remainder of this manuscript:

$$\ell(k) = k^2\zeta(2), \quad \text{where} \quad \zeta(2) = \sum_{j=1}^{\infty} j^{-2} \approx 1.645.$$

We also introduce the following "iterated logarithm" factor that captures the small excess error inherent to our anytime-valid bounds:

$$\mathsf{IL}_t := \log(\ell(\log_2(2t))) < 2\log\log 2t + 1.3. \tag{29}$$

Additionally, throughout this section we set

$$\bar{t} := 2^{\lfloor \log_2(t) \rfloor}. \tag{30}$$

With these definitions in hand, we now state our time-uniform version of Proposition 19.

Corollary 22 (General anytime bound for convex functions) *Let (Z_t) be exchangeable. Let $\varphi : \mathbb{R}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ be convex and $\nu \in \mathcal{M}(\Theta)$ be a prior. Let (λ_t) be a sequence of positive values. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all $\rho \in \mathcal{M}(\Theta)$ and at all times $t \geqslant 1$,*

$$\mathbb{E}_\rho \varphi_t(\theta) \leqslant \frac{\log \mathbb{E}_\nu \mathbb{E}_{\mathcal{D}} \exp\left\{\lambda_{\bar{t}} \varphi_{\bar{t}}(\theta)\right\}}{\lambda_{\bar{t}}} + \frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta) + \mathsf{IL}_t}{\lambda_{\bar{t}}}, \tag{31}$$

for IL_t as in (29), $\bar{t}$ as in (30), and $\varphi_t(\theta) = \varphi(\widehat{R}_t(\theta), R(\theta))$.

Ideally, the subscripts $\bar{t}$ above would have been equal to t, but we were not able to prove such a result. Since $t/2 \leqslant \bar{t} \leqslant t$, this results in a slight looseness; see Remark 23.

Proof Recall the shorthand $\varphi_t(\theta) = \varphi(\widehat{R}_t(\theta), R(\theta))$. For $j \in \mathbb{N}$, define

$$M_t^j(\theta) = \lambda_j \varphi_t(\theta) - \log \mathbb{E}_{\nu,\mathcal{D}} \exp(\lambda_j \varphi_j(\theta)). \tag{32}$$

the second term on the right hand side is deterministic, so Lemma 21 implies that $(M_t^j(\theta))_{t \geqslant 1}$ is a reverse submartingale with respect to $(\mathcal{E}_t)$. Hence, by Jensen's inequality, so is the process $(\exp M_t^j(\theta))_{t \geqslant 1}$. Moreover, note that $\mathbb{E}_{\mathcal{D}} \exp(M_j^j(\theta)) = 1$. Therefore, Theorem 4 implies that, for all ρ,

$$\mathbb{P}(\exists t \geqslant j : \mathbb{E}_\rho M_t^j(\theta) - D_{\mathrm{KL}}(\rho\|\nu) \geqslant \log(u/\delta)) \leqslant \delta/u, \tag{33}$$

for $u > 0$. Suppose that for some $t^* \geqslant 1$ and some $\rho \in \mathcal{M}(\theta)$ we have the inequality $\mathbb{E}_\rho M_{t^*}^{\bar{t^*}}(\theta) - D_{\mathrm{KL}}(\rho\|\nu) \geqslant \log(\ell(\log_2(2t^*))/\delta)$. By construction, $\bar{t^*} = 2^{k^*}$ where $k^* = \lfloor \log_2(t^*) \rfloor \in \mathbb{N}$. Therefore,

$$\begin{aligned}\mathbb{E}_\rho M_{t^*}^{2^{k^*}}(\theta) - D_{\mathrm{KL}}(\rho\|\nu) &= \mathbb{E}_\rho M_{t^*}^{\bar{t^*}}(\theta) - D_{\mathrm{KL}}(\rho\|\nu) \\ &\geqslant \log(\ell(\log_2(2t^*))/\delta) \geqslant \log(\ell(k^*+1)/\delta),\end{aligned}$$

where the final inequality follows since $\log_2(2t^*) = \log_2(t^*) + 1 \geqslant \lfloor \log_2(t^*) \rfloor + 1 = k^* + 1$, and ℓ is an increasing function. We have thus shown that the event

$$\{\exists \rho, \exists t \geqslant 1 : \mathbb{E}_\rho M_t^{\bar{t}}(\theta) - D_{\mathrm{KL}}(\rho\|\nu) \geqslant \log(\ell(\log_2(2t))/\delta)\},$$

is contained in

$$\bigcup_{k=0}^{\infty} \left\{\exists \rho, \exists t \geqslant 2^k : \mathbb{E}_\rho M_t^{2^k}(\theta) - D_{\mathrm{KL}}(\rho\|\nu) \geqslant \log(\ell(k+1)/\delta))\right\},$$

implying that

$$\begin{aligned}&\mathbb{P}\left(\exists \rho, \exists t \geqslant 1 : \mathbb{E}_\rho M_t^{\bar{t}}(\theta) - D_{\mathrm{KL}}(\rho\|\nu) \geqslant \log(\ell(\log_2(2t))/\delta)\right) \\ &\leqslant \mathbb{P}\left(\bigcup_{k=0}^{\infty} \left\{\exists \rho, \exists t \geqslant 2^k : \mathbb{E}_\rho M_t^{2^k}(\theta) - D_{\mathrm{KL}}(\rho\|\nu) \geqslant \log(\ell(k+1)/\delta))\right\}\right) \\ &\leqslant \sum_{k=0}^{\infty} \mathbb{P}(\exists \rho, \exists t \geqslant 2^k : \mathbb{E}_\rho M_t^{2^k}(\theta) - D_{\mathrm{KL}}(\rho\|\nu) \geqslant \log(\ell(k+1)/\delta)) \leqslant \sum_{k=0}^{\infty} \frac{\delta}{\ell(k+1)} = \delta,\end{aligned}$$

where we've applied (33) with $u = \ell(k+1)$. In other words, expanding $M_t^{\bar{t}}(\theta)$, we have that with probability at least $1-\delta$, for all ρ and $t \geqslant 1$,

$$\begin{aligned}\lambda_{\bar{t}} \mathbb{E}_\rho \varphi_t(\theta) &\leqslant \mathbb{E}_\rho \log \mathbb{E}_{\nu,\mathcal{D}} \exp(\lambda_{\bar{t}} \varphi_{\bar{t}}(\theta)) + D_{\mathrm{KL}}(\rho\|\nu) + \log(\ell(\log_2(2t))/\delta) \\ &\leqslant \log \mathbb{E}_{\nu,\mathcal{D}} \exp(\lambda_{\bar{t}} \varphi_{\bar{t}}(\theta)) + D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta) + \mathsf{IL}_t,\end{aligned}$$

using the definition of ℓ and the fact that log is concave. Dividing by $\lambda_{\bar{t}}$ completes the argument. ■

Remark 23 *Notice that our choice of $\bar{t}$ may leave us with a bound that is approximately twice as big as the fixed time counterpart. Indeed, if $t = 1023$ then $\bar{t} = 512 \approx t/2$. Thus if $\lambda_j = j$ (say), then $\lambda_{\bar{1023}} = 512 \approx \lambda_{1023}/2$. The culprit is our choice of $\bar{t}$ and the constant 2 therein. We chose 2 simply as a matter of convenience, but the analysis can be modified. In particular, for any fixed $s > 1$, we may consider $\bar{t} = s^{\lfloor \log_s(t) \rfloor}$, which obeys $t/s \leqslant \bar{t} \leqslant t$. As $s \to 1$, $\bar{t}$ thus lags behind t less and less. The price we pay is that the iterated logarithm error term must be modified to $\mathsf{IL}_t = \log(\ell(\log_s(st)))$ which grows as $s \to 1$. We leave the choice of s and the appropriate trade-off between the lag and the additive error to the practitioner.*

Comparing Corollary 22 and Proposition 19, we see there are several differences aside from IL_t. For one, our expectation is on the outside of φ_t on the left hand side. Of course, because φ is convex, $\mathbb{E}_\rho \varphi(\widehat{R}_t(\theta), R(\theta)) \geqslant \varphi(\mathbb{E}_\rho \widehat{R}_t(\theta), \mathbb{E}_\rho R(\theta))$, so our result implies a bound on the latter term. Second, as noted in the remark above, our log-MGF term is based on $\bar{t} \in [t/2, t]$ instead of t, thus "lags behind" the fixed-time result. This is a consequence of stitching. However, if there is a fixed time n of special interest, we can obtain the following time-uniform bound for all $t \geqslant n$, which is just as tight as Proposition 19.

Corollary 24 *Let (Z_t) be exchangeable. Let $\varphi : \mathbb{R}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ be convex and $\nu \in \mathcal{M}(\Theta)$ be a prior. Fix $\lambda > 0$ and $n \in \mathbb{N}$. Then, for all $\delta \in (0,1)$, with probability at least $1 - \delta$ over the random draw of (Z_t), for all $\rho \in \mathcal{M}(\Theta)$ and at all times $t \geqslant n$,*

$$\mathbb{E}_\rho \varphi(\widehat{R}_t(\theta), R(\theta)) \leqslant \frac{\log \mathbb{E}_\nu \mathbb{E}_{\mathcal{D}} \exp\left\{\lambda \varphi(\widehat{R}_n(\theta), R(\theta))\right\}}{\lambda} + \frac{D_{\mathrm{KL}}(\rho \| \nu) + \log(1/\delta)}{\lambda}. \tag{34}$$

Proof Similarly to the proof of Corollary 22, set

$$M_t^n(\theta) = \exp\left\{\lambda \varphi_t(\theta) - \log \mathbb{E}_{\nu, \mathcal{D}} \exp(\lambda \varphi_n(\theta))\right\}.$$

Then $(M_t^n(\theta))_{t \geqslant n}$ is a reverse submartingale with respect to $(\mathcal{E}_t)_{t \geqslant n}$ with $\mathbb{E}_{\mathcal{D}} M_n^n(\theta) = 1$. Therefore, Theorem 4 gives

$$\mathbb{P}(\exists \rho, \exists t \geqslant n : \mathbb{E}_\rho M_t^n(\theta) - D_{\mathrm{KL}}(\rho \| \nu) \geqslant \log(1/\delta)) \leqslant \delta,$$

which rearranges to the claimed result. ∎

Corollary 24 requires some interpretation. The right hand side of (34) is constant with respect to t. While such a bound might be more straightforward for a fixed $t \geqslant n$, our bound shows that it holds simultaneously for all $t \geqslant n$. These bounds are in some sense analogous to Freedman-style deviation inequalities (which hold for all $t \leqslant n$, but with tightness only depending on n and not improving for $t \ll n$) and perhaps even more analogous to de la Peña-style deviation inequalities (which hold for all $t \geqslant n$, but with tightness only depending on time n and not improving for $t \gg n$) — see Howard et al. (2020) for a detailed discussion and a unification of the two types of boundaries (in particular Figures 1, 4 and 5 for intuition).

5.2 A Time-Uniform Seeger Bound

By choosing particular convex functions φ and applying Corollary 22 (or 24), we recover time-uniform versions of several classical PAC-Bayes inequalities. We present several of them here, but refer the reader to resources such as Alquier (2021) and Germain et al. (2009) for more comprehensive discussions. A particularly famous result is that of Langford and Seeger (2001) and Maurer (2004). To state it, let us define, for any $p, q \in (0,1)$,

$$\mathsf{kl}(p\|q) := p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right),$$

which is the KL-divergence between Bernoulli distributions with means p and q, respectively. That kl is convex (in each argument, p and q) thus follows from the fact that $D_{\mathrm{KL}}(\cdot\|\cdot)$ is convex (in distribution space). Indeed,

$$\begin{aligned}\mathsf{kl}(\lambda p_1 + (1-\lambda)p_2\|q) &= D_{\mathrm{KL}}(\lambda \mathrm{Ber}(p_1) + (1-\lambda)\mathrm{Ber}(p_2)\|\mathrm{Ber}(q)) \\ &\leqslant \lambda D_{\mathrm{KL}}(\mathrm{Ber}(p_1)\|\mathrm{Ber}(q)) + (1-\lambda) D_{\mathrm{KL}}(\mathrm{Ber}(p_2)\|\mathrm{Ber}(q)) \\ &= \lambda\, \mathsf{kl}(p_1\|q) + (1-\lambda)\,\mathsf{kl}(p_2\|q),\end{aligned}$$

for any $\lambda \in [0,1]$, where $\mathrm{Ber}(p)$ is a Bernoulli distribution with mean p. An identical argument holds for the second argument of kl. Now, for $k \in \mathbb{N}$, define

$$\xi(k) := \sum_{\ell=0}^{k} \mathbb{P}_{Y\sim\mathrm{Bin}(k,\ell/k)}(Y=\ell) = \sum_{\ell=0}^{k} \binom{k}{\ell} (\ell/k)^{\ell}(1-\ell/k)^{k-\ell}.$$

As noted by Maurer (2004); Germain et al. (2015), $\sqrt{k} \leqslant \xi(k) \leqslant 2\sqrt{k}$ for all $k \in \mathbb{N}$. Employing Corollary 22 leads to the following bound, which relates $\xi(k)$ to the log-MGF. Recall that $\bar{t} = 2^{\lfloor \log_2(t)\rfloor}$ and $\mathsf{IL}_t < 2\log\log 2t + 1.3$. The proof of the following bound can be found in Appendix A.10.

Corollary 25 (Anytime-valid Langford-Seeger Bound) *Let (Z_t) be i.i.d. and consider stationary losses bounded in $[0,1]$. Let $\nu \in \mathcal{M}(\Theta)$ be a data-free prior. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all $\rho \in \mathcal{M}(\Theta)$ and at all times $t \geqslant 1$,*

$$\mathbb{E}_\rho\, \mathsf{kl}(\widehat{R}_t(\theta)\|R(\theta)) \leqslant \frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(\xi(\bar{t})/\delta) + \mathsf{IL}_t}{\bar{t}}. \tag{35}$$

Moreover, for any fixed n, we obtain that for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all $\rho \in \mathcal{M}(\Theta)$ and at all times $t \geqslant n$,

$$\mathbb{E}_\rho\, \mathsf{kl}(\widehat{R}_t(\theta)\|R(\theta)) \leqslant \frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(\xi(n)/\delta)}{n}. \tag{36}$$

For ease of comparison, let us recall the usual fixed-time version of this bound, which reads: For all $n \in \mathbb{N}$, with probability at least $1-\delta$, for all ρ,

$$\mathsf{kl}(\mathbb{E}_\rho \widehat{R}_n(\theta)\|\mathbb{E}_\rho R(\theta)) \leqslant \frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(\xi(n)/\delta)}{n}. \tag{37}$$

Noting that $\mathsf{kl}(\mathbb{E}_\rho \widehat{R}_n(\theta) \| \mathbb{E}_\rho R(\theta)) \leqslant \mathbb{E}_\rho \mathsf{kl}(\widehat{R}_n(\theta) \| R(\theta))$ due to Jensen's inequality, we see that at time $t = n$, (36) recovers the fixed-time Seeger bound. Moreover, by noting that $\bar{t} \in [t/2, t]$, (35) provides a guarantee for all $t \geqslant 1$, that is at most a constant factor worse than (36). We emphasize that this constant factor can be changed by altering the definition of $\bar{t}$; see Remark 23. As a brief historical note, (37) was first stated in that form by Germain et al. (2015, Lemma 20). Langford and Seeger (2001, Theorem 3) use n in place of $\xi(n)$ and Maurer (2004, Theorem 5) then tightens this to $2\sqrt{n}$.

A time-uniform McAllester bound (McAllester, 1998, 2003) — distinct from that derived in Section 4 — follows immediately by applying Jensen's inequality and Pinsker's inequality: For all $x, y \in (0,1)$, $2(x-y)^2 \leqslant \mathsf{kl}(x\|y)$. This implies that $2[\mathbb{E}_\rho(\widehat{R}_t(\theta) - R(\theta))]^2 \leqslant 2\mathbb{E}_\rho(\widehat{R}_t(\theta) - R(\theta))^2 \leqslant \mathbb{E}_\rho \mathsf{kl}(\widehat{R}_t(\theta)\|R(\theta))$. Using this in conjunction with the fact that $\xi(k) \leqslant 2\sqrt{k}$ yields the following.

