Novel Change of Measure Inequalities with Applications to PAC-Bayesian Bounds and Monte Carlo Estimation

Yuki Ohnishi
Purdue University

Jean Honorio
Purdue University

Abstract

We discuss several novel change of measure inequalities for two families of divergences: f-divergences and α-divergences. We show how the variational representation for f-divergences leads to novel change of measure inequalities. We also present a multiplicative change of measure inequality for α-divergences and a generalized version of Hammersley-Chapman-Robbins inequality. Finally, we present several applications of our change of measure inequalities, including PAC-Bayesian bounds for various classes of losses and non-asymptotic intervals for Monte Carlo estimates.

1 Introduction

The Probably Approximate Correct (PAC) Bayesian inequality was introduced by Shawe-Taylor and Williamson (1997) and McAllester (1999). This framework allows us to produce PAC performance bounds (in the sense of a loss function) for Bayesian-flavored estimators Guedj (2019), and several extensions have been proposed to date (see e.g., Seeger (2003); Catoni (2007); McAllester (2003b,a); Seldin et al. (2012); Ambroladze et al. (2007)). The core of these theoretical results is summarized by a *change of measure inequality.* The change of measure inequality is an expectation inequality involving two probability measures where the expectation with respect to one measure is upper-bounded by the divergence between the two measures and the moments with respect to the other measure. The change of measure inequality also plays a major role in information theory. For instance, Katsoulakis et al. (2017) derived robust uncertainty quantification bounds for statistical estimators of interest with change of measure inequalities. Recent research efforts have been put into more generic perspectives to get PAC-Bayes bounds and to get rid of assumptions such as boundedness of the loss function (see e.g., Lever et al. (2013); Tolstikhin and Seldin (2013); Mhammedi et al. (2019); Holland (2019); Grünwald and Mehta (2019); Shalaeva et al. (2020)). All PAC-Bayesian bounds contained in these works massively rely on one of the most famous change of measure inequalities, which are based on the variational representation of the KL-divergence (Donsker and Varadhan (1975); Csiszar (1975)). Several change of measure inequalities had been proposed along with PAC-Bayes bounds lately. Bégin et al. (2016) proposed a proof scheme of PAC-Bayesian bounds based on the Rényi divergence. Honorio and Jaakkola (2014) proposed an inequality for the χ^2 divergence and derived a PAC-Bayesian bound for linear classification. Alquier and Guedj (2018) proposed a novel change of measure inequality and PAC-Bayesian bounds based on the α-divergence. While most works in the PAC-Bayesian literature propose an analysis based on a specific change of measure inequality, a comprehensive study of change of measure inequalities has not been performed, to the best of our knowledge. Our work proposes several novel and general change of measure inequalities for two families of divergences: f-divergences and α-divergences. It is a well-known fact that the f-divergence can be variationally characterized as the maximum of an optimization problem rooted in the convexity of the function f. This variational representation has been recently used in various applications of information theory, such as f-divergence estimation Nguyen et al. (2010b) and quantification of the bias in adaptive data analysis Jiao et al. (2017). On the theoretical side, Ruderman et al. (2012) showed that the variational representation of the f-divergence can be tightened when the convex dual is constrained to the space of probability measures, as opposed to the space of all measures.

Our main contributions are as follows:

- We derive several change of measure inequali-

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

ties, based on the variational representation for f-divergences. We perform the analysis for the constrained regime (to the space of probability densities) as well as the unconstrained regime (the space of all integrable functions).

- We present a multiplicative change of measure inequality for the family of α-divergences. This generalizes the previous results Alquier and Guedj (2018); Honorio and Jaakkola (2014) for the α-divergence, which in turns enables PAC-Bayes inequalities for types of losses not considered before.
- We also generalize prior results for the Hammersley-Chapman-Robbins inequality Lehmann and Casella (1998) from the particular χ^2 divergence, to the family of α-divergences.
- We provide new PAC-Bayesian bounds with the α-divergence and the χ^2-divergence from our novel change of measure inequalities for bounded, sub-Gaussian, sub-exponential and bounded-variance loss functions. Our results are either novel, or have a tighter complexity term than existing results in the literature, and pertain to important machine learning prediction problems, such as regression, classification and structured prediction.
- We provide a new scheme for estimation of non-asymptotic intervals for Monte Carlo estimates. Our results indicate that the empirical mean over a sampling distribution concentrates around an expectation with respect to any arbitrary distribution.