Corollary 26 (Anytime-valid McAllester Bound) *Let (Z_t) be i.i.d. and consider stationary losses bounded in $[0,1]$. Let $\nu \in \mathcal{M}(\Theta)$ be a data-free prior. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all $\rho \in \mathcal{M}(\Theta)$ and at all times $t \geqslant 1$, we have*

$$\mathbb{E}_\rho R(\theta) \leqslant \mathbb{E}_\rho \widehat{R}_t(\theta) + \left(\frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(2\sqrt{\bar{t}}/\delta) + \mathsf{IL}_t}{2\bar{t}} \right)^{1/2}.$$

Moreover, for any fixed n, we obtain that for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all $\rho \in \mathcal{M}(\Theta)$ and at all times $t \geqslant n$,

$$\mathbb{E}_\rho R(\theta) \leqslant \mathbb{E}_\rho \widehat{R}_t(\theta) + \left(\frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(2n/\delta)}{2n} \right)^{1/2}. \tag{38}$$

As above, at the fixed time $t = n$, (38) recovers McAllester's bound in (12). Other bounds follow from other choices of φ. Bégin et al. (2016) note that $\varphi(x,y) = -cx - \log(1 - y(1-e^{-c})$ leads to Theorem 1.2.6 of Catoni (2007). Meanwhile, as pointed out by Alquier (2021) and Pérez-Ortiz et al. (2021) we can also generate the bounds of Tolstikhin and Seldin (2013) and Thiemann et al. (2017) by using other inequalities involving kl.

5.3 U- and V-statistics

We end this section by discussing the representation of U- and V-statistics as reverse submartingales, thus opening the door for bounds based on these quantities. We will assume the data (Z_t) are drawn i.i.d. and that the loss function is stationary. We consider functionals of the form $\Phi : \mathcal{P}(\mathcal{Z}) \times \Theta \to \mathbb{R}$, where

$$\Phi(P, \theta) = \iint h(f(z_1, \theta), f(z_2, \theta)) \mathrm{d}P(z_1) \mathrm{d}P(z_2).$$

Here, P is a distribution over the data, and $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{>0}$ is symmetric, continuous, and positive semi-definite. For instance, if we take $h(a,b) = (a-b)^2/2$, then $\Phi(P,\theta) =$

$\mathsf{Var}_{Z\sim P}[f(Z,\theta)]$. The U- and V-statistics of Φ are, respectively,

$$U_t(\theta) := \frac{2}{t(t-1)} \sum_{1\leqslant i<j\leqslant t} h(f(Z_i,\theta), f(Z_j,\theta)), \tag{39}$$

$$V_t(\theta) := \frac{1}{t^2} \sum_{1\leqslant i,j\leqslant t} h(f(Z_i,\theta), f(Z_j,\theta)). \tag{40}$$

Berk (1966) and, more recently, Manole and Ramdas (2023) observe that $U_t(\theta)$ is a reverse martingale with respect to $(\mathcal{E}_t)$. Indeed, this can be seen by appealing to the leave-one-property (Lemma 20). As for $V_t(\theta)$, it can be seen to be a reverse submartingale with respect to $(\mathcal{E}_t)$ if, in addition to the conditions on h above, the range of f is a compact subset of $\mathbb{R}$ (Manole and Ramdas, 2023, Proposition 16).

In conjunction with Theorem 4, these properties enable us to give time-uniform PAC-Bayes bounds involving functionals and their U- and V-statistics. To illustrate, we recover time-uniform versions of Theorems 3 and 4 in Tolstikhin and Seldin (2013). Recalling Remark 23, this result will employ stitching and thus lags behind the fixed-time result by a small constant. As was discussed in Section 4.2.4, a tighter time-uniform bound (i.e., one that uses forward supermartingales and thus does not have an iterated logarithm factor) can be obtained by using a betting-style martingale. However, we state and prove the following bound as an example of how U-statistics and their properties can be applied in the PAC-Bayes setting.

Corollary 27 *Let (Z_t) be drawn i.i.d. and f be stationary. Let $\mathsf{Var}_t(\theta)$ be the unbiased empirical variance, i.e., $\mathsf{Var}_t(\theta) := \frac{1}{t(t-1)}\sum_{1\leqslant i<j\leqslant t}(f(Z_i,\theta)-f(Z_j,\theta))^2$. Let (λ_t) be a sequence of positive scalars. Denote the true variance as $\mathsf{Var}(\theta) = \mathsf{Var}_{Z\sim P}[f(Z,\theta)]$. Then, for any $\delta\in(0,1)$, with probability at least $1-\delta$ over (Z_t), for all $t\geqslant 1$ and $\rho\in\mathcal{M}(\Theta)$,*

$$\mathbb{E}_\rho[\mathsf{Var}(\theta)-\mathsf{Var}_t(\theta)] \leqslant \frac{D_{\mathrm{KL}}(\rho\|\nu)+\log(1/\delta)+\mathsf{IL}_t+\frac{\lambda_{\bar{t}}^2}{2}\frac{\bar{t}^2}{\bar{t}-1}\mathbb{E}_\rho\mathsf{Var}(\theta)}{\bar{t}\lambda_{\bar{t}}}, \tag{41}$$

where $\bar{t}=2^{\lfloor\log_2(t)\rfloor}$ and $\mathsf{IL}_t<2\log\log 2t+1.3$.

The proof may be found in Appendix A.11. Theorem 3 of Tolstikhin and Seldin (2013) follows from optimizing over λ with a union bound (for a fixed $t=n$). In particular, they consider a grid of λs designed as a geometric progression. Then, for each ρ, they find the optimal value of λ from the grid. Finally, they apply a union bound over the grid, a technique inspired by Seldin et al. (2012) (and which we will see again in Section 6). Doing so yields a bound with dependence $\widetilde{O}(\sqrt{\frac{\mathsf{Var}_t(\theta)(D_{\mathrm{KL}}(\rho\|\nu)+\log(1/\delta))}{n}}+\frac{D_{\mathrm{KL}}(\rho\|\nu)+\log(1/\delta)}{n})$, where $\widetilde{O}$ ignores iterated logarithm factors. If $\mathsf{Var}_t(\theta)$ is sufficiently small, the bound then scales at a rate of roughly $1/n$. In the anytime setting meanwhile, we might consider taking $\lambda_j=\frac{1}{\sqrt{j}}$, in which case the right hand side of (41) becomes proportional to $\frac{D_{\mathrm{KL}}(\rho\|\nu)+\log(1/\delta)+\mathbb{E}_\rho\mathsf{Var}(\theta)}{\sqrt{\bar{t}}}$, thus shrinking at a rate of roughly $1/\sqrt{\bar{t}}$. Similary to the fixed-time setting, Theorem 4 of Tolstikhin and Seldin (2013) follows from combining the above result with Corollary 38, a time-uniform extension of results in Seldin et al. (2012).

6. Extensions

Owing in part to the ability for PAC-Bayes bounds to provide insight into the performance of neural networks (Dziugaite and Roy, 2017; Biggs and Guedj, 2022), recent years have seen a surge of interest in and progress on the topic. In this section, we provide some comments on the ability of our unified framework to incorporate some of these advances. In particular, we discuss replacing the KL divergence with integral probability metrics, ϕ-divergences, and Rényi divergences, in addition to how Theorem 4 enables us to replace the loss function with martingale difference sequences. We also discuss how many of the bounds in the two previous sections give rise to confidence sequences (i.e., time-uniform confidence intervals), and provide some general advice on choosing (λ_t) in the supermartingale bounds.

6.1 Replacing the KL Divergence with IPMs

Given that all the bounds provided thus far rely on the KL divergence between ρ and ν, a natural question is whether we can replace this term with an alternative distributional metric? Here we answer in the affirmative and demonstrate that recent work by Amit et al. (2022), which replaces the KL divergence with a variety of *Integral Probability Metrics* (IPMs), can be made time-uniform.

Definition 28 *Let $\mathcal{G}$ be a family of functions which map Θ to $\mathbb{R}$. The Integral Probability Metric with respect to $\mathcal{G}$ between two distributions ρ and ν over Θ is*

$$\gamma_{\mathcal{G}}(\rho, \nu) := \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{\theta \sim \rho} g(\theta) - \mathbb{E}_{\theta \sim \nu} g(\theta) \right|. \tag{42}$$

IPMs are a large class of divergences. By choosing the appropriate family $\mathcal{G}$, one can recover the Total Variation distance, the Wasserstein distance, the Dudley metric, and the Maximum Mean Discrepancy (Sriperumbudur et al., 2009). We note that the KL divergence is not a special case of an IPM.

The following theorem is our main result for IPMs. Just as Theorem 4 provided a general framework for generating PAC-Bayes bounds with a KL-divergence term, Theorem 29 provides a framework for generating PAC-Bayes bounds with an IPM. The main idea is to replace the use of the Donsker-Varadhan formula with an assumption on the family of functions $\mathcal{G} : \Theta \to \mathbb{R}$ (or, more precisely, *families* of functions).

Theorem 29 *Let $(\mathcal{G}_t)_{t \geqslant 1}$ be a predictable sequence, where each $\mathcal{G}_t$ is a family of functions from $\Theta \to \mathbb{R}$. Let (h_t) be a sequence of functions such that $h_t \in \mathcal{G}_t$ for all $t \geqslant t_0$. Suppose that $(\exp h_t(\theta))_{t \geqslant t_0}$ is a supermartingale or reverse submartingale (adapted to some filtration) for all $\theta \in \Theta$ such that $\mathbb{E}_{\mathcal{D}} \exp h_{t_0}(\theta) \leqslant 1$. Then, for any $\delta \in (0, 1)$ and prior $\nu \in \mathcal{M}(\Theta)$, with probability at least $1 - \delta$,*

$$\mathbb{E}_{\theta \sim \rho} h_t(\theta) \leqslant \gamma_{\mathcal{G}_t}(\rho, \nu) + \log(1/\delta), \tag{43}$$

for all $\rho \in \mathcal{M}(\Theta)$ and times $t \geqslant t_0$.

Proof By assumption, $h_t \in \mathcal{G}_t$ for all t. Hence $\gamma_{\mathcal{G}_t}(\rho, \nu) \geqslant \mathbb{E}_\rho h_t(\theta) - \mathbb{E}_\nu h_t(\theta)$. Rearranging and exponentiating gives

$$\exp(\mathbb{E}_\rho h_t(\theta) - \gamma_{\mathcal{G}_t}(\rho, \nu)) \leqslant \exp \mathbb{E}_\nu h_t(\theta) \leqslant \mathbb{E}_\nu \exp h_t(\theta).$$

Since $(\exp h_t(\theta))_{t\geqslant t_0}$ is a super or submartingale by assumption and ν is data-free, the process $(\mathbb{E}_\nu \exp h_t(\theta))_{t\geqslant t_0}$ is also a super or submartingale by Lemma 46. Therefore, Ville's inequality gives

$$\mathbb{P}\big(\exists t \geqslant t_0 : \exp\left\{\mathbb{E}_\rho h_t(\theta) - \gamma_{\mathcal{G}_t}(\rho,\nu)\right\} \geqslant 1/\delta\big) \leqslant \mathbb{P}\big(\exists t \geqslant 1 : \mathbb{E}_\nu \exp h_t(\theta) \geqslant 1/\delta\big) \leqslant \delta.$$

Since ρ was arbitrary, this yields that with probability at least $1-\delta$,

$$\exp\left\{\mathbb{E}_\rho h_t(\theta) - \gamma_{\mathcal{G}_t}(\rho,\nu)\right\} \leqslant 1/\delta,$$

for all $t \geqslant t_0$ and ρ. Rearranging gives the desired result. ■

Following Amit et al. (2022), we let the family of functions $\mathcal{G}_t$ be a function of the timestep (hence possibly dependent on data $Z_1, \dots, Z_t$). Sections 4 and 5 are replete with processes $(\exp h_t(\theta))$ which are super and submartingales, each of which furnishes a separate bound after applying Theorem 29. We will not list them all here, trusting that practitioners can combine results as befits their problem of interest. We will, however, state the following consequence of Theorem 29 in order to compare our results with those of Amit et al. (2022). In what follows, we use notation and concepts introduced in Section 5, such as $\bar{t} = 2^{\lfloor \log_2(t) \rfloor}$, $\mathsf{IL}_t = \log(\log_2(2t)\zeta(2))$, $\varphi_t(\theta) = \varphi(\widehat{R}_t(\theta), R(\theta))$, and the exchangeable reverse filtration $(\mathcal{E}_t)$. We also assume a stationary loss function.

Corollary 30 *Let (Z_t) be exchangeable, and let $\varphi : \mathbb{R}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ be a convex function. Fix a prior $\nu \in \mathcal{M}(\Theta)$. Consider a family of functions $(\mathcal{G}_t)$ with $\mathcal{G}_t : \Theta \to \mathbb{R}$ and let (λ_t) be a positive sequence such that, for all natural numbers $k \geqslant 0$,*

$$\lambda_{2^k}\varphi_t(\theta) - \log \mathbb{E}_{\mathcal{D}} \exp(\lambda_{2^k}\varphi_{2^k}(\theta)) \in \mathcal{G}_t, \quad \text{for all } t \geqslant 1.$$

Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all $\rho \in \mathcal{M}(\Theta)$ and at all times $t \geqslant 1$,

$$\mathbb{E}_\rho \varphi_t(\theta) \leqslant \frac{\log \mathbb{E}_\nu \mathbb{E}_{\mathcal{D}} \exp(\lambda_{\bar{t}}\varphi_{\bar{t}}(\theta))}{\lambda_{\bar{t}}} + \frac{\gamma_{\mathcal{G}_t}(\rho,\nu) + \log(1/\delta) + \mathsf{IL}_t}{\lambda_{\bar{t}}}. \tag{44}$$

Moreover, suppose n is some fixed time of interest, and that $\lambda\varphi_t(\theta) - \log \mathbb{E}_{\mathcal{D}} \exp(\lambda\varphi_n(\theta)) \in \mathcal{G}_t$ for all $t \geqslant n$ and some $\lambda > 0$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all $t \geqslant n$:

$$\mathbb{E}_\rho \varphi_t(\theta) \leqslant \frac{\log \mathbb{E}_\nu \mathbb{E}_{\mathcal{D}} \exp(\lambda\varphi_n(\theta))}{\lambda} + \frac{\gamma_{\mathcal{G}_t}(\rho,\nu) + \log(1/\delta)}{\lambda}. \tag{45}$$

A proof sketch is provided in Appendix A.12. The previous result parallels Corollaries 22 and 24 but using IPMs instead of the KL divergence. The reliance of (44) on $\bar{t}$ and IL_t once again arises from stitching. For the fixed time $t = n$, (45) gives a generalized version of Corollaries 4 and 5 in Amit et al. (2022). Those results are obtained by considering particular functions φ, as was done in Section 5.2. As noted by Amit et al. (2022), the above bounds are merely "templates" in the sense that, to be insightful, one must choose a family of functions $\mathcal{G}_t$. A bound based on the total variation distance can be achieved by considering the family $\mathcal{G}_t = \{g : \Theta \to [0,\infty) : \|g\|_\infty \leqslant 1\}$, and one based on the Wasserstein distance can be achieved by appealing to Kantorovich-Rubinstein duality. We refer the reader to Amit et al. (2022) for the details of these bounds.

6.2 ϕ-divergences and Rényi divergences

The KL divergence is a member of a more general family of divergences termed ϕ-divergences (Ali and Silvey, 1966) (often called f-divergences, but we have reserved f for our loss). For a convex function $\phi : \mathbb{R} \to \mathbb{R}$, the *$\phi$-divergence* between measures ρ and ν over Θ such that $\rho \ll \nu$ is

$$D_\phi(\rho\|\nu) = \int_\Theta \phi\left(\frac{\mathrm{d}\rho}{\mathrm{d}\nu}\right)\mathrm{d}\nu = \mathbb{E}_{\theta\sim\nu}\left[\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(\theta)\right]. \tag{46}$$

The KL divergence is recovered by considering $\phi(x) = x\log x$. ϕ-divergences are a nearly orthogonal set of divergences from IPMs, considered in the previous section. Indeed, the total variation distance is the only (non-trivial) divergence which is both an IPM and a ϕ-divergence (Sriperumbudur et al., 2012).

The Donsker-Varadhan formula for the KL divergence is an improvement on a more general variational representation of ϕ-divergences (e.g., Sriperumbudur et al., 2009), which states the following. For any measures ρ and ν and any convex function $\phi : \mathbb{R} \to \mathbb{R}$,

$$D_\phi(\rho\|\nu) \geqslant \mathbb{E}_\rho[h(\theta)] - \mathbb{E}_\nu[\phi^*(h(\theta))], \tag{47}$$

where ϕ^* is the convex conjugate of ϕ, i.e.,

$$\phi^*(y) = \sup_{x\in\mathbb{R}}\{xy - \phi(x)\}.$$

We can use (47) in place of the Donsker-Varadhan formula in Theorem 4, where the term $\mathbb{E}_\nu\phi^*(h(\theta))$ replaces $\log\mathbb{E}_\nu\exp h(\theta)$.

Theorem 31 (Anytime PAC-Bayes with ϕ-divergences) *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. Let $P(\theta) = (P_t(\theta))_{t=1}^\infty$ be a stochastic process such that, for all $\theta \in \Theta$, $\exp\mathbb{E}_\nu[\phi^*(P_t(\theta))]$ is a supermartingale or reverse submartingale (adapted to some underlying filtration). Suppose that $\exp\mathbb{E}_\nu[\phi^*(P_1(\theta))] \leqslant 1$. Then, for any $\delta \in (0,1)$ and prior $\nu \in \mathcal{M}(\Theta)$, with probability at least $1-\delta$,*

$$\mathbb{E}_\rho P_t(\theta) \leqslant D_\phi(\rho\|\nu) + \log(1/\delta), \tag{48}$$

for all times $t \geqslant 1$ and $\rho \in \mathcal{M}(\Theta)$.