For quick reference, our contributions are summarized in Tables 1, 3 and 4.

2 Change of Measure Inequalities

In this section, we formalize the definition of f-divergences and present the constrained representation (to the space of probability measures) as well as the unconstrained representation. Then, we provide different change of measure inequalities for several divergences. We also provide multiplicative bounds as well as a generalized Hammersley-Chapman-Robbins bound. Table 1 summarizes our results.

2.1 Change of Measure Inequality from the Variational Representation of f-divergences

Let $f : (0, +\infty) \to \mathbb{R}$ be a convex function. The convex conjugate f^* of f is defined by:

$$f^*(y) = \sup_{x \in \mathbb{R}} (xy - f(x)). \tag{1}$$

The definition of f^* yields the following Young-Fenchel inequality

$$f(x) \geq xy - f^*(y)$$

which holds for any y. Using the notation of convex conjugates, the f-divergence and its variational representation is defined as follows.

Definition 1 (f-divergence)**.** *Let $\mathcal{H}$ be any arbitrary domain. Let P and Q denote the probability measures over the Borel σ-field on $\mathcal{H}$. Additionally, let f: $[0, \infty) \to \mathbb{R}$ be a convex and lower semi-continuous function that satisfies $f(1) = 0$.*

$$D_f(Q\|P) := \mathbb{E}_P\left[f\left(\frac{dQ}{dP}\right)\right]$$

For simplicity, we denote $\mathbb{E}_P[\cdot] \equiv \mathbb{E}_{h\sim P}[\cdot]$ in the sequel. Many common divergences, such as the KL-divergence, the χ^2-divergence and the Hellinger divergence, are members of the family of f-divergences, coinciding with a particular choice of $f(t)$. Table 2 presents the definition of each divergence with the corresponding generator $f(t)$. It is well known that the f-divergence can be characterized as the following variational representation.

Lemma 1 (Variational representation of f-divergence, Lemma 1 in Nguyen et al. (2010a))**.** *Let $\mathcal{H}$, P, Q and f be defined as in Definition 1. The f-divergence from P to Q is characterized as*

$$D_f(Q\|P) \geq \sup_{\phi} \mathbb{E}_Q[\phi] - \mathbb{E}_P[f^*(\phi)]$$

where the supremum is over all real-valued functions ϕ: $\mathcal{H} \to \mathbb{R}$.

Ruderman et al. (2012) shows that this variational representation for f-divergences can be tightened.

Theorem 1 (Change of measure inequality from the constrained variational representation for f-divergences Ruderman et al. (2012))**.** *Let $\mathcal{H}$, P, Q and f be as in Definition 1. Let ϕ: $\mathcal{H} \to \mathbb{R}$ be a real-valued function. Let $\Delta(\mu) := \{g : \mathcal{H} \to \mathbb{R} : g \geq 0, \|g\|_1 = 1\}$ denote the space of probability densities with respect to μ, where the norm is defined as $\|g\|_1 := \int_{\mathcal{H}} |g| d\mu$, given a measure μ over $\mathcal{H}$. The general form of the change of measure inequality for f-divergences is given by*

$$\mathbb{E}_Q[\phi] \leq D_f(Q\|P) + (\mathbb{I}^R_{f,P})^*(\phi)$$
$$(\mathbb{I}^R_{f,P})^*(\phi) = \sup_{p \in \Delta(P)} \mathbb{E}_P[\phi p] - \mathbb{E}_P[f(p)]$$

where p is constrained to be a probability density function.

Table 1: Our novel change of measure inequalities. For simplicity, we denote $\mathbb{E}_P[\cdot] \equiv \mathbb{E}_{h\sim P}[\cdot]$ and $\phi \equiv \phi(h)$.