Proof Set $V_t^{\mathrm{mix}} := \exp\sup_\rho\left\{\mathbb{E}_{\theta\sim\rho}[P_t(\theta)] - D_\phi(\rho\|\nu)\right\}$. The variational formula for D_ϕ gives $V_t^{\mathrm{mix}} \leqslant \exp\mathbb{E}_\nu[\phi^*(P_t(\theta))]$, so by assumption, V_t^{mix} is upper bounded by a nonnegative supermartingale or reverse submartingale. From here, the proof follows that of Theorem 4. ■

The key distinction between this result and Theorem 4 is that while the latter posits that $\exp P(\theta)$ is (upper bounded by) a nonnegative super/submartingale, here we assume that $\exp\mathbb{E}_\nu[\phi^*(P(\theta)]$ plays this role. We consider establishing functions ϕ and processes $P(\theta)$ such that $\exp\mathbb{E}_\nu\phi^*(P(\theta))$ has this property to be an interesting line of future research. We note that Theorem 31 cannot strictly be called a generalization of Theorem 4 as the latter relies on the Donsker-Varadhan formula which is tighter than the variational formula for the KL divergence given by (47).

Another (related) family of distances is the *Rényi divergence*. Here, for measures $\rho \ll \nu$ and any $\alpha \in (0,1) \cup (1,\infty)$, we define

$$D_\alpha(\rho\|\nu) := \frac{1}{1-\alpha}\mathbb{E}_{\theta\sim\rho}\left[\left(\frac{\rho(\theta)}{\nu(\theta)}\right)^\alpha\right].$$

As $\alpha \to 1$, $D_\alpha(\rho\|\nu) \to D_{\mathrm{KL}}(\rho\|\nu)$, so by continuity we define $D_1(\rho\|\nu) = D_{\mathrm{KL}}(\rho\|\nu)$. The Rényi divergence yields the following variational formula, which can be seen as an extension of the Donsker-Varadhan formula (Lemma 3). It was given by Bégin et al. (2016).

Lemma 32 *Let $h : \Theta \to \mathbb{R}$ be measurable. For any measures ρ and ν, with $\rho \ll \nu$, we have*

$$\log \mathbb{E}_\nu[h(\theta)^{\frac{\alpha}{\alpha-1}}] \geqslant \frac{\alpha}{\alpha-1}\log \mathbb{E}_\rho h(\theta) - D_\alpha(\rho\|\nu),$$

for all $\alpha \in (0,1) \cup (1,\infty)$.

Using this formula, one can give a Theorem in the style of Theorem 4 and 31 for α-divergences.

Theorem 33 *Set $\alpha > 1$. Let $P(\theta) = (P_t(\theta))_{t\geqslant t_0}$ be a stochastic process such that, for all $\theta \in \Theta$, $\exp(P^{\frac{\alpha}{\alpha-1}}(\theta))$ is a supermartingale or reverse submartingale (adapted to some underlying filtration) obeying $\mathbb{E}_{\mathcal{D}} \exp P_{t_0}^{\frac{\alpha}{\alpha-1}}(\theta) \leqslant 1$. Then, for any $\delta \in (0,1)$ and prior $\nu \in \mathcal{M}(\Theta)$, with probability at least $1-\delta$,*

$$\mathbb{E}_\rho[P_t(\theta)] \leqslant \frac{\alpha-1}{\alpha}(D_\alpha(\rho\|\nu) + \log(1/\delta)), \tag{49}$$

for all times $t \geqslant t_0$ and $\rho \in \mathcal{M}(\Theta)$.

Proof Following Theorems 4 and 31, put $V_t^{\mathrm{mix}} = \exp\sup_\rho\{\frac{\alpha}{\alpha-1}\log \mathbb{E}_\rho \exp P_t(\theta) - D_\alpha(\rho\|\nu)\}$. Then $V_t^{\mathrm{mix}} \leqslant \mathbb{E}_\nu \exp P_t^{\frac{\alpha}{\alpha-1}}(\theta)$ by Lemma 32, where the process $(\mathbb{E}_\nu \exp P_t^{\frac{\alpha}{\alpha-1}}(\theta))_{t\geqslant t_0}$ is a nonnegative supermartingale or reverse submartingale by assumption and Lemma 46. It also has initial expected value at most 1 by assumption. Therefore, $\mathbb{P}(\exists t \geqslant t_0 : V_t^{\mathrm{mix}} \geqslant 1/\delta) \leqslant \mathbb{P}(\exists t \geqslant t_0 : \mathbb{E}_\nu \exp P_t^{\frac{\alpha}{\alpha-1}}(\theta) \geqslant 1/\delta) \leqslant \delta$ by Ville's inequality. Rearranging the inequality $V_t^{\mathrm{mix}} \leqslant 1/\delta$, we obtain that with probability at least $1-\delta$,

$$\mathbb{E}_\rho P_t(\theta) \leqslant \log \mathbb{E}_\rho \exp P_t(\theta) \leqslant \frac{\alpha-1}{\alpha}\bigg(D_\alpha(\rho\|\nu) + \log(1/\delta)\bigg),$$

for all ρ and $t \geqslant t_0$, as claimed. ■

Theorem 33 suggests the question: When is $\exp(P^{\frac{\alpha}{\alpha-1}}(\theta))$ a supermartingale or reverse submartingale? There are several candidates. By Jensen's inequality, a sufficient condition for this quantity to be a reverse submartingale is for $P(\theta)$ to also be a reverse submartingale. Indeed, if (N_t) is a reverse submartingale with respect to $(\mathcal{R}_t)$, then

$$\mathbb{E}[N_t^{\frac{\alpha}{\alpha-1}}|\mathcal{R}_{t+1}] \geqslant \mathbb{E}[N_t|\mathcal{R}_{t+1}]^{\frac{\alpha}{\alpha-1}} \geqslant N_{t+1}^{\frac{\alpha}{\alpha-1}}, \tag{50}$$

since $x \mapsto x^{\frac{\alpha}{\alpha-1}}$ is convex. However, to apply Ville's inequality, one would also need to control $\mathbb{E}_{\mathcal{D}} N_1^{\frac{\alpha}{\alpha-1}}$ which is less easily done, even if $\mathbb{E}_{\mathcal{D}} N_1 \leqslant 1$. One might also consider using the processes employed in the proof of Corollary 22, but raised to the $(\alpha-1)/\alpha$. In that case, of course, raising the result to the $\alpha/(\alpha-1)$ power would result in the original process. However, in this case we achieve the same bound as Corollary 22, but with $D_{\mathrm{KL}}(\rho\|\nu)$ replaced by $D_\alpha(\rho\|\nu)$. This a weaker result since $D_\alpha(\rho\|\nu) \geqslant D_{\mathrm{KL}}(\rho\|\nu)$ for all $\alpha > 0$. Instead, to take advantage of Lemma 32, we construct an altogether different process. This leads to the following result. As in Section 5 we consider a stationary loss function and exchangeable data. Recall the shorthand $\varphi_t(\theta) = \varphi(\widehat{R}_t(\theta), R(\theta))$ for a convex function φ, as well as the quantities $\bar{t} = 2^{\lfloor \log_2(t) \rfloor}$ and $\mathsf{IL}_t = \log(\log_2^2(2t)\zeta(2))$.

Corollary 34 *Let (Z_t) be exchangeable. Let $\varphi : \mathbb{R}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{>0}$ be a convex function and $\nu \in \mathcal{M}(\Theta)$ be a prior. Put $\alpha > 1$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all $\rho \in \mathcal{M}(\Theta)$ and at all times $t \geqslant 1$,*

$$\log \mathbb{E}_\rho \varphi_t(\theta) \leqslant \frac{\alpha-1}{\alpha}\left(D_\alpha(\rho\|\nu) + \log \mathbb{E}_{\nu,\mathcal{D}}[\varphi_{\bar{t}}^{\frac{\alpha}{\alpha-1}}(\theta)] + \log(1/\delta) + \mathsf{IL}_t \right). \tag{51}$$

The proof is provided in Appendix A.13. Similarly to Corollary 24, we can obtain a version of the above result which holds for all times $t \geqslant n$ for some pre-selected time n. These results constitute a time-uniform extension of Theorem 9 in Bégin et al. (2016), who give a fixed-time version for binary classification. By taking $\alpha = 2$, we obtain a PAC-Bayes bound using the χ^2 divergence (see Bégin et al., 2016, Corollary 10). We note that unlike Corollary 22, the above result is a bound on the logarithm of φ. By exponentiating both sides, one obtains an intriguing PAC-Bayes bound in multiplicative form.

6.3 Confidence Sequences and Choice of (λ_t)

Our anytime-valid bounds enable us, under some circumstances, to construct time-uniform *confidence sequences*, i.e., sequences of sets which contain the true parameter of interest at all times with high probability (Darling and Robbins, 1967a; Lai, 1976). In our setting, the parameter of interest is the conditional mean $\frac{1}{t}\sum_{i=1}^t \mathbb{E}_{\theta\sim\rho}\mu_i(\theta)$, where $\mu_i(\theta) = \mathbb{E}_{\mathcal{D}}[f_i(Z_i,\theta)|\mathcal{F}_{i-1}]$. A $(1-\delta)$-confidence sequence is then a random sequence $(C_t(\rho,\nu))_{t=1}^\infty$ such that

$$\mathbb{P}\left(\forall t \geqslant 1 : \frac{1}{t}\sum_{i=1}^t \mathbb{E}_{\theta\sim\rho}\mu_i(\theta) \in C_t(\rho,\nu) \right) \geqslant 1-\delta. \tag{52}$$

Observe that the confidence sequence depends on the prior ν and posterior ρ. It does not hold simultaneously across all such distributions.

While we allow the conditional mean $t^{-1}\sum_{i\leqslant t}\mu_i(\theta)$ to change over time in general, let us begin the discussion with the case of a common conditional mean and stationary loss function f. More precisely, we assume that $\mu(\theta) = \mu_t(\theta) = \mathbb{E}_{\mathcal{D}}[f(Z_t,\theta)|\mathcal{F}_{t-1}]$ is unchanging as a function of time. Many of the bounds generated in previous sections are based on processes which are themselves based on tail bounds on the term $\lambda\Delta_i(\theta) = \lambda(\mu_i(\theta) - f(Z_i,\theta))$. By considering $-\Delta_i(\theta)$ and applying the union bound, we may obtain a confidence sequence. For instance, the following confidence sequence may be derived from Corollary 7.

Corollary 35 *Let f be σ-subGaussian and let $(Z_t) \sim \mathcal{D}$ be such that $\mu(\theta) = \mathbb{E}_{\mathcal{D}}[f(Z,\theta)|\mathcal{F}_{t-1}]$ is constant for all $t \geqslant 1$. Fix a prior $\nu \in \mathcal{M}(\Theta)$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all ρ and $t \geqslant 1$,*

$$\mathbb{E}_\rho \mu(\theta) \in \left(\frac{\sum_{i=1}^t \lambda_i f(Z_i,\theta)}{\sum_{i=1}^t \lambda_i} \pm W_t \right), \quad \textit{where} \quad W_t := \frac{\log(2/\delta) + D_{\mathrm{KL}}(\rho\|\nu) + \frac{\sigma^2}{2}\sum_{i=1}^t \lambda_i^2}{\sum_{i=1}^t \lambda_i} \tag{53}$$

We note the factor of 2 in $\log(2/\delta)$ comes from the union bound. We state the above proposition as an example only; many other confidence sequences may be derived from the arguments throughout Sections 4 and 5.

Studying confidence sequences provides an opportunity to demonstrate why we allow λ_t to change as a function of time. It is desirable that the width of the sequence, W_t, goes to 0 as $t \to \infty$ so that the confidence sequence asymptotically converges on the correct value with high probability. This would not be possible with fixed λ, as W_t would converge to $\sigma^2\lambda/2 \neq 0$. On the other hand, following Waudby-Smith and Ramdas (2023), if we instead consider $\lambda_t \asymp (t \log t)^{-1/2}$, then we have $W_t = \widetilde{O}(\sqrt{\log(t)/t})$, where $\widetilde{O}$ hides log-log factors. Further, we can attain the optimal rate $O(\sqrt{\log\log t/t})$ due to the Law of the Iterated Logarithm (LIL) (Darling and Robbins, 1967b) by the same technique of geometrically spaced union bounds that was used in Section 5.1. Such a result applied to Corollary 35 is stated and proved in Appendix A.15, but is omitted here in favor of the following discussion which is more general and also provides a LIL bound.

Let us turn now to the case when $\mu_t(\theta)$ is not assumed to be independent of t. Similarly to Corollary 35, a union bound applied to Corollary 7 tells us that

$$\sum_{i=1}^t \lambda_i \mathbb{E}_\rho \mu_i(\theta) \in \left(\sum_{i=1}^t \frac{\lambda_i^2 \sigma_i^2}{2} \pm [D_{\mathrm{KL}}(\rho||\nu) + \log(2/\delta)] \right),$$

for all $t \geqslant 1$ with probability at least $1-\delta$. However, this does not yield a closed-form expression for a confidence sequence. To construct an explicit confidence sequence with optimal width, we turn once again to stitching. The technique we use is applicable to general sub-ψ processes (Section 4.1), but we demonstrate it in the case of 1-subGaussian losses for simplicity.

Corollary 36 *Let f_i be $1-$subGaussian and fix a prior $\nu \in \mathcal{M}(\Theta)$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of (Z_t), for all ρ and $t \geqslant 1$:*

$$\frac{1}{t}\sum_{i=1}^t \mathbb{E}_\rho \mu_i(\theta) \in \left(\frac{1}{t}\sum_{i=1}^t \mathbb{E}_\rho f_i(Z_i,\theta) \pm W_t \right),$$

where

$$W_t \lesssim \frac{\sqrt{\log(\log(t)) + \log(1/\delta)}}{\sqrt{t}} + \frac{D_{\mathrm{KL}}(\rho\|\nu)}{\sqrt{t\log(\log(t)) + t\log(1/\delta)}}.$$

The proof can be found in Appendix A.14. There has been much recent work on developing sequences (λ_t) which achieve optimal shrinkage rates; we refer the interested reader

to Catoni (2012); Howard et al. (2021); Waudby-Smith and Ramdas (2023); Wang and Ramdas (2023a,b) for further discussion on this point.

We end this section by noting that we have now deployed the stitching technique in two capacities. In Section 5 it was used to apply a different reverse submartingale in each epoch, whereas in the above result it was used to choose appropriate constants in each epoch. While the intuition behind stitching is similar, the two applications yield different results. The former loses some tightness compared to fixed-time bounds, while the latter enables us to achieve optimal rates.

6.4 Martingale Difference Sequences

Throughout this work we've considered loss functions f_t acting on $\mathcal{Z}$ and Θ. While this is a natural setting for PAC-Bayes analysis owing to its connections to learning theory, different settings have been considered. Seldin et al. (2012) and Balsubramani (2015), for instance, consider PAC-Bayesian inequalities for martingale difference sequences. In this section we briefly demonstrate that our results extend to this setting. This is due to the fact that our workhorse, Theorem 4, holds for general stochastic processes.

We consider a sequence of random functions (F_t) such that $F_t : \Theta \to \mathbb{R}$. We suppose that (F_t) is a *martingale difference sequence*, i.e., $\mathbb{E}[F_t|\mathcal{F}_{t-1}] = 0$ for all $t \geqslant 1$, where $\mathcal{F}_t = \sigma(F_1, \dots, F_t)$. That is, $\mathbb{E}[F_t(\theta)|\mathcal{F}_{t-1}] = 0$ for all $\theta \in \Theta$. Note that the expectation is over the functions themselves, not over θ. Let $S_t = \sum_{i=1}^t F_i$ (and, by extension, $S_t(\theta) = \sum_{i=1}^t F_i(\theta)$).

First, suppose the F_t are bounded, say $F_t : \Theta \to [\alpha_t, \beta_t]$. Just as we did in Corollary 7, we can consider the nonnegative process $N_t(\theta) = \exp\left\{ \sum_{i=1}^t \lambda_i F_i(\theta) - \frac{1}{8}\sum_{i=1}^t \lambda_i^2(\beta_i - \alpha_i)^2 \right\}$, which is a supermartingale since $\mathbb{E}[F_i|\mathcal{F}_{i-1}] = 0$. (Note that we have substituted $(\beta_i - \alpha_i)^2/4$ for σ_i^2 in (9), since F_i is $(\beta_i - \alpha_i)/2$-subGaussian.) This process, in conjunction with Theorem 4, leads to the following result, which is the time-uniform extension of Theorem 5 of Seldin et al. (2012).

Corollary 37 (Anytime bound for bounded MDSs I) *Let (F_t) be a martingale difference sequence where $F_t : \Theta \to [\alpha_t, \beta_t]$. Let $\nu \in \mathcal{M}(\Theta)$ be a prior and (λ_t) a nonnegative predictable sequence. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the sequence of functions, for all $t \geqslant 1$ and $\rho \in \mathcal{M}(\Theta)$,*

$$\sum_{i=1}^t \lambda_i \mathbb{E}_\rho F_i(\theta) \leqslant \frac{1}{8}\sum_{i=1}^t \lambda_i^2(\beta_i - \alpha_i)^2 + D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta).$$

Using similar techniques, we can provide a time-uniform version of Theorem 7 in Seldin et al. (2012), a result which also undergirds the main theorem of Balsubramani (2015).

Corollary 38 (Anytime bound for bounded MDSs II) *Let (F_t) be a martingale difference sequence where $F_t : \Theta \to \mathbb{R}$ such that $|F_t(\theta)| \leqslant H$ for all $\theta \in \Theta$. Let $\nu \in \mathcal{M}(\Theta)$ be a prior and $\lambda \in [0, 1/H]$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ over the sequence of functions,*

$$\sum_{i=1}^t F_i(\theta) \leqslant \lambda(e-2)\sum_{i=1}^t \mathbb{E}[F_i^2(\theta)|\mathcal{F}_{i-1}] + \frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta)}{\lambda},$$

for all $t \geqslant 1$ and $\rho \in \mathcal{M}(\Theta)$.

Note that because F_t is bounded, all (conditional) moments exists. The bound is therefore non-vacuous by assumption. The proof of the above result (and the statement itself) is very similar to that of Corollary 12, and is thus omitted. Like that proposition, here λ could be taken to be a sequence $\{\lambda_t\} \subseteq [0, 1/H]$, but we leave it stationary for easier to comparison to prior work.

Theorem 1 of Balsubramani (2015) is based on Corollary 38 and then choosing λ strategically (and stochastically) in order to tighten the bound. Such techniques have also been used to generate sharp martingale concentration bounds. Seldin et al. (2012) also optimize over λ in the fixed-time version of Corollary 37 in order to provide a tighter bound (see their Theorems 5 and 6). An anytime version of this result would follow from applying the same procedure to Corollary 37, though we note that their optimization procedure employs knowledge of the sample size and thus cannot be replicated precisely in the anytime setting.

Our final result generalizes Theorem 4 of Seldin et al. (2012), by providing a version of Corollary 22 for difference sequences. Here we will broaden the setting slightly from martingale difference sequences and let $\mathbb{E}[F_t|\mathcal{F}_{t-1}] = G$ for all $t \geqslant 1$ some $G : \Theta \to \mathbb{R}$, meaning that $\mathbb{E}[F_t(\theta)|\mathcal{F}_{t-1}] = G(\theta)$ for all $\theta \in \Theta$. The proof of the following bound uses precisely the same mechanics as that of Corollary 22, so we do not provide it. Recall that $\bar{t} = 2^{\lfloor \log_2(t) \rfloor}$ and $\mathsf{IL}_t = \log(\log_2^2(2t)\zeta(2))$.

Corollary 39 *Let (F_t) be a random exchangeable sequence of functions with $F_t : \Theta \to \mathbb{R}$ such that $\mathbb{E}[F_t|\mathcal{F}_{t-1}] = G$ for all $t \geqslant 1$ and some fixed $G : \Theta \to \mathbb{R}$. Let $\varphi : \mathbb{R}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ be a convex function and set $S_t = \sum_{i=1}^t F_i$. Fix a prior $\nu \in \mathcal{M}(\Theta)$ and let (λ_t) be a positive sequence of real numbers. Then, for all $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the sequence of functions, for all $\rho \in \mathcal{M}(\Theta)$ and at all times $t \geqslant 1$,*

$$\mathbb{E}_\rho \varphi\left(\frac{1}{t} S_t(\theta), G(\theta)\right) \leqslant \frac{\log \mathbb{E}_\nu \mathbb{E} \exp(\lambda_{\bar{t}} \varphi(\frac{1}{\bar{t}} S_{\bar{t}}(\theta), G(\theta)))}{\lambda_{\bar{t}}} + \frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta) + \mathsf{IL}_t}{\lambda_{\bar{t}}}.$$

A time-uniform version of Theorem 4 of Seldin et al. (2012) follows from the above bound by taking $\varphi = \mathsf{kl}$ (and taking $\lambda_t = \lambda$ for all t) as was done in both Sections 5.2 and 6.1. Finally, we note that Kuzborskij and Szepesvári (2019) also discuss bounds based on martingale difference sequences. In particular, they use the Doob decomposition to construct a canonical difference sequence when estimating a general function, and then bound the increments using an empirical Efron-Stein like term. We discuss their work more thoroughly in Section 4, where we relate it to sub-ψ processes.

6.5 Data-dependent Priors

Several recent works have investigated the role of data-dependent priors (Rivasplata et al., 2020; Awasthi et al., 2020). The appeal is clear: A well chosen prior with mass close to the true parameter will typically enable much tighter bounds. Historically, this is often achieved with sample splitting, i.e., reserving some fraction of the sample to choose the prior, and then computing the bound on the remaining data (Parrado-Hernández et al., 2012; Dziugaite and Roy, 2017). Here we provide some remarks on how to extend our analysis to allow for data-dependent priors.

To set the stage, let $P(\theta) = (P_t(\theta))_{t \geqslant 1}$ be a stochastic process such that $\exp P(\theta)$ is a supermartingale. In our results thus far, the prior ν is data-free to ensure that

Fubini-Tonelli can be applied so that the mixture $\mathbb{E}_{\theta\sim\nu}[\exp P(\theta)]$ is also a supermartingale. We can, however, weaken this condition slightly. If ν is $\mathcal{F}_{t_0-1}$-measurable, then the process $(\mathbb{E}_{\theta\sim\nu}\exp P_t(\theta))_{t\geqslant t_0}$ remains a supermartingale, since ν is deterministic at time t_0. Of course, different priors result in different processes: for $\nu_i \in \mathcal{M}(\Theta)$ which is $\mathcal{F}_{t_i-1}$-measurable, the process $(\mathbb{E}_{\nu_1}\exp U_t(\theta))_{t\geqslant t_1}$ can be distinct from $(\mathbb{E}_{\nu_2}\exp U_t(\theta))_{t\geqslant t_2}$. Thus, a bound which covers changing priors must cover different processes. We can ensure such coverage but at the price of a union bound.

More formally, let $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$, $N \in \mathbb{N}\cup\{\infty\}$ be a set of times at which we change priors. We allow the times t_i to be stopping times (i.e., $\{t_i \leqslant n\}$ is $\mathcal{F}_n$-measurable for all n), and we allow $|\mathcal{T}|$ to be infinite (e.g., $t_i = 2^i$ is an example of a deterministic set of times satisfying the condition). Suppose we begin with prior ν_0. At time $t_i \in \mathcal{T}$ a new prior ν_i is used, where ν_i is $\mathcal{F}_{t_i-1}$-measurable. Define, for each $\theta\in\Theta$,

$$\mathfrak{M}_t(\theta) = \begin{cases} P_t(\theta) - D_{\mathrm{KL}}(\rho\|\nu_0), & 1 \leqslant t < t_1, \\ P_t(\theta) - D_{\mathrm{KL}}(\rho\|\nu_1), & t_1 \leqslant t < t_2, \\ \vdots & \vdots \end{cases} \tag{54}$$

When $\mathcal{T} = \emptyset$, a bound on $\mathbb{E}_{\theta\sim\rho}\mathfrak{M}_t(\theta)$ for all ρ and $t \geqslant 1$ is given by Theorem 4. The following theorem generalizes this result and provides a bound for an arbitrary number of priors. However, we pay a price (a loosening of the bound) for each change of prior. Interestingly, our result does not allow for the stochastic process to be a reverse submartingale, only a forward supermartingale. This is because, if $(\mathcal{R}_t)$ is a reverse filtration and ν_i is $\mathcal{R}_{t_0-1}$-measurable, it may not be $\mathcal{R}_t$ measurable for $t \geqslant t_0$ since $\mathcal{R}_t \supseteq \mathcal{R}_{t+1}$. Thus, the mixtures may cease to be submartingales as time advances.

Theorem 40 *For each $\theta\in\Theta$, let $P(\theta) = (P_t(\theta))_{t=1}^{\infty}$ be a stochastic process such that $\exp P(\theta)$ is a supermartingale adapted to the filtration $(\mathcal{F}_t)$ and $\mathbb{E}\exp P_1(\theta) \leqslant 1$. Let $\mathcal{T}$ and $\mathfrak{M}_t$ be as above. Then, for any $\delta\in(0,1)$, with probability at least $1-\delta$, for all $t\geqslant 1$ and $\rho\in\mathcal{M}(\Theta)$,*

$$\mathbb{E}_\rho[\mathfrak{M}_t(\theta)] \leqslant \log\left(\frac{(t^\dagger+1)(t^\dagger+2)}{\delta}\right), \tag{55}$$

where $t^\dagger$ is the number of times the prior has been changed up to and including time t.

Proof Let us briefly clarify notation. We begin with prior ν_0, and switch to ν_1 at time t_1, switch to ν_2 at time t_2, and so on. We set $t_0 = 1$ for convenience. The optional stopping theorem for nonnegative supermartingales implies that for each t_i, $\mathbb{E}\exp P_{t_i}(\theta) \leqslant \mathbb{E}\exp P_1(\theta) \leqslant 1$ (Durrett, 2019, Theorem 5.7.6). Moreover, for each θ, the process $\exp(P_{t\vee t_i}(\theta))_{t\geqslant 0}$ is a supermartingale adapted to the filtration $(\mathcal{F}_{t\vee t_i})$ (Klenke, 2013, Theorem 10.15). Combining this with Lemma 46 and the fact that ν_i is $\mathcal{F}_{t_i-1}$-measurable implies that the process $(\mathbb{E}_{\nu_i}[\exp(P_{t\vee t_i}(\theta))])_{t\geqslant 0}$ is a supermartingale on the filtration $(\mathcal{F}_{t\vee t_i})$. Applying Theorem 4, we have that for all times t_i,

$$\mathbb{P}\left\{\exists t\geqslant t_i, \exists\rho\in\mathcal{M}(\Theta) : \mathbb{E}_\rho P_t(\theta) - D_{\mathrm{KL}}(\rho\|\nu_i) \geqslant \log\left(\frac{s(i)}{\delta}\right)\right\} \leqslant \frac{\delta}{s(i)},$$

where $s(i) = (i+1)(i+2)$. If at time t we are using the prior ν_i, then we have switched priors i times, so $t^\dagger = i$. Therefore,

$$\left\{\exists t \geqslant 1, \exists \rho : \mathbb{E}_\rho \mathfrak{M}_t(\theta) \geqslant \log(s(t^\dagger)/\delta)\right\} \subseteq \left\{\exists i \geqslant 0, \exists t \geqslant t_i, \exists \rho : \mathbb{E}_\rho \mathfrak{M}_t(\theta) \geqslant \log(s(i)/\delta)\right\}.$$

The union bound then implies that $\mathbb{P}(\exists t \geqslant 1, \exists \rho : \mathbb{E}_\rho \mathfrak{M}_t(\theta) \geqslant \log(s(t^\dagger)/\delta))$ is bounded by

$$\sum_{i \geqslant 0} \mathbb{P}\left(\{\exists t \geqslant t_i, \exists \rho : \mathbb{E}_\rho P_t(\theta) - D_{\mathrm{KL}}(\rho \| \nu_i) \geqslant \log(s(i)/\delta)\}\right) \leqslant \sum_{i \geqslant 0} \frac{\delta}{s(i)} = \delta.$$

completing the proof. ■

It's perhaps worth noting that there is nothing special about our function s in the above proof, and it only needs to satisfy $\sum_{i=0}^{\infty} \frac{1}{s(i)} \leqslant 1$. Given such a function, the resulting bound reads $\mathbb{E}_\rho[\mathfrak{M}_t(\theta)] \leqslant \log(s(t^\dagger)/\delta)$ in place of (55).

Remark 41 *If $\mathcal{T}$ is finite and deterministic, then we can replace* (55) *with $\mathbb{E}_\rho \mathfrak{M}_t(\theta) \leqslant \log(|\mathcal{T}|/\delta)$, thus reducing the quadratic dependence on the number of priors used to a linear dependence. This can be seen by setting $s(i) = |\mathcal{T}|$ for all i in the above analysis, and noting that the final union bound need only cover $|\mathcal{T}|$ events.*

Our result has a different flavor than those of Rivasplata et al. (2020) and Awasthi et al. (2020). However, we believe its general form is useful and interpretable: For every switch of the prior, the bound suffers an additional additive logarithmic factor.

6.6 Application: Gaussian Process Classification

Here we follow Seeger (2002) and apply the PAC-Bayes framework to Gaussian process classification. We take $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ and consider a supervised classification problem with features $x \in \mathcal{X}$ and binary labels $y(x)$. For a prediction $\hat{y}(x)$ our loss is the 0-1 loss $\mathbf{1}(\hat{y}(x) \neq y(x))$. We assume that the labels $y(x)$ are generated as $y = \operatorname{sgn} \theta(x)$, where $\theta : \mathcal{X} \to \mathbb{R}$ is some function in a nonparametric family Θ. We adopt a Bayesian perspective and consider θ to be a random function. Accordingly, we place a zero-mean Gaussian process prior ν over Θ, i.e., $\theta \sim \mathcal{GP}(0, k)$, where k is a Mercer kernel. More precisely, given $x = (x_1, \ldots, x_t)$, we have

$$\nu(\theta(x)) = \frac{|K_x|^{-1/2}}{(2\pi)^{t/2}} \exp\left(-\frac{1}{2} x^\intercal K_x^{-1} x\right),$$

where K_x is the symmetric matrix whose ij-th entry is given by $k(x_i, x_j)$. Given a distribution ρ over Θ, the empirical risk at time t is

$$\mathbb{E}_{\theta \sim \rho} \widehat{R}_t(\theta) = \frac{1}{t} \sum_{i \leqslant t} \mathbb{P}_{\theta \sim \rho}(\operatorname{sgn} \theta(x_i) \neq y_i), \tag{56}$$

and the expected risk is

$$\mathbb{E}_{\theta \sim \rho} R(\theta) = \mathbb{E}_{\theta \sim \rho} \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbf{1}(\operatorname{sgn} \theta(x) \neq y). \tag{57}$$

Note that the expected risk is constant because the data are i.i.d. and the loss function is stationary. The posterior predictions $\theta(x^t)$ given training data $x^t = (x_1, \dots, x_t)$ and $y^t = (y_1, \dots, y_t)$ depend on the distribution of $y|\theta$, but can be written abstractly as

$$\rho_t(\theta(x^t)|x^t, y^t) = N(K_{x^t}\mu_t, \Sigma_t),$$

for some $\mu_t \in \mathbb{R}^t$, $\Sigma_t \in \mathbb{R}^{t\times t}$. We have introduced the subscript t on ρ_t to emphasize that this is the posterior at time t. Here $N(\cdot,\cdot)$ is a multivariate normal. Having introduced μ_t and Σ_t, we can now write down the KL divergence between the prior ν and posterior ρ_t, which reduces to the KL divergence between their finite-dimensional distributions on the training data (given that ρ_t is defined via the Bayesian update rule), and can be thus calculated via the usual KL divergence between multivariate Gaussians (see Seeger, 2002 for the derivation):

$$D_{\mathrm{KL}}(\rho_t\|\nu) = D_{\mathrm{KL}}(\rho_t(x^t)\|\nu(x^t)) = \frac{1}{2}\log|\Sigma_t^{-1}K_{x^t}| + \frac{1}{2}\mathrm{tr}(\Sigma_t^{-1}K_{x^t})^{-1} + \frac{1}{2}\mu_t^\intercal K_{x^t}\mu_t - t/2.$$

Note that our bound holds simultaneously for the entire sequence of posteriors (ρ_t). As was mentioned following the statement of Theorem 4, this showcases how one would typically make full use our anytime valid bounds. Plugging this into Corollary 25 gives the following result.

Corollary 42 (Anytime PAC-Bayes bound for GP Classification) *Consider the classification setting described above with GP prior ν and posterior ρ. With probability at least $1-\delta$ over (X_t, Y_t), for all times $t \geqslant 1$,*

$$\mathsf{kl}(\mathbb{E}_{\rho_t}\widehat{R}_t(\theta)\|\mathbb{E}_{\rho_t}R(\theta)) \leqslant \frac{\log|\Sigma_t^{-1}K_{x^t}| + \mathrm{tr}(\Sigma_t^{-1}K_{x^t})^{-1} + \mu_t^\intercal K_{x^t}\mu_t - t}{2\bar{t}} + \frac{\log(\xi(\bar{t})/\delta) + \mathsf{IL}_t}{\bar{t}},$$

where the empirical risk and expected risk are as in (56) *and* (57), *$\bar{t} = 2^{\lfloor\log_2(t)\rfloor}$ and $\mathsf{IL}_t < 2\log\log 2t + 1.3$.*

We recall from Section 5 that $\bar{t}$ and IL_t capture the "lag" in our anytime bounds. While we've chosen $\bar{t}$ such that $t/2 \leqslant \bar{t} \leqslant t$ for convenience, we may choose it to lie in $[t/s, t]$ for any $s > 1$, though we pay price in the IL_t term. See Remark 23 for details.