Bound Type	**Divergence**	**Uppper-Bound for Every Q and a Fixed P**	**Reference**
Constrained Variational Representation	KL	$\mathbb{E}_Q[\phi] \le KL(Q\|P) + \log(\mathbb{E}_P[e^{\phi}])$	McAllester (1999)
	Pearson χ^2	$\mathbb{E}_Q[\phi] \le \chi^2(Q\|P) + \mathbb{E}_P[\phi] + \frac{1}{4}\mathbb{V}\mathrm{ar}_P[\phi]$	Lemma 2 **[New]**
	Total Variation	$\mathbb{E}_Q[\phi] \le TV(Q\|P) + \mathbb{E}_P[\phi]$ for $\phi \in [0,1]$	Lemma 4
Unconstrained Variational Representation	KL	$\mathbb{E}_Q[\phi] \le KL(Q\|P) + (\mathbb{E}_P[e^{\phi}] - 1)$	McAllester (1999)
	Pearson χ^2	$\mathbb{E}_Q[\phi] \le \chi^2(Q\|P) + \mathbb{E}_P[\phi] + \frac{1}{4}\mathbb{E}_P[\phi^2]$	Lemma 3 **[New]**
	Total Variation	$\mathbb{E}_Q[\phi] \le TV(Q\|P) + \mathbb{E}_P[\phi]$ for $\phi \in [0,1]$	
	α	$\mathbb{E}_Q[\phi] \le D_\alpha(Q\|P) + \frac{(\alpha-1)^{\frac{\alpha}{\alpha-1}}}{\alpha}\mathbb{E}_P[\phi^{\frac{\alpha}{\alpha-1}}] + \frac{1}{\alpha(\alpha-1)}$	Lemma 5 **[New]**
	Squared Hellinger	$\mathbb{E}_Q[\phi] \le H^2(Q\|P) + \mathbb{E}_P[\frac{\phi}{1-\phi}]$ for $\phi\ < 1$	Lemma 6 **[New]**
	Reverse KL	$\mathbb{E}_Q[\phi] \le \overline{KL}(Q\|P) + \mathbb{E}_P[\log(\frac{1}{1-\phi})]$ for $\phi\ < 1$	Lemma 7 **[New]**
	Neyman χ^2	$\mathbb{E}_Q[\phi] \le \overline{\chi^2}(Q\|P) + 2 - 2\mathbb{E}_P[\sqrt{1-\phi}]$ for $\phi\ < 1$	Lemma 8 **[New]**
Multiplicative	Pearson χ^2	$\mathbb{E}_Q[\phi] \le \sqrt{(\chi^2(Q\|P)+1)\mathbb{E}_P[\phi^2]}$	Honorio and Jaakkola (2014)
	α	$\mathbb{E}_Q[\phi] \le (\alpha(\alpha-1)D_\alpha(Q\|P)+1)^{\frac{1}{\alpha}}(\mathbb{E}_P[\lvert\phi\rvert^{\frac{\alpha}{\alpha-1}}])^{\frac{\alpha-1}{\alpha}}$	Alquier and Guedj (2018)
Generalized HCR	Pearson χ^2	$\chi^2(Q\|P) \ge \frac{(\mathbb{E}_Q[\phi]-\mathbb{E}_P[\phi])^2}{\mathbb{V}\mathrm{ar}_P[\phi]}$	Lehmann and Casella (1998)
	Pseudo α	$\lvert\mathbb{E}_Q[\phi]-\mathbb{E}_P[\phi]\rvert \le \widetilde{\mathcal{D}_\alpha}(Q\|P)^{\frac{1}{\alpha}}(\mathbb{E}_P[\lvert\phi-\mu_P\rvert^{\frac{\alpha}{\alpha-1}}])^{\frac{\alpha-1}{\alpha}}$	Lemma 12 **[New]**

Table 2: Some common f-divergences with corresponding generator.