Seeger's fixed time bound for the same problem reads: For any fixed n, with probability at least $1-\delta$,

$$\mathsf{kl}(\mathbb{E}_{\rho_n}\widehat{R}_n(\theta)\|\mathbb{E}_{\rho_n}R(\theta)) \leqslant \frac{\log|\Sigma_n^{-1}K_{x^n}| + \mathrm{tr}(\Sigma_n^{-1}K_{x^n})^{-1} + \mu_n^\intercal K_{x^n}\mu_n - n}{2n} + \frac{\log(\frac{n+1}{\delta})}{n},$$

though we note that Theorem 2 of Seeger (2002) uses a quasi-inverse of the kl function to isolate $\widehat{R}_n(\theta)$ and $R(\theta)$. Any such inversion also applies here. We refer the reader to Seeger (2002) for an extensive study on the behavior of the RHS for various GP models. All of his analyses—theoretical and empirical—apply here. Finally, let us note that we might have applied any other bound that handles bounded losses (e.g., Corollary 7). However, the kl based bound is often acknowledged as the tightest, often provably so (Biggs and Guedj, 2023; Foong et al., 2021).

7. Summary

We have demonstrated that underlying many PAC-Bayes bounds is a (typically implicit) supermartingale or reverse submartingale structure. Such structure, when coupled with the method of mixtures and Ville's inequalities, provides a general method of deriving new bounds and illuminates the connection between existing ones (Table 1). For instance, we are able to generate PAC-Bayes bounds for sub-ψ processes (Howard et al., 2020, Tables 3 and 4), a broad class of processes which itself encapsulates a large swath of existing concentration inequalities. More generally, as soon as one identifies a nonnegative supermartingale or reverse submartingale with bounded initial value, our framework supplies a PAC-Bayes bound. We hope this serves to both ease the search for future bounds and to provide a more unified view of the existing literature.

Beyond such unification, our martingale-based approach provides time-uniform bounds (i.e., valid at all stopping times), whereas the majority of previous bounds in the literature are fixed-time results. Moreover, we are able to shed many traditional distribution assumptions. Many of our bounds do not require i.i.d. data, and those based on supermartingales require no explicit assumptions (Table 2). We hope that the anytime nature of our bounds is not just a theoretical curiosity, but useful for computing generalization bounds in practice. Indeed, because they allow for adaptive stopping and continuous monitoring of data, practitioners are able to repeatedly compute the bounds as more data are used without sacrificing statistical validity. We hope, for instance, that these properties can serve efforts to generate non-vacuous generalization bounds for neural networks (Dziugaite and Roy, 2017; Biggs and Guedj, 2022; Liao et al., 2020). Instead of computing bounds off a fixed test set, our methods enable the collection of more data and the monitoring of the evolution of the bound over time.

Aside from applications to neural networks, there are additional practical benefits of the time-uniform nature of our results. Beyond those applications mentioned in the introduction, PAC-Bayes bounds have been used in bandit problems (Seldin et al., 2012; Flynn et al., 2022, 2023), policy evaluation (Fard et al., 2011; Sakhi et al., 2023), multiple testing (Blanchard and Fleuret, 2007), estimating means of random vectors (Catoni and Giulini, 2017), and domain adaptation (Germain et al., 2016). Several of these applications could benefit from our techniques. For instance, Seldin et al. (2012) rely on a union bound to cover all time steps which might be circumvented by our analysis, leading to tighter bounds for contextual bandits. Similarly, our techniques applied to those of Blanchard and Fleuret (2007) may lead to better bounds for sequential multiple testing procedures (bandit multiple testing for instance, see Jamieson and Jain, 2018 and Xu et al., 2021), and anytime-valid off-policy evaluation, in which there has been recent interest (Waudby-Smith et al., 2023). Overall, we hope our work contributes to the increasing interest in applying PAC-Bayes ideas to interactive learning settings, which have a wider scope for applications.

Acknowledgments

We graciously thank Felix Biggs who pointed out a flaw in the first version of the paper. We also thank the anonymous referees for valuable feedback which improved the paper. BC was

partially supported by the Natural Sciences and Engineering Research Council of Canada. The authors acknowledge support from NSF grants IIS-2229881 and DMS-2310718

Appendix A. Omitted Proofs

A.1 Proof of Corollary 7

Let $\psi(\lambda) = \lambda^2/2$. Define the process $P(\theta) = (P_t(\theta))_{t\geqslant 1}$ as $P_t(\theta) = \sum_{i=1}^t \lambda_i \Delta_i(\theta) - \sum_{i=1}^t \psi(\lambda_i)\sigma_i^2$. We claim that $\exp(P(\theta))$ is a supermartingale. Since λ_i and $f_i(Z_i,\theta)$ are $\mathcal{F}_{t-1}$ measurable for all $i \leqslant t-1$, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\exp(P_t(\theta))|\mathcal{F}_{t-1}] &= \mathbb{E}_{\mathcal{D}}\left[\prod_{i=1}^t \exp(\lambda_i\Delta_i(\theta) - \psi(\lambda_i)\sigma_i^2)\middle|\mathcal{F}_{t-1}\right] \\
&= \mathbb{E}_{\mathcal{D}}[\exp(\lambda_t\Delta_t(\theta) - \psi(\lambda_t)\sigma_t^2)|\mathcal{F}_{t-1}]\prod_{i=1}^{t-1}\exp(\lambda_i\Delta_i(\theta) - \psi(\lambda_i)\sigma_i^2) \\
&= \mathbb{E}_{\mathcal{D}}[\exp(\lambda_t\Delta_t(\theta) - \psi(\lambda_t)\sigma_t^2)|\mathcal{F}_{t-1}]\exp(P_{t-1}(\theta)),
\end{aligned}$$

Now, the final line is at most $\exp(P_{t-1}(\theta))$ due to Hoeffding's lemma:

$$\mathbb{E}_{\mathcal{D}}[\exp(\lambda_t\Delta_t(\theta))|\mathcal{F}_{t-1}] = \mathbb{E}_{\mathcal{D}}[\exp(\lambda_t(\mu_i(\theta) - f_i(Z_i,\theta)))|\mathcal{F}_{t-1}] \leqslant \exp(\lambda_t^2\sigma_t^2/8),$$

for all $\lambda_t \in \mathbb{R}$. This proves that $\exp(P_t(\theta))$ is a supermartingale, and also that $\mathbb{E}_{\mathcal{D}}[\exp P_1(\theta)|\mathcal{F}_0] \leqslant 1$. Consequently, we may apply Corollary 6 to obtain that with probability at least $1-\delta$,

$$\sum_{i=1}^t \lambda_i \mathbb{E}_\rho \Delta_i(\theta) - \sum_{i=1}^t \psi(\lambda_i)\sigma_i^2 \leqslant D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta),$$

for all $\rho \in \mathcal{M}(\Theta)$. The result follows from rearranging.

A.2 Proof of Corollary 9

Let $g(\lambda; a, b^2)$ be the density of a Gaussian with mean a and variance b^2. We are interested in the mixing distribution F with $\mathrm{d}F(\lambda) = g(\lambda; 0, \gamma^2)\mathrm{d}\lambda$ for some fixed γ. Before proving the PAC-Bayes bound, we prove the following lemma. Let $D_t = \sum_{i=1}^t \Delta_i(\theta)$ and $H_t = \sum_{i=1}^t \sigma_i^2$.

Lemma 43 *For*

$$M_t(\lambda,\theta) := \exp\left\{\lambda\sum_{i=1}^t \Delta_i(\theta) - \frac{\lambda^2}{2}\sum_{i=1}^t \sigma_i^2\right\},$$

we have

$$M_t(\theta) = \int_{\lambda\in\mathbb{R}} M_t(\lambda,\theta)\mathrm{d}F(\lambda) = \frac{1}{\sqrt{1+\gamma^2 H_t}}\exp\left(\frac{\gamma^2 D_t^2}{1+\gamma^2 H_t}\right).$$

Proof Compute

$$\begin{aligned}
M_t(\theta) &= \frac{1}{\gamma\sqrt{2\pi}} \int_{\lambda\in\mathbb{R}} \exp\left(\lambda D_t - \frac{\lambda^2 H_t}{2}\right) \exp\left(-\frac{\lambda^2}{2\gamma^2}\right) \mathrm{d}\lambda \\
&= \frac{1}{\gamma\sqrt{2\pi}} \int_{\lambda\in\mathbb{R}} \exp\left(\frac{2\lambda\gamma^2 D_t - \lambda^2\gamma^2 H_t - \lambda^2}{2\gamma^2}\right) \mathrm{d}\lambda \\
&= \frac{1}{\gamma\sqrt{2\pi}} \int_{\lambda\in\mathbb{R}} \exp\left(\frac{-\lambda^2(1+\gamma^2 H_t) + 2\lambda\gamma^2 D_t}{2\gamma^2}\right) \mathrm{d}\lambda.
\end{aligned}$$

Define $u = 1 + \gamma^2 H_t$ and $v = \gamma^2 D_t$. Now rewrite the above expression as

$$\begin{aligned}
M_t(\theta) &= \frac{1}{\gamma\sqrt{2\pi}} \int_{\lambda\in\mathbb{R}} \exp\left(\frac{-u(\lambda^2 - 2\lambda v/u)}{2\gamma^2}\right) \mathrm{d}\lambda \\
&= \frac{1}{\gamma\sqrt{2\pi}} \int_{\lambda\in\mathbb{R}} \exp\left(\frac{-(\lambda - v/u)^2 + (v/u)^2}{2\gamma^2/u}\right) \mathrm{d}\lambda \\
&= \frac{1}{\gamma\sqrt{2\pi}} \int_{\lambda\in\mathbb{R}} \exp\left(\frac{-(\lambda - v/u)^2}{2\gamma^2/u}\right) \mathrm{d}\lambda \exp\left(\frac{v^2}{2u\gamma^2}\right) \\
&= \frac{\sqrt{1/u}}{\sqrt{1\gamma^2/u}\sqrt{2\pi}} \int_{\lambda\in\mathbb{R}} \exp\left(\frac{-(\lambda - v/u)^2}{2\gamma^2/u}\right) \mathrm{d}\lambda \exp\left(\frac{v^2}{2u\gamma^2}\right) \\
&= \sqrt{1/u} \exp\left(\frac{v^2}{2u\gamma^2}\right),
\end{aligned}$$

where the final equality follows because

$$\frac{1}{\sqrt{\gamma^2/u}\sqrt{2\pi}} \int_{\lambda\in\mathbb{R}} \exp\left(\frac{-(\lambda - v/u)^2}{2\gamma^2/u}\right) \mathrm{d}\lambda = \int_{\lambda\in\mathbb{R}} g(\lambda; v/u, 2\gamma^2/u) \mathrm{d}\lambda = 1,$$

where $g(\lambda; v/u, \gamma^2/u)$ is the density of a Gaussian with mean v/u and variance $2\gamma^2/u$. Thus, we have obtained that

$$M_t(\theta) = \frac{1}{\sqrt{u}} \exp\left(\frac{v^2}{2u\gamma^2}\right) = \frac{1}{\sqrt{1+\gamma^2 H_t}} \exp\left(\frac{\gamma^2 D_t^2}{1+\gamma^2 H_t}\right).$$

This completes the proof of the lemma. ■

From here, in order to apply Corollary 6, write this as

$$\begin{aligned}
M_t(\theta) &= \frac{1}{\sqrt{1+\gamma^2 H_t}} \exp\left(\frac{\gamma^2 D_t^2}{1+\gamma^2 H_t}\right) \\
&= \exp\left(\frac{\gamma^2 D_t^2}{1+\gamma^2 H_t} + \log([\sqrt{1+\gamma^2 H_t}]^{-1})\right) \\
&= \exp\left(\frac{\gamma^2 D_t^2}{1+\gamma^2 H_t} - \frac{1}{2}\log(1+\gamma^2 H_t)\right).
\end{aligned}$$

Corollary 6 yields that with probability at least $1-\delta$, for all t and ρ,

$$\mathbb{E}_\rho\left[\frac{\gamma^2 D_t^2}{1+\gamma^2 H_t}\right] \leqslant \frac{1}{2}\log(1+\gamma^2 H_t) + D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta).$$

Rearranging and taking square roots gives

$$\begin{aligned}
\mathbb{E}_\rho[D_t] &\leqslant \left[(\gamma^{-2}+H_t)\log(1+\gamma^2 H_t) + (\gamma^{-2}+H_t)\left(D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta)\right)\right]^{1/2} \\
&= \left[(\gamma^{-2}+H_t)\left(D_{\mathrm{KL}}(\rho\|\nu) + \log((1+\gamma^2 H_t)/\delta)\right)\right]^{1/2} \\
&= \left[\frac{s_t(\beta)}{\beta}\left(D_{\mathrm{KL}}(\rho\|\nu) + \log(s_t(\beta))/\delta)\right)\right]^{1/2},
\end{aligned}$$

where we've taken $\beta = \gamma^2$ and recalled that $s_t(c) = 1 + cH_t$. Expanding the definition of D_t completes the proof.

A.3 Proof of Corollary 12

Set

$$\xi_t(\theta) := \lambda_t \Delta_t(\theta) - \lambda_t^2 (e-2)\mathbb{E}[\Delta_t^2(\theta)|\mathcal{F}_{t-1}],$$

for all $t \geqslant 1$. First we claim that the process in Equation (15), i.e., $B_t(\theta) = \prod_{i=1}^t \exp \xi_i(\theta)$, is a nonnegative supermartingale. To see this, we recall the inequality

$$e^x \leqslant 1 + x + (e-2)x^2, \tag{58}$$

for all $x \leqslant 1$. Since $\lambda_t \leqslant |\frac{1}{2H}|$ by assumption, we have

$$|\lambda_t \Delta_t(\theta)| \leqslant \lambda_t(|\mu_t(\theta)| + |f_t(Z_t,\theta)|) \leqslant \lambda_t 2H \leqslant 1,$$

so we may apply (58) with $x = \lambda_t \Delta_t(\theta)$. This gives

$$\begin{aligned}
\mathbb{E}[\exp(\lambda_t \Delta_t(\theta)|\mathcal{F}_{t-1}] &\leqslant 1 + \lambda_t \mathbb{E}[\Delta_t(\theta)|\mathcal{F}_{t-1}] + \lambda^2 (e-2)\mathbb{E}[\Delta_t^2(\theta)|\mathcal{F}_{t-1}] \\
&= 1 + \lambda_t^2 (e-2)\mathbb{E}[\Delta_t^2(\theta)|\mathcal{F}_{t-1}] \\
&\leqslant \exp(\lambda_t^2 (e-2)\mathbb{E}[\Delta_t^2(\theta)|\mathcal{F}_{t-1}]),
\end{aligned}$$

where the equality in the second line follows by definition of $\Delta_t(\theta)$. Hence,

$$\mathbb{E}[\exp(\xi_t(\theta))|\mathcal{F}_{t-1}] = \mathbb{E}[\exp(\lambda_t \Delta_t(\theta) - \lambda_t^2 (e-2)\mathbb{E}[\Delta_t^2(\theta)|\mathcal{F}_{t-1}])] \leqslant 1.$$

It follows that $(B_t(\theta))$ is a nonnegative supermartingale and the result is then obtained by applying Theorem 4.

A.4 Proof of Corollary 13

Recall that $\psi_P(x) = e^x - x - 1$. Let $v_i^2(\theta) = \mathbb{E}_{\mathcal{D}}[f_i^2(Z_i, \theta)|\mathcal{F}_{t-1}]$. Consider the nonnegative process

$$S_t(\theta) = \prod_{i=1}^{t} \exp\left\{\lambda_i(f_i(Z_i,\theta) - \mu_i(\theta)) - \frac{v_i^2(\theta)}{H_i^2}\psi_P(\lambda_i H_i)\right\}.$$

The function $x^{-2}\psi_P(x)$ is nondecreasing (at $x = 0$ we continuously extend the function to $1/2$ following the proof of Corollary 17). Since f_i is bounded by H_i, we have

$$\frac{1}{(\lambda_i f_i(Z_i,\theta))^2}\psi_P(\lambda_i f_i(Z_i,\theta)) \leqslant \frac{1}{(\lambda_i H_i)^2}\psi_P(\lambda_i H_i),$$

that is,

$$e^{\lambda_i f_i(Z_i,\theta)} \leqslant \frac{f_i^2(Z_i,\theta)}{H_i^2}\psi_P(\lambda_i H_i) + \lambda_i f_i(Z_i,\theta) + 1.$$

Taking expectations,

$$\mathbb{E}[e^{\lambda_i f_i(Z_i,\theta)}|\mathcal{F}_{t-1}] \leqslant \frac{v_i^2(\theta)}{H_i^2}\psi_P(\lambda_i H_i) + \lambda_i\mu_i(\theta) + 1.$$

Note that $\psi_P(x) \geqslant 0$ for all x, so

$$\frac{v_i^2(\theta)}{H_i^2}\psi_P(\lambda_i H_i) + \lambda_i\mu_i(\theta) \geqslant \lambda_i\mu_i(\theta) \geqslant \lambda_i v_i(\theta) > -1,$$

since $\lambda_i < 1/v_i(\theta)$ by assumption. Note that we've used $\mu_i^2(\theta) \leqslant v_i^2(\theta)$ (by Jensen) so $|\mu_i(\theta)| \leqslant v_i(\theta)$. Therefore, we may take the logarithm of the above and applying the inequality $\log(1+x) \leqslant x$ for $x > -1$ to obtain

$$\log \mathbb{E}[e^{\lambda_i f_i(Z_i,\theta)}|\mathcal{F}_{t-1}] \leqslant \log\left(\frac{v_i^2(\theta)}{H_i^2}\psi_P(\lambda_i H_i) + \lambda_i\mu_i(\theta) + 1\right) \leqslant \frac{v_i^2(\theta)}{H_i^2}\psi_P(\lambda_i H_i) + \lambda_i\mu_i(\theta).$$

Adding $-\lambda_i\mu_i(\theta) = \log e^{-\lambda_i\mu_i(\theta)}$ to each side gives

$$\log \mathbb{E}[e^{\lambda_i \Delta_i(\theta)}|\mathcal{F}_{t-1}] \leqslant \frac{v_i^2(\theta)}{H_i^2}\psi_P(\lambda_i H_i).$$

Exponentiating and rearranging implies that

$$\exp\left\{\lambda_i(f_i(Z_i,\theta) - \mu_i(\theta)) - \frac{v_i^2(\theta)}{H_i^2}\psi_P(\lambda_i H_i)\right\} \leqslant 1,$$

thus implying that $(S_t(\theta))$ is a supermartingale, and the result thus follows from applying Theorem 4.

A.5 Proof of Corollary 14

Due to the constraints on (λ_t) and (c_t), (16) implies that the process defined by

$$M_t(\theta) = \prod_{i=1}^{t} \exp\left\{\lambda_i(\mu_i(\theta) - f_i(Z_i,\theta) - c_i f_i^2(Z_i,\theta))\right\},$$

is a nonnegative supermartingale. Proceed as usual and apply Theorem 4.

A.6 Proof of Corollary 11

Recall our assumption: $\mathbb{E}[(f_i(Z_i,\theta)-\mu_i(\theta))^k] \leqslant \frac{1}{2}k!\sigma_i^2 c_i^{k-2}$. By Wainwright (2019, Proposition 2.10), this implies that

$$\mathbb{E}[\exp(\lambda(\mu_i(\theta)-f_i(Z_i,\theta)))|\mathcal{F}_{i-1}] \leqslant \exp\left(\frac{\lambda^2\sigma_i^2}{2(1-c_i|\lambda|)}\right), \tag{59}$$

whenever $|\lambda| < 1/c_t$. Consider the quantity $N_t(\theta) = \prod_{i=1}^t \exp\left\{\lambda_i\Delta_i(\theta) - \frac{\lambda_i^2\sigma_i^2}{2(1-c_i\lambda_i)}\right\}$. Similarly to the proof in Appendix A.1, $(N_t(\theta))$ is a supermartingale by appealing to (59), since $0<\lambda_i<1/c_i$ by assumption. From here we apply Theorem 4.

A.7 Proof of Corollary 16

Consider $W_t(\theta) = \lambda_t f_t(Z_t,\theta) - \log \mathbb{E}_{\mathcal{D}} \exp(\lambda_t f_t(Z,\theta))$. Observe that the conditional expectation of $W_t(\theta)$ is precisely 1:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\exp(W_t(\theta))|\mathcal{F}_{t-1}] &= \mathbb{E}_{\mathcal{D}}[\exp(\lambda_t f_t(Z_t,\theta) - \log \mathbb{E}_{\mathcal{D}} \exp(\lambda_t f_t(Z,\theta))|\mathcal{F}_{t-1}] \\
&= \mathbb{E}_{\mathcal{D}}[\exp(\lambda_t f_t(Z_t,\theta)\cdot[\mathbb{E}_{\mathcal{D}} \exp(\lambda_t f_t(Z,\theta))]^{-1}|\mathcal{F}_{t-1}] \\
&= [\mathbb{E}_{\mathcal{D}} \exp(\lambda_t f_t(Z,\theta))]^{-1}\mathbb{E}_{\mathcal{D}}[\exp(\lambda_t f_t(Z_t,\theta)|\mathcal{F}_{t-1}] = 1.
\end{aligned}$$

Therefore, $\mathbb{E}[\sum_{i\leqslant t} W_i(\theta)|\mathcal{F}_{t-1}] = \mathbb{E}[W_t(\theta)|\mathcal{F}_{t-1}]\sum_{i=1}^t W_i(\theta) = \sum_{i=1}^{t-1} W_i(\theta)$, so the process $(\sum_{i\leqslant t} W_i(\theta))_t$ is a nonnegative supermartingale. Applying Theorem 4 we obtain that, with probability at least $1-\delta$, for all t and $\rho \in \mathcal{M}(\Theta)$,

$$\mathbb{E}_{\theta\sim\rho}\sum_{i=1}^t \lambda_i f_i(Z_i,\theta) \leqslant \mathbb{E}_{\theta\sim\rho}\sum_{i=1}^t \log \mathbb{E}_{\mathcal{D}} \exp(\lambda_i f_i(Z,\theta)) + D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta).$$

Using the concavity of the logarithm then completes the argument.

A.8 Proof of Corollary 17

First we prove a self-contained result concerning the relevant supermartingale. From here, the result follows immediately from an application of Theorem 4.

Lemma 44 *Let (X_t) be nonnegative random variables where X_i has conditional mean $\mathbb{E}_{i-1}[X_i] = \mathbb{E}[X_i|\mathcal{F}_{i-1}]$ and conditional variance $\mathsf{Var}_{i-1}(X_i) = \mathsf{Var}(X_i|\mathcal{F}_{i-1}) < \infty$. For any predictable sequence of positive real numbers $\{\lambda_i\}$, the following process is a nonnegative supermartingale:*

$$L_t := \prod_{i=1}^t \exp\left\{\lambda_i(\mathbb{E}_{i-1}[X_i]-X_i) - \frac{\lambda_i^2}{2}\mathbb{E}_{i-1}[X_i^2]\right\}.$$

Proof Since L_{t-1} is $\mathcal{F}_{t-1}$ measurable, we obtain

$$\mathbb{E}[L_t|\mathcal{F}_{t-1}] = L_{t-1}\cdot\exp\left\{\lambda_t(\mathbb{E}_{t-1}[X_t]-X_t) - \frac{\lambda_t^2}{2}\mathbb{E}_{t-1}[X_t^2]\middle|\mathcal{F}_{t-1}\right\}.$$

Since λ_t is predictable, in order to show the above term is bounded by L_{t-1} it suffices to show that for any nonnegative random variable X with finite mean μ and second moment we have

$$\mathbb{E}[\exp(\lambda(\mu - X))] \leqslant \exp(\lambda^2 \mathbb{E}[X^2]/2),$$

for all $\lambda > 0$. This fact follows from applying a one-sided Bernstein inequality to $-X$, but we supply the proof for completeness. Let $Z = -X$ and put $\psi(s) = e^s - s - 1$. Let

$$g(s) = \begin{cases} \psi(s)/s^2, & s \neq 0, \\ 1/2, & s = 0. \end{cases}$$

Note that $g(s)$ simply defines the continuous extension of $\psi(s)/s^2$ at $s = 0$. Indeed, $\lim_{s\to 0^+} \psi(s)/s^2 = \lim_{s\to 0^-} \psi(s)/s^2 = 1/2$. Note also $g(s)$ is an increasing function for all $s \in \mathbb{R}$. Therefore, for all $s \leqslant 0$, $\psi(s) = s^2 g(s) \leqslant s^2 g(0) = \frac{s^2}{2}$. Since $Z \leqslant 0$ and $\lambda > 0$, we may take $s = \lambda Z$ to obtain $\phi(\lambda Z) \leqslant (\lambda Z)^2/2$. Thus, $\mathbb{E}[e^{\lambda Z}] - \lambda \mathbb{E}[Z] - 1 \leqslant \frac{\lambda^2}{2}\mathbb{E}[Z^2]$, and

$$\begin{aligned} \mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] &\leqslant e^{-\lambda \mathbb{E}[Z]}(1 + \lambda \mathbb{E}[Z] + \lambda^2 \mathbb{E}[Z^2]/2) \\ &\leqslant e^{-\lambda \mathbb{E}[Z]} \exp(\lambda \mathbb{E}[Z] + \lambda^2 \mathbb{E}[Z^2]/2) = \exp(\lambda^2 \mathbb{E}[Z^2]/2). \end{aligned}$$

Replacing Z with $-X$ completes the proof. ■

A.9 Proof of Corollary 18

Let (λ_i) be a predictable sequence. (Delyon, 2009, Proposition 12) shows that for all $x \in \mathbb{R}$, $\exp(x - x^2/6) \leqslant 1 + x + x^2/3$. Applying this with $x = \lambda_t \Delta_t(\theta)$ and taking expectations, we obtain that

$$\begin{aligned} \mathbb{E}[\exp\left\{\lambda_t \Delta_t(\theta) - \lambda_t^2 \Delta_t^2(\theta)/6\right\} | \mathcal{F}_{t-1}] &\leqslant 1 + \mathbb{E}[\lambda_t \Delta_t(\theta) | \mathcal{F}_{t-1}] + \mathbb{E}[\lambda_t^2 \Delta_t^2(\theta)/3 | \mathcal{F}_{t-1}] \\ &= 1 + \mathbb{E}[\lambda_t^2 \Delta_t^2(\theta)/3 | \mathcal{F}_{t-1}] \\ &\leqslant \exp\left\{\mathbb{E}[\lambda_t^2 \Delta_t^2(\theta)/3 | \mathcal{F}_{t-1}]\right\}, \end{aligned}$$

where the equality in the second line follows since $\Delta_t(\theta)$ is mean zero. Therefore,

$$\mathbb{E}\left[\exp\left\{ \lambda_t \Delta_t(\theta) - \frac{\lambda_t^2}{6}(\Delta_t^2(\theta) + 2\mathbb{E}[\Delta_t^2(\theta)|\mathcal{F}_{t-1}]) \right\} \middle| \mathcal{F}_{t-1} \right] \leqslant 1,$$

and we conclude that

$$M_t(\theta) = \exp\left\{ \sum_{i \leqslant t} \lambda_i \Delta_i(\theta) - \frac{1}{6} \sum_{i \leqslant t} \lambda_i^2 (\Delta_i^2(\theta) + 2\mathbb{E}[\Delta_i^2(\theta)|\mathcal{F}_{i-1}]) \right\},$$

is a nonnegative supermartingale with initial value $\mathbb{E}[M_1(\theta)] \leqslant 1$. Applying Theorem 4 gives that with probability at least $1 - \delta$, for all t and ρ,

$$\sum_{i \leqslant t} \lambda_i \mathbb{E}_\rho \Delta_i(\theta) \leqslant \frac{1}{6} \sum_{i \leqslant t} \left(\lambda_i^2 \mathbb{E}_\rho[(\Delta_i^2(\theta) + 2\mathbb{E}[\Delta_i^2(\theta)|\mathcal{F}_{i-1}])] \right) + \log(1/\delta) + D_{\mathrm{KL}}(\rho \| \nu).$$

This proves the first part of the result. From here, we can simplify the bound by observing that

$$\begin{aligned}
&\sum_{i\leqslant t}\Delta_i^2(\theta) + 2\sum_{i\leqslant t}\mathbb{E}[\Delta_i^2(\theta)|\mathcal{F}_{i-1}] \\
&= \sum_{i=1}^{t}(\mu_i(\theta) - f_i(Z_i,\theta))^2 + 2\sum_{i=1}^{t}\mathbb{E}[(\mu_i(\theta) - f_i(Z,\theta))^2|\mathcal{F}_{i-1}] \\
&= \sum_{i=1}^{t}\left\{f_i^2(Z_i,\theta) - 2\mu_i(\theta)f_i(Z_i,\theta) + 2\mathbb{E}[f_i^2(Z,\theta)|\mathcal{F}_{i-1}] - \mu_i^2(\theta)\right\} \\
&\leqslant \sum_{i=1}^{t}[f_i^2(Z_i,\theta) + 2\mathbb{E}_{\mathcal{D}}[f_i^2(Z,\theta)|\mathcal{F}_{i-1}]],
\end{aligned}$$

where we've used that the loss is nonnegative (therefore so is $\mu_i(\theta)$). This gives that with probability at least $1-\delta$, for all t and ρ,

$$\sum_{i\leqslant t}\lambda_i\mathbb{E}_\rho\Delta_i(\theta) \leqslant \frac{1}{6}\sum_{i\leqslant t}\lambda_i^2\mathbb{E}_\rho\left(f_i(Z_i,\theta) + 2\mathbb{E}_{\mathcal{D}}[f_i^2(Z,\theta)|\mathcal{F}_{i-1}]\right) + \log(1/\delta) + D_{\mathrm{KL}}(\rho\|\nu).$$

If we take $f = f_i$ and $\lambda = \lambda_i$ as constants and divide both sides by t we obtain (25).

A.10 Proof of Corollary 25

For $Z_1, \ldots, Z_n$ i.i.d, Maurer (2004, Theorem 1) proved the inequality,

$$\mathbb{E}_{(Z_t)\sim\mathcal{D}}\exp\left\{n\,\mathsf{kl}(\widehat{R}_n(\theta)\|R(\theta))\right\} \leqslant \mathbb{E}_{B\sim\mathrm{Bin}(n,R(\theta))}\exp\left\{n\,\mathsf{kl}(B/n\|R(\theta))\right\},$$

where Bin denotes the binomial distribution. Following Germain et al. (2015), the latter quantity is equal to $\xi(n)$. Indeed,

$$\begin{aligned}
&\mathbb{E}_{B\sim\mathrm{Bin}(n,R(\theta))}\exp\left(n\,\mathsf{kl}\left(\frac{B}{n}\middle\|R(\theta)\right)\right) \\
&= \mathbb{E}_{B\sim\mathrm{Bin}(n,R(\theta))}\left(\frac{B/n}{R(\theta)}\right)^B\left(\frac{1-B/n}{1-R(\theta)}\right)^{n-B} \\
&= \sum_{k=0}^{n}\mathbb{P}(B=k)\left(\frac{k/n}{R(\theta)}\right)^k\left(\frac{1-k/n}{1-R(\theta)}\right)^{n-k} \\
&= \sum_{k=0}^{n}\binom{n}{k}R(\theta)^k(1-R(\theta))^{n-k}\left(\frac{k/n}{R(\theta)}\right)^k\left(\frac{1-k/n}{1-R(\theta)}\right)^{n-k} \\
&= \sum_{k=0}^{n}\binom{n}{k}(k/n)^k(1-k/n)^{n-k} = \xi(n).
\end{aligned}$$

Therefore, applying Corollary 22 with $\varphi = \mathsf{kl}$ and $\lambda_{\bar{t}} = \bar{t}$ gives

$$\begin{aligned}
\mathbb{E}_\rho\varphi_t(\theta) &\leqslant \frac{\log\mathbb{E}_{\rho,\mathcal{D}}\exp(\bar{t}\varphi_{\bar{t}}(\theta))}{\bar{t}} + \frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(1/\delta) + \mathsf{IL}_t}{\bar{t}} \\
&\leqslant \frac{D_{\mathrm{KL}}(\rho\|\nu) + \log(\xi(\bar{t})/\delta) + \mathsf{IL}_t}{\bar{t}},
\end{aligned}$$

as desired. Finally, (36) follows from similar arguments and applying Corollary 24.