Divergence	**Formula with probability measures P and Q defined on a common space $\mathcal{H}$**	**Corresponding Generator$f(t)$**
KL	$KL(Q\|P) = \int_{\mathcal{H}} \log\frac{dQ}{dP}dQ$	$t\log t - t + 1$
Reverse KL	$\overline{KL}(Q\|P) = \int_{\mathcal{H}} \log\frac{dP}{dQ}dP$	$-\log t$
Pearson χ^2	$\chi^2(Q\|P) = \int_{\mathcal{H}}(\frac{dQ}{dP}-1)^2dP$	$(t-1)^2$
Neyman χ^2	$\overline{\chi^2}(Q\|P) = \int_{\mathcal{H}}(\frac{dP}{dQ}-1)^2dQ$	$\frac{(1-t)^2}{t}$
Total Variation	$TV(Q\|P) = \frac{1}{2}\int_{\mathcal{H}}\lvert\frac{dQ}{dP}-1\rvert dP$	$\frac{1}{2}\lvert t-1\rvert$
Squared Hellinger	$H^2(Q\|P) = \int_{\mathcal{H}}(\sqrt{\frac{dQ}{dP}}-1)^2dP$	$(\sqrt{t}-1)^2$
α	$D_\alpha(Q\|P) = \frac{1}{\alpha(\alpha-1)}\int_{\mathcal{H}}(\lvert\frac{dQ}{dP}\rvert^\alpha - 1)dP$	$\frac{t^\alpha-1}{\alpha(\alpha-1)}$
Pseudo α	$\widetilde{\mathcal{D}_\alpha}(Q\|P) = \int_{\mathcal{H}}\lvert\frac{dQ}{dP}-1\rvert^\alpha dP$	$\lvert t-1\rvert^\alpha$
ϕ_p (Alquier and Guedj, 2018)	$D_{\phi_p-1}(Q\|P) = \int_{\mathcal{H}}(\lvert\frac{dQ}{dP}\rvert^p - 1)dP$	$t^p - 1$

The famous Donsker-Varadhan representation for the KL-divergence, which is used in most PAC-Bayesian bounds, can be actually derived from this tighter representation by setting $f(t) = t\log(t)$. However, it is not always easy to find a closed-form solution for Theorem 1, as it requires to resort to variational calculus, and in some cases, there is no closed-form solution. In such a case, we can use the following corollary to obtain looser bounds, but only requires to find a convex conjugate.

Corollary 1 (Change of measure inequality from the unconstrained variational representation for f-divergences)**.** *Let P, Q, f and ϕ be as in Theorem 1. By Definition 1, we have*

$$\forall Q \text{ on } \mathcal{H}: \ \mathbb{E}_Q[\phi] \le D_f(Q\|P) + \mathbb{E}_P[f^*(\phi)]$$

Detailed proofs can be found in Appendix A. By choosing a right function f and deriving the constrained maximization term $(\mathbb{I}^R_{f,P})^*(\phi)$ with the help of variational calculus, we can create an upper-bound based on the corresponding divergence $D_f(Q\|P)$. Next, we discuss the case of the χ^2 divergence.

Lemma 2 (Change of measure inequality from the constrained representation of the χ^2-divergence)**.** *Let P, Q and ϕ be as in Theorem 1, we have*

$$\forall Q \text{ on } \mathcal{H}: \ \mathbb{E}_Q[\phi] \le \chi^2(Q\|P) + \mathbb{E}_P[\phi] + \frac{1}{4}\mathbb{V}\mathrm{ar}_P[\phi]$$

The bound in Lemma 2 is slightly tighter than the one without the constraint. The change of measure inequality without the constraint is given as follows.

Lemma 3 (Change of measure inequality from the unconstrained representation of the Pearson χ^2-divergence)**.** *Let P, Q and ϕ be as in Theorem 1, we have*

$$\forall Q \text{ on } \mathcal{H}: \ \mathbb{E}_Q[\phi] \leq \chi^2(Q\|P) + \mathbb{E}_P[\phi] + \frac{1}{4}\mathbb{E}_P[\phi^2]$$

As might be apparent, the bound in Lemma 2 is tighter than the one in Lemma 3 by $(\mathbb{E}_P[\phi])^2$ because $\mathbb{V}\text{ar}_P[\phi] \leq \mathbb{E}_P[\phi^2]$. Next, we discuss the case of the total variation divergence.

Lemma 4 (Change of measure inequality from the constrained representation of the total variation divergence)**.** *Let ϕ: $\mathcal{H} \to [0,1]$ be a real-valued function. Let P and Q be as in Theorem 1, we have*

$$\forall Q \text{ on } \mathcal{H}: \ \mathbb{E}_Q[\phi] \leq TV(Q\|P) + \mathbb{E}_P[\phi]$$

Interestingly, we can obtain the same bound on the total variation divergence even if we use the unconstrained variational representation. Next, we state our result for α-divergences.