A.11 Proof of Corollary 27

Fix $\lambda > 0$. Put

$$M_t^j(\theta) = \exp\left\{ j\lambda_j(\mathsf{Var}(\theta) - \mathsf{Var}_t(\theta)) - \frac{\lambda_j^2}{2}\frac{j^2}{j-1}\mathsf{Var}(\theta) \right\}.$$

Note that $\mathsf{Var}_t(\theta) = U_t(\theta)$, i.e., it is the U-statistic for the functional $\Phi(P,\theta) = \mathsf{Var}_P(f(P,\theta))$. Jensen's inequality combined with the fact that $U_t(\theta)$ is a reverse martingale with respect to $(\mathcal{E}_t)$ implies that $(M_t^j(\theta))_{t\geqslant j}$ is a reverse submartingale with respect to $(\mathcal{E}_t)$. (Note that everything else inside the exponential is constant with respect to t.) Moreover, we note that $\mathbb{E}_P[M_j^j(\theta)] \leqslant 1$ due to the self-bounding property of $U_t(\theta)$ (Tolstikhin and Seldin, 2013, Equation (9)). From here, the argument resembles that of Corollary 22. Theorem 4 implies that

$$\mathbb{P}(\exists t \geqslant j : \mathbb{E}_\rho M_t^j(\theta) - D_{\mathrm{KL}}(\rho\|\nu) \geqslant \log(u/\delta)) \leqslant \delta/u.$$

We then apply a union bound over the events $\{\exists t \geqslant 2^k : \mathbb{E}_\rho M_t^{2^k}(\theta) - D_{\mathrm{KL}}(\rho\|\nu) \geqslant \log(\ell(k+1)/\delta)\}$ implying that

$$\mathbb{P}(\exists t \geqslant 1 : \mathbb{E}_\rho M_t^{\bar{t}}(\theta) - D_{\mathrm{KL}}(\rho\|\nu) \geqslant \log(\ell(\log_2(2t)/\delta))) \leqslant \delta,$$

completing the proof.

A.12 Proof of Corollary 30

The proof follows that of Corollary 22 very closely, so we provide only the outline. Define

$$h_t^j(\theta) = \lambda_j\varphi_t(\theta) - \log \mathbb{E}_{\nu,\mathcal{D}} \exp(\lambda_j\varphi_j(\theta)).$$

Then $(\exp h_t^j(\theta))$ is a reverse submartingale with respect to $(\mathcal{E}_t)$ obeying $\mathbb{E}_{\mathcal{D}}[\exp h_j^j(\theta)] = 1$. Theorem 29 along with our assumption implies that

$$\mathbb{P}(\exists t \geqslant 2^k : \mathbb{E}_\rho h_t^{2^k}(\theta) - \gamma_{\mathcal{G}_t}(\rho,\nu) \geqslant \log(u/\delta)) \leqslant \delta/u.$$

The event $\{\exists t \geqslant 1 : \mathbb{E}_\rho h_t^{\bar{t}}(\theta) - \gamma_{\mathcal{G}_t}(\rho,\nu) \geqslant \log(\ell(\log_2(2t))/\delta)\}$ is contained in the event $\bigcup_{k=0}^\infty \{\exists t \geqslant 2^k : \mathbb{E}_\rho h_t^{2^k} - \gamma_{\mathcal{G}_t}(\rho) \geqslant \log(\ell(k+1)/\delta)\}$, where ℓ is the stitching function introduced in Section 5.1. The union bound over all such events implies that

$$\mathbb{P}(\exists t \geqslant 1 : \mathbb{E}_\rho h_t^{\bar{t}}(\theta) - \gamma_{\mathcal{G}_t}(\rho,\nu) \geqslant \log(\ell(\log_2(2t))/\delta)) \leqslant \delta.$$

This proves the first part the result. The second part comes from applying Theorem 29 to the process $(h_t^n(\theta))$ with $t_0 = n$.

A.13 Proof of Corollary 34

Let $\alpha_0 = \alpha/(\alpha-1)$. Define the quantity

$$S_t^j(\theta) = \log\varphi_t(\theta) - \frac{1}{\alpha_0}\log \mathbb{E}_{\vartheta\sim\nu,\mathcal{D}}[\varphi_j^{\alpha_0}(\vartheta)].$$

Note that the final term is not a function of θ. First, we claim that $(\exp S_t^j(\theta))_{t\geqslant 1}$ is a reverse submartingale with respect to $(\mathcal{E}_t)$. Recalling that $\varphi_t(\theta)$ is reverse submartingale with respect to the same filtration, we have

$$\mathbb{E}_{\mathcal{D}}[\exp S_t^j(\theta)|\mathcal{E}_{t+1}] = \frac{\mathbb{E}_{\mathcal{D}}[\varphi_t(\theta)|\mathcal{E}_{t+1}]}{\mathbb{E}_{\nu,\mathcal{D}}[\varphi_j^{\alpha_0}(\vartheta)]^{\frac{1}{\alpha_0}}} \geqslant \frac{\varphi_{t+1}(\theta)}{\mathbb{E}_{\nu,\mathcal{D}}[\varphi_j^{\alpha_0}(\vartheta)]^{\frac{1}{\alpha_0}}} = \exp S_{t+1}^j(\theta).$$

Therefore, it follows from (50) that $([\exp S_t^j(\theta)]^{\alpha_0})_{t\geqslant 1}$ is a reverse submartingale with respect to $(\mathcal{E}_t)$. Next we observe that, by construction, $\mathbb{E}_{\mathcal{D}}[\exp S_j^j(\theta)^{\alpha_0}] = 1$. Therefore, by Theorem 33, for all ρ,

$$\mathbb{P}(\exists t \geqslant j : \mathbb{E}_\rho S_t^j(\theta) \geqslant \alpha_0^{-1}(D_\alpha(\rho\|\nu) + \log(u/\delta))) \leqslant \delta/u,$$

for $u > 0$. Let $\ell(k) = k^2\zeta(2)$ be the stitching function introduced in Section 5.1. Following the proof of Corollary 22, we claim that

$$\begin{aligned}&\left\{\exists t \geqslant 1 : \mathbb{E}_\rho S_t^{\bar{t}}(\theta) \geqslant \alpha_0^{-1}(D_\alpha(\rho\|\nu) + \log(\ell(\log_2(2t))/\delta))\right\}\\ \subseteq &\bigcup_{k=0}^{\infty}\left\{\exists t \geqslant 2^k : \mathbb{E}_\rho S_t^{2^k}(\theta) \geqslant \alpha_0^{-1}(D_\alpha(\rho\|\nu) + \log(\ell(k+1)/\delta))\right\}.\end{aligned}$$

The argument is identical to before: if there is some t^* such that the first event holds, then $n(t^*) = 2^{k^*}$ for some k^* so

$$\begin{aligned}\mathbb{E}_\rho S_{t^*}^{2^{k^*}}(\theta) = \mathbb{E}_\rho S_{t^*}^{\bar{t^*}}(\theta) &\geqslant \alpha_0(D_\alpha(\rho\|\nu) + \log(\ell(\log_2(2t^*))/\delta))\\ &\geqslant \alpha_0(D_\alpha(\rho\|\nu) + \log(\ell(k^*+1))/\delta)),\end{aligned}$$

since $\log_2(2t^*) = 1 + \log_2(t^*) \geqslant 1 + \lfloor\log_2(t^*)\rfloor = 1 + k^*$. Applying the union bound, we conclude that

$$\mathbb{P}(\exists t \geqslant 1 : \mathbb{E}_\rho S_t^{\bar{t}}(\theta) \geqslant \alpha_0^{-1}(D_\alpha(\rho\|\nu) + \log(1/\delta) + \mathsf{IL}_t) \leqslant \sum_{k=1}^{\infty}\frac{\delta}{\ell(k)} = \delta.$$

That is, with probability at least $1-\delta$, for all $t \geqslant 1$ and $\rho \in \mathcal{M}(\Theta)$,

$$\mathbb{E}_\rho \log\varphi_t(\theta) - \frac{1}{\alpha_0}\log\mathbb{E}_{\vartheta\sim\nu,\mathcal{D}}[\varphi_{\bar{t}}^{\alpha_0}(\vartheta)] \leqslant \frac{1}{\alpha_0}(D_\alpha(\rho\|\nu) + \log(1/\delta) + \mathsf{IL}_t).$$

The desired result then follows by rearranging, and by noting that $\log\mathbb{E}_\rho\varphi_t(\theta) \geqslant \mathbb{E}_\rho\log\varphi_t(\theta)$.

A.14 Proof of Corollary 36

As in Section 5, we will make use of a nondecreasing function $\ell : \{0,1,2,\dots\} \to \mathbb{R}_{>0}$ such that $\sum_{k=0}^{\infty}\frac{1}{\ell(k)} \leqslant 1$. For concreteness, the reader is encouraged to keep $\ell(k) = (k+1)^2\zeta(2)$ in mind, but other options are available. We note that the domain of this function differs slightly from that in Section 5. This is a matter of convenience only.

Recall the notation $\Delta_i(\theta) = \mu_i(\theta) - f_i(Z_i, \theta)$ and set $S_t(\theta) = \sum_{i=1}^t \Delta_i(\theta)$. Let $\psi(\lambda) = \lambda^2/2$ be the ψ function for subGaussian random variables. The process $(\exp\{\lambda S_t(\theta) - \psi(\lambda)t\})_{t\geqslant 1}$ is a nonnegative supermartingale, so Corollary 6 implies that, for all $\lambda \in \mathbb{R}$,

$$\mathbb{P}\left(\exists t \geqslant 1 : \mathbb{E}_\rho S_t(\theta) \geqslant \frac{\psi(\lambda)t + D_{\mathrm{KL}}(\rho||\nu) + \log(1/\delta)}{\lambda}\right) \leqslant \delta.$$

Let $r = \log(1/\delta)$ and take $g_{\lambda,r}$ to be the lower bound on $\mathbb{E}_\rho S_t(\theta)$:

$$g_{\lambda,r}(u) = \frac{\psi(\lambda)u + D_{\mathrm{KL}}(\rho||\nu) + r}{\lambda}.$$

We can rewrite the time-uniform bound on $\mathbb{E}_\rho S_t(\theta)$ as

$$\mathbb{P}(\exists t \geqslant 1 : \mathbb{E}_\rho S_t(\theta) \geqslant g_{\lambda,r}(t)) \leqslant e^{-r}. \tag{60}$$

As in the proof of Corollary 22, we consider geometrically spaced epochs in time: $[2^k, 2^{k+1})$ for $k = 0, 1, \ldots$. We wish to employ (60) in each epoch $[2^k, 2^{k+1})$ with carefully chosen parameters r_k and λ_k and then take the union bound over all epochs to obtain our result. Following Theorem 1 of Howard et al. (2021), we select λ_k such that $g_{\lambda_k,r_k}(2^k)/2^k = g_{\lambda_k,r_k}(2^{k+1})/2^{k+1}$. This gives $\lambda_k = \psi^{-1}(r_k/2^{k+1/2}) = \sqrt{2r_k/2^{k+1/2}}$. Plugging this into g gives

$$g_{\lambda_k,r_k}(u) = \frac{\sqrt{r_k u}}{\sqrt{2}}\left(\sqrt{\frac{u}{2^{k+1/2}}} + \sqrt{\frac{2^{k+1/2}}{u}}\right) + \frac{D_{\mathrm{KL}}(\rho||\nu)\sqrt{2^{k+1/2}}}{\sqrt{2r_k}}.$$

The first term on the right hand side can be bounded by $2\sqrt{r_k u}$ by maximizing $\sqrt{\frac{u}{2^{k+1/2}}} + \sqrt{\frac{2^{k+1/2}}{u}}$ over $u \in [2^k, 2^{k+1}]$. Consider taking $r_k = \log(\ell(k)/\delta)$. Then $k \leqslant \log_2(u)$, so $r_k = \log(\ell(k)/\delta) \leqslant \log(\ell(\log_2(u)/\delta))$, implying the first term can be upper bounded as $2\sqrt{u\log(\log_2(u)/\delta)}$. For the KL divergence term, note that $2^k \leqslant u$ so $\sqrt{2^{k+1/2}} < \sqrt{2u}$. Furthermore, $k + 1 \geqslant \log_2(u)$, so $r_k = \log(\ell(k)/\delta) \geqslant \log(\ell(\log_2(u) - 1)/\delta)$. Putting this all together yields

$$g_{\lambda_k,r_k}(u) \leqslant 2\sqrt{u\log(\ell(\log_2(u))/\delta)} + D_{\mathrm{KL}}(\rho||\nu)\sqrt{\frac{u}{\log(\ell(\log_2(u) - 1)/\delta)}} = B_\delta(u).$$

That is, we have shown that for $2^k \leqslant u < 2^{k+1}$, $g_{\lambda_k,r_k}(u) \leqslant B_\delta(u)$.

Now, consider the event $\mathbb{E}_\rho S_{t^*}(\theta) > B_\delta(t^*)$. Let k^* be such that $t^* \in [2^{k^*}, 2^{k^*+1})$. Then $\mathbb{E}_\rho S_{t^*}(\theta) > g_{\lambda_{k^*},r_{k^*}}(t^*)$, implying that the event $\{\exists t \geqslant 1 : \mathbb{E}_\rho S_t(\theta) > B_\delta(t)\}$ is contained in the event $\bigcup_{k=0}^\infty \{\exists t \in [2^k, 2^{k+1}) : \mathbb{E}_\rho S_t(\theta) > g_{\lambda_k,r_k}(t)\}$. Consequently, (60) in conjunction with the union bound implies that

$$\mathbb{P}(\exists t \geqslant 1 : \mathbb{E}_\rho S_t(\theta) > B_\delta(t)) \leqslant \sum_{k=0}^\infty e^{-r_k} = \delta\sum_{k=0}^\infty \frac{1}{\ell(k)} \leqslant \delta.$$

We have thus shown that, with probability at least $1 - \delta$, for all $t \geqslant 1$,

$$\frac{1}{t}\sum_{i=1}^t \mathbb{E}_\rho \mu_i(\theta) \leqslant \frac{1}{t}\sum_{i=1}^t \mathbb{E}_\rho f_i(Z_i, \theta) + \frac{2\sqrt{\log(\ell(\log_2(t))/\delta)}}{\sqrt{t}} + \frac{D_{\mathrm{KL}}(\rho\|\nu)}{\sqrt{t\log(\ell(\log_2(t) - 1)/\delta)}}.$$

By considering $-\mathbb{E}_\rho S_t(\theta)$ and taking a union bound we conclude that

$$\begin{aligned}\frac{1}{t}|\mathbb{E}_\rho S_t(\theta)| &\leqslant \frac{2\sqrt{\log(2\ell(\log_2(t))/\delta)}}{\sqrt{t}} + \frac{D_{\mathrm{KL}}(\rho\|\nu)}{\sqrt{t\log(2\ell(\log_2(t)-1)/\delta)}} \\ &\lesssim \frac{\sqrt{\log(\log(t))+\log(1/\delta)}}{\sqrt{t}} + \frac{D_{\mathrm{KL}}(\rho\|\nu)}{\sqrt{t\log(\log(t))+t\log(1/\delta)}},\end{aligned}$$

as claimed.

A.15 LIL Bound for a Constant Mean

The following is obtained via an ingredient of stitching similar to both Howard et al. (2021, Theorem 1) and Wang and Ramdas (2023a, Corollary 10.2). The resulting width of the boundary is the same as in Corollary 36, but the argument is simpler as the mean is constant.

Corollary 45 *Let f be 1-subGaussian and let $(Z_t) \sim \mathcal{D}$ be such that $\mu(\theta) = \mathbb{E}_{\mathcal{D}}[f(Z,\theta)|\mathcal{F}_{t-1}]$ is constant for all $t \geqslant 1$. Fix a prior $\nu \in \mathcal{M}(\Theta)$. Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$, for all ρ and $t \geqslant 1$,*

$$\mathbb{E}_\rho \mu(\theta) \in \left(\frac{\sum_{i=1}^t f(Z_i,\theta)}{t} \pm W_t^{\mathsf{stch}}\right),$$

where the width W_t^{stch} is

$$2\sqrt{\frac{\log(6.3/\delta)+1.4\log\log_2 2t}{t}} + \frac{D_{\mathrm{KL}}(\rho\|\nu)}{\sqrt{(\log(6.3/\delta)+1.4\log\log_2(t+1))t}}.$$

Proof Let

$$W_t(\Lambda,\delta) = \frac{\log(2/\delta) + D_{\mathrm{KL}}(\rho\|\nu) + \frac{1}{2}t\Lambda^2}{t\Lambda} \tag{61}$$

be the width of the CS in (53) when the error level is set to δ, the sequence $\{\lambda_t\}$ is set to a constant $\Lambda > 0$, and σ is set to 1. Let $t_j = 2^j$, $\delta_j = \frac{\delta(1+j)^{-1/4}}{3.15}$, and $\Lambda_j = \sqrt{\log(2/\delta_j)2^{-j}}$. Note that $\sum_{j=0}^\infty \delta_j < \delta$. By Corollary 35, with probability at least $1-\delta_j$, for all ρ and integers $t \in [t_j, t_{j+1})$, $\mathbb{E}_\rho\mu(\theta) \in \left(\frac{\sum_{i=1}^t f(Z_i,\theta)}{t} \pm W_t(\Lambda_j,\delta_j)\right)$. Therefore, by the union bound, we have for all ρ and t,

$$\mathbb{E}_\rho \mu(\theta) \in \left(\frac{\sum_{i=1}^t f(Z_i,\theta)}{t} \pm W_t^{\mathsf{stch}*}\right), \quad \text{where} \quad W_t^{\mathsf{stch}*} := W_t(\Lambda_j,\delta_j) \text{ for } t_j \leqslant t < t_{j+1}.$$

Next, we show the straightforward fact that $W_t^{\mathsf{stch}*}$ satisfies an iterated logarithmic rate. Note that $\log(6.3/\delta) + 1.4\log\log_2(t+1) \leqslant \log(2/\delta_j) \leqslant \log(6.3/\delta) + 1.4\log\log_2 2t$, so

$$\begin{aligned}
W_t^{\mathsf{stch}*} &= \frac{\log(2/\delta_j) + D_{\mathrm{KL}}(\rho\|\nu) + \frac{1}{2}t\Lambda_j^2}{t\Lambda_j} \\
&\leqslant \frac{2\log(2/\delta_j) + D_{\mathrm{KL}}(\rho\|\nu)}{\sqrt{\log(2/\delta_j)t}} \\
&= 2\sqrt{\frac{\log(2/\delta_j)}{t}} + \frac{D_{\mathrm{KL}}(\rho\|\nu)}{\sqrt{\log(2/\delta_j)t}} \\
&\leqslant 2\sqrt{\frac{\log(6.3/\delta) + 1.4\log\log_2 2t}{t}} + \frac{D_{\mathrm{KL}}(\rho\|\nu)}{\sqrt{(\log(6.3/\delta) + 1.4\log\log_2(t+1))t}}.
\end{aligned}$$

This concludes the proof. ■

Appendix B. Mixtures of Martingales

Lemma 46 (Mixture of martingales) *Let $\{(M_t(\theta))_{t\in\mathbb{Z}} : \theta \in \Theta\}$ be a family of martingales (resp., super/submartingales) on a filtered probability space $(\Omega, \mathcal{A}, (\mathcal{F}_t)_{t\in\mathbb{Z}}, \mathbb{P})$, indexed by θ in a measurable space $(\Theta, \mathcal{B})$ such that*

(i) each $M_t(\theta)$ is $\mathcal{F}_t \otimes \mathcal{B}$-measurable; and

(ii) each $\mathbb{E}[M_t(\theta)|\mathcal{F}_{t-1}]$ is $\mathcal{F}_{t-1} \otimes \mathcal{B}$-measurable.

Let μ be a finite measure on $(\Theta, \mathcal{B})$ such that for all t,

$$\mathbb{P}\otimes\mu\text{-almost everywhere } M_t(\theta) \geqslant 0, \quad \text{or} \quad \mathbb{E}_{\theta\sim\mu}\mathbb{E}[|M_t(\theta)|] < \infty.$$

Then the mixture $(M_t^{\mathsf{mix}})_{t\in\mathbb{Z}}$, where $M_t^{\mathsf{mix}} = \mathbb{E}_{\theta\sim\mu}M_t(\theta)$, is also a martingale (or super/submartingale).

Proof First consider the case of supermartingales. Take any $A \in \mathcal{F}_{t-1}$. Employing assumptions (i) and (ii) we can apply Fubini's theorem to $M_t(\theta)$ on $\mathbb{P}|_A \otimes \mu$:

$$\mathbb{E}\left[\mathbf{1}_A \int M_t(\theta)\mu(\mathrm{d}\theta)\right] = \int \mathbb{E}\left[\mathbf{1}_A M_t(\theta)\right]\mu(\mathrm{d}\theta) = \int \mathbb{E}\left[\mathbf{1}_A \mathbb{E}\left[M_t(\theta)|\mathcal{F}_{t-1}\right]\right]\mu(\mathrm{d}\theta).$$

Next, again by the assumptions, either $\mathbb{P}|_A \otimes \mu$-a.e., $\mathbb{E}\left[M_t(\theta)|\mathcal{F}_{t-1}\right] \geqslant 0$, or

$$\begin{aligned}
\int \mathbb{E}\left[\left|\mathbb{E}\left[M_t(\theta) \mid \mathcal{F}_{t-1}\right]\right|\right]\mu(\mathrm{d}\theta) &\leqslant \int \mathbb{E}\left[\mathbb{E}\left[|M_t(\theta)| \mid \mathcal{F}_{t-1}\right]\right]\mu(\mathrm{d}\theta) \\
&= \int \mathbb{E}\left[|M_t(\theta)|\right]\mu(\mathrm{d}\theta) < \infty.
\end{aligned}$$

Hence we can apply Fubini's theorem to $\mathbb{E}\left[M_t(\theta)|\mathcal{F}_{t-1}\right]$ on $\mathbb{P}|_A \otimes \mu$:

$$\int \mathbb{E}\left[\mathbf{1}_A \mathbb{E}\left[M_t(\theta)|\mathcal{F}_{t-1}\right]\right]\mu(\mathrm{d}\theta) = \mathbb{E}\left[\mathbf{1}_A \int \mathbb{E}\left[M_t(\theta)|\mathcal{F}_{t-1}\right]\mu(\mathrm{d}\theta)\right].$$

Therefore, for all $A \in \mathcal{F}_{t-1}$, we have $\mathbb{E}\left[\mathbf{1}_A \int M_t(\theta)\mu(\mathrm{d}\theta)\right] = \mathbb{E}\left[\mathbf{1}_A \int \mathbb{E}\left[M_t(\theta)|\mathcal{F}_{t-1}\right]\mu(\mathrm{d}\theta)\right]$. Further, by Fubini's theorem, $\int \mathbb{E}\left[M_t(\theta)|\mathcal{F}_{t-1}\right]\mu(\mathrm{d}\theta)$ is $\mathcal{F}_{t-1}$-measurable. Hence,

$$\mathbb{E}\bigg[\int M_t(\theta)\mu(\mathrm{d}\theta)|\mathcal{F}_{t-1}\bigg] = \int \mathbb{E}[M_t(\theta)|\mathcal{F}_{t-1}]\mu(\mathrm{d}\theta),$$

and so,

$$\begin{aligned}\mathbb{E}[M_t^{\mathsf{mix}}|\mathcal{F}_{t-1}] &= \mathbb{E}\left[\int M_t(\theta)\mu(\mathrm{d}\theta)\middle|\mathcal{F}_{t-1}\right] \\ &= \int \mathbb{E}\left[M_t(\theta)|\mathcal{F}_{t-1}\right]\mu(\mathrm{d}\theta) \leqslant \int M_{t-1}(\theta)\mu(\mathrm{d}\theta) = M_{t-1}^{\mathsf{mix}}.\end{aligned}$$

The fact that M_t^{mix} is $\mathcal{F}_t$-measurable is again guaranteed by Fubini's theorem. Hence (M_t^{mix}) is a supermartingale. The case with submartingales can be proven by considering $-M_t(\theta)$. The case with martingales is proven by combining the cases with supermartingales and submartingales. ■

We remark that the above lemma, albeit stated in terms of forward (super/sub)martingales, immediately implies that the mixture of *reverse* (super/sub)martingales is again a reverse (super/sub)martingale. This is because we allow the indices of the process to run through $t \in \mathbb{Z}$. To wit, letting $\{(N_t(\theta))_{t=1}^{\infty} : \theta \in \Theta\}$ be a family of reverse submartingales on a reverse filtered probability space $(\Omega, \mathcal{A}, (\mathcal{G}_t)_{t=1}^{\infty}, \mathbb{P})$ satisfying the equivalent measurability assumptions, we may set $M_{-t}(\theta) = N_t(\theta)$ and $\mathcal{F}_{-t} = \mathcal{G}_t$ for $t = 1, 2, \ldots$, and trivially extrapolate $M_0(\theta) = M_1(\theta) = \cdots = N_1(\theta)$, $\mathcal{G}_0 = \mathcal{G}_1 = \cdots = \mathcal{F}_1$ to make each $(M_t(\theta))_{t\in\mathbb{Z}}$ a forward submartingale on the forward filtration $(\mathcal{F}_t)_{t\in\mathbb{Z}}$. Lemma 46 is therefore applicable.

References

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1): 131–142, 1966.

P. Alquier. PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008.

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

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

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. *The 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.

R. Amit, B. Epstein, S. Moran, and R. Meir. Integral probability metrics PAC-Bayes bounds. *Advances in Neural Information Processing Systems*, 36, 2022.

J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. In *Annales de l'IHP Probabilités et statistiques*, volume 40, pages 685–736, 2004.

P. Awasthi, S. Kale, S. Karp, and M. Mohri. PAC-Bayes learning bounds for sample-dependent priors. *Advances in Neural Information Processing Systems*, 33:4403–4414, 2020.

A. Balsubramani. PAC-Bayes iterated logarithm bounds for martingale mixtures. *arXiv:1506.06573*, 2015.

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

B. Bercu and A. Touati. Exponential inequalities for self-normalized martingales with applications. *The Annals of Applied Probability*, 18(5):1848–1869, 2008.

R. H. Berk. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

F. Biggs and B. Guedj. Non-vacuous generalisation bounds for shallow neural networks. In *International Conference on Machine Learning*, pages 1963–1981. PMLR, 2022.

F. Biggs and B. Guedj. Tighter PAC-Bayes generalisation bounds by leveraging example difficulty. In *International Conference on Artificial Intelligence and Statistics*, pages 8165–8182. PMLR, 2023.

G. Blanchard and F. Fleuret. Occam's hammer. In *International Conference on Computational Learning Theory*, pages 112–126. Springer, 2007.

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

O. Catoni. A PAC-Bayesian approach to adaptive classification. Technical report, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2003.

O. Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*, volume 1851. Springer Science & Business Media, 2004.

O. Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. Monograph, Institute of Mathematical Statistics lecture notes, 2007.

O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et Statistiques*, volume 48, pages 1148–1185, 2012.

O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv:1712.02747*, 2017.

O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector. *(Almost) 50 Shades of Bayesian Learning: PAC-Bayesian Trends and Insights. NeurIPS Workshop.*, 2018.

P. Chen, X. Jin, X. Li, and L. Xu. A generalized Catoni's M-estimator under finite α-th moment assumption with $\alpha \in (1, 2)$. *Electronic Journal of Statistics*, 15(2):5523–5544, 2021.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158, 1975.

D. A. Darling and H. Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967a.

D. A. Darling and H. Robbins. Iterated logarithm inequalities. *Proceedings of the National Academy of Sciences*, 57(5):1188–1192, 1967b.

B. Delyon. Exponential inequalities for sums of weakly dependent variables. *Electronic Journal of Probability*, 14:752–779, 2009.

M. Donsker and S. Varadhan. Large deviations for Markov processes and the asymptotic evaluation of certain Markov process expectations for large times. In *Probabilistic Methods in Differential Equations*, pages 82–88. Springer, 1975.

R. Durrett. *Probability: theory and examples*. Cambridge university press, 2019.

G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *Uncertainty in Artificial Intelligence*, 2017.

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

M. M. Fard, J. Pineau, and C. Szepesvári. PAC-Bayesian policy evaluation for reinforcement learning. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 195–202, 2011.

H. Flynn, D. Reeb, M. Kandemir, and J. Peters. PAC-Bayesian lifelong learning for multi-armed bandits. *Data Mining and Knowledge Discovery*, 36(2):841–876, 2022.

H. Flynn, D. Reeb, M. Kandemir, and J. Peters. PAC-Bayes bounds for bandit problems: A survey and experimental comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

A. Foong, W. Bruinsma, D. Burt, and R. Turner. How tight can PAC-Bayes be in the small data regime? *Advances in Neural Information Processing Systems*, 34:4093–4105, 2021.

P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 353–360, 2009.

P. Germain, A. Lacasse, F. Laviolette, M. March, and J.-F. Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(26):787–860, 2015.

P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A new PAC-Bayesian perspective on domain adaptation. In *International conference on machine learning*, pages 859–868. PMLR, 2016.

P. Grünwald, R. de Heide, and W. M. Koolen. Safe testing. *Journal of the Royal Statistical Society, Series B. To appear with discussion*, 2023.

B. Guedj. A primer on PAC-Bayesian learning. *Proceedings of the second congress of the French Mathematical Society*, 33, 2019.

B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, 7:264–291, 2013.

M. Haddouche and B. Guedj. PAC-Bayes generalisation bounds for heavy-tailed losses through supermartingales. *Transactions on Machine Learning Research*, 2023.

M. Haddouche, B. Guedj, O. Rivasplata, and J. Shawe-Taylor. PAC-Bayes unleashed: generalisation bounds with unbounded losses. *Entropy*, 23(10):1330, 2021.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

K. G. Jamieson and L. Jain. A bandit approach to sequential experimental design with false discovery control. *Advances in Neural Information Processing Systems*, 31, 2018.

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

A. Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

S. Kullback. *Information theory and statistics*. Wiley, New York, 1959.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

I. Kuzborskij and C. Szepesvári. Efron-Stein PAC-Bayesian inequalities. *arXiv preprint arXiv:1909.01931*, 2019.

T. L. Lai. On confidence sequences. *The Annals of Statistics*, pages 265–280, 1976.

J. Langford and M. Seeger. *Bounds for averaging classifiers*. School of Computer Science, Carnegie Mellon University, 2001.

A. J. Lee. *U-statistics: Theory and Practice*. Routledge, 2019.

G. Letarte, P. Germain, B. Guedj, and F. Laviolette. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

G. Lever, F. Laviolette, and J. Shawe-Taylor. Distribution-dependent PAC-Bayes priors. In *International Conference on Algorithmic Learning Theory*, pages 119–133. Springer, 2010.

R. Liao, R. Urtasun, and R. Zemel. A PAC-Bayesian approach to generalization bounds for graph neural networks. In *International Conference on Learning Representations*, 2020.

R. Livni and S. Moran. A limitation of the PAC-Bayes framework. *Advances in Neural Information Processing Systems*, 33:20543–20553, 2020.

T. Manole and A. Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Transactions on Information Theory*, 2023.

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

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

D. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 164–170, 1999.

D. McAllester. Simplified PAC-Bayesian margin bounds. In *Learning Theory and Kernel Machines*, pages 203–215. Springer, 2003.

Z. Mhammedi, P. Grünwald, and B. Guedj. PAC-Bayes un-expected Bernstein inequality. *Advances in Neural Information Processing Systems*, 32, 2019.

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*, pages 1711–1719. PMLR, 2021.

F. Orabona and K.-S. Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Transactions on Information Theory*, 2023.

E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. PAC-Bayes bounds with data dependent priors. *The Journal of Machine Learning Research*, 13(1):3507–3531, 2012.

M. Pérez-Ortiz, O. Rivasplata, J. Shawe-Taylor, and C. Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22, 2021.

A. Ramdas, J. Ruf, M. Larsson, and W. Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167*, 2020.

A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science (forthcoming)*, 2023.

O. Rivasplata, I. Kuzborskij, C. Szepesvári, and J. Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. *Advances in Neural Information Processing Systems*, 33:16833–16845, 2020.

H. Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

O. Sakhi, P. Alquier, and N. Chopin. PAC-Bayesian offline contextual bandits with guarantees. In *International Conference on Machine Learning*, pages 29777–29799. PMLR, 2023.

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

M. Seeger. Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations. Technical report, University of Edinburgh, 2003.

Y. Seldin and N. Tishby. PAC-Bayesian generalization bound for density estimation with application to co-clustering. In *Artificial Intelligence and Statistics*, pages 472–479. PMLR, 2009.

Y. Seldin and N. Tishby. PAC-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11(12), 2010.

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. *Game-theoretic foundations for probability and finance*, volume 455. John Wiley & Sons, 2019.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational learning theory*, pages 2–9, 1997.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet. On integral probability metrics, ϕ-divergences and binary classification. *arXiv:0901.2698*, 2009.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

N. Thiemann, C. Igel, O. Wintenberger, and Y. Seldin. A strongly quasiconvex PAC-Bayesian bound. In *International Conference on Algorithmic Learning Theory*, pages 466–492. PMLR, 2017.

I. O. Tolstikhin and Y. Seldin. PAC-Bayes-empirical-Bernstein inequality. *Advances in Neural Information Processing Systems*, 26, 2013.

J. Ville. Étude critique de la notion de collectif. *Bull. Amer. Math. Soc*, 45(11):824, 1939.

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

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

H. Wang and A. Ramdas. Catoni-style confidence sequences for heavy-tailed mean estimation. *Stochastic Processes and their Applications*, 2023a.

H. Wang and A. Ramdas. Huber-robust confidence sequences. *International Conference on Artificial Intelligence and Statistics*, 2023b.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B (Methodological), to appear with discussion*, 2023.

I. Waudby-Smith, D. Arbour, R. Sinha, E. H. Kennedy, and A. Ramdas. Time-uniform central limit theory and asymptotic confidence sequences. *arXiv:2103.06476*, 2021.

I. Waudby-Smith, L. Wu, A. Ramdas, N. Karampatziakis, and P. Mineiro. Anytime-valid off-policy inference for contextual bandits. *ACM/IMS Journal of Data Science (forthcoming)*, 2023.

Z. Xu, R. Wang, and A. Ramdas. A unified framework for bandit multiple testing. *Advances in Neural Information Processing Systems*, 34:16833–16845, 2021.