Lemma 5 (Change of measure inequality from the unconstrained representation of the α-divergence)**.** *Let P, Q and ϕ be as in Theorem 1. For $\alpha > 1$, we have*

$$\forall Q \text{ on } \mathcal{H}:$$

$$\mathbb{E}_Q[\phi] \leq D_\alpha(Q\|P) + \frac{(\alpha-1)^{\frac{\alpha}{\alpha-1}}}{\alpha}\mathbb{E}_P[\phi^{\frac{\alpha}{\alpha-1}}] + \frac{1}{\alpha(\alpha-1)}$$

We can obtain the bounds based on the squared Hellinger divergence $H^2(Q\|P)$, the reverse KL-divergence $\overline{KL}(Q\|P)$ and the Neyman χ^2-divergence $\overline{\chi^2}(Q\|P)$ in a similar fashion.

Lemma 6 (Change of measure inequality from the unconstrained representation of the squared Hellinger divergence)**.** *Let ϕ: $\mathcal{H} \to (-\infty, 1)$ be a real-valued function. Let P and Q be as in Theorem 1, we have*

$$\forall Q \text{ on } \mathcal{H}: \ \mathbb{E}_Q[\phi] \leq H^2(Q\|P) + \mathbb{E}_P\left[\frac{\phi}{1-\phi}\right]$$

Similarly, we obtain the following bound for the reverse-KL divergence.

Lemma 7 (Change of measure inequality from the unconstrained representation of the reverse KL-divergence)**.** *Let ϕ: $\mathcal{H} \to (-\infty, 1)$ be a real-valued function. Let P and Q be as in Theorem 1, we have*

$$\forall Q \text{ on } \mathcal{H}: \ \mathbb{E}_Q[\phi] \leq \overline{KL}(Q\|P) + \mathbb{E}_P\left[\log\left(\frac{1}{1-\phi}\right)\right]$$

Finally, we prove our result for the Neyman χ^2 divergence based on a similar approach.

Lemma 8 (Change of measure inequality from the unconstrained representation of the Neyman χ^2-divergence)**.** *Let ϕ: $\mathcal{H} \to (-\infty, 1)$ be a real-valued function. Let P and Q be as in Theorem 1, we have*

$$\forall Q \text{ on } \mathcal{H}: \ \mathbb{E}_Q[\phi] \leq \overline{\chi^2}(Q\|P) + 2 - 2\mathbb{E}_P[\sqrt{1-\phi}]$$

2.2 Multiplicative Change of Measure Inequality for α-divergences

First, we state a known result for the χ^2 divergence.

Lemma 9 (Multiplicative change of measure inequality for the χ^2-divergence (Honorio and Jaakkola, 2014))**.** *Let P, Q and ϕ be as in Theorem 1, we have*

$$\forall Q \text{ on } \mathcal{H}: \ \mathbb{E}_Q[\phi] \leq \sqrt{(\chi^2(Q\|P)+1)\mathbb{E}_P[\phi^2]}$$

First, we note that the χ^2 divergence is an α-divergence for $\alpha = 2$. Next, we generalize the above bound for any α-divergence.

Lemma 10 (Multiplicative change of measure inequality for the α-divergence)**.** *Let P, Q and ϕ be as in Theorem 1. For any $\alpha > 1$, we have*

$$\forall Q \text{ on } \mathcal{H}:$$

$$\mathbb{E}_Q[\phi] \leq \big(\alpha(\alpha-1)D_\alpha(Q\|P) + 1\big)^{\frac{1}{\alpha}}\big(\mathbb{E}_P[|\phi|^{\frac{\alpha}{\alpha-1}}]\big)^{\frac{\alpha-1}{\alpha}}$$

Our bound is stated in the form of α-divergence. By choosing $\alpha = 2$, we have the same bound as Lemma 9 where $\chi^2(Q\|P) = 2D_\alpha(Q\|P)$. We will later apply the above α-divergence change of measure to obtain PAC-Bayes inequalities for types of losses not considered before (Alquier and Guedj, 2018; Honorio and Jaakkola, 2014).

2.3 A Generalized Hammersley-Chapman-Robbins (HCR) Inequality

The HCR inequality is a famous information theoretic inequality for the χ^2-divergence.

Lemma 11 (HCR inequality (Lehmann and Casella, 1998))**.** *Let P, Q and ϕ be as in Theorem 1, we have*

$$\forall Q \text{ on } \mathcal{H}: \ \chi^2(Q\|P) \geq \frac{(E_Q[\phi] - \mathbb{E}_P[\phi])^2}{\mathbb{V}\text{ar}_P[\phi]}$$

Next, we generalize the above bound for α-divergence.

Lemma 12 (The generalization of HCR inequality.)**.** *Let P, Q and ϕ be as in Theorem 1. For any $\alpha > 1$, we have*

$$\forall Q \text{ on } \mathcal{H}:$$

$$\big|\mathbb{E}_Q[\phi] - \mathbb{E}_P[\phi]\big| \leq \widetilde{\mathcal{D}_\alpha}(Q\|P)^{\frac{1}{\alpha}}\big(\mathbb{E}_P[|\phi-\mu_P|^{\frac{\alpha}{\alpha-1}}]\big)^{\frac{\alpha-1}{\alpha}}$$

where

$$\widetilde{\mathcal{D}_\alpha}(Q\|P) = \int_{\mathcal{H}} |\frac{dQ}{dP} - 1|^\alpha dP \text{ and } \mu_P = \mathbb{E}_P[\phi]$$

We call $\widetilde{\mathcal{D}_\alpha}(Q\|P)$ a pseudo α-divergence, which is a member of the family of f-divergences with $f(t) = |t-1|^\alpha$ for $\alpha > 1$. See Appendix B for the formal proof. Straightforwardly, we can obtain Lemma 11 by choosing $\alpha = 2$.

3 Application to PAC-Bayesian Theory

In this section, we will explore the applications of our change of measure inequalities. We consider an arbitrary input space $\mathcal{X}$ and a output space $\mathcal{Y}$. The samples $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are input-output pairs. Each example (x, y) is drawn according to a fixed, but unknown, distribution D on $\mathcal{X} \times \mathcal{Y}$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ denote a generic loss function. The risk $R_D(h)$ of any predictor $h : \mathcal{X} \to \mathcal{Y}$ is defined as the expected loss induced by samples drawn according to D. Given a training set S of m samples, the empirical risk $R_S(h)$ of any predictor h is defined by the empirical average of the loss. That is

$$R_D(h) = \mathop{\mathbb{E}}_{(x,y)\sim D} \ell(h(x), y)$$

$$R_S(h) = \frac{1}{|S|} \sum_{(x,y)\in S} \ell(h(x), y)$$

In the PAC-Bayesian framework, we consider a hypothesis space $\mathcal{H}$ of predictors, a prior distribution P on $\mathcal{H}$, and a posterior distribution Q on $\mathcal{H}$. In the classical PAC-Bayes framework, the prior is specified before exploiting the information contained in S, while the posterior is obtained by running a learning algorithm on S. There are advances on extending the classical framework to data-dependent P, e.g., Rivasplata et al. (2020). The PAC-Bayesian theory usually studies the stochastic Gibbs predictor G_Q. Given a distribution Q on $\mathcal{H}$, G_Q predicts an example x by drawing a predictor h according to Q, and returning $h(x)$. The risk of G_Q is then defined as follows. For any probability distribution Q on a set of predictors, the Gibbs risk $R_D(G_Q)$ is the expected risk of the Gibbs predictor G_Q relative to D. Hence,

$$R_D(G_Q) = \mathop{\mathbb{E}}_{(x,y)\sim D} \mathop{\mathbb{E}}_{h\sim Q} \ell(h(x), y) \tag{2}$$

Usual PAC-Bayesian bounds give guarantees on the generalization risk $R_D(G_Q)$. Typically, these bounds rely on the empirical risk $R_S(G_Q)$ defined as follows.

$$R_S(G_Q) = \frac{1}{|S|} \sum_{(x,y)\in S} \mathop{\mathbb{E}}_{h\sim Q} \ell(h(x), y) \tag{3}$$

Due to space constraints, we fully present PAC-Bayes generalization bounds for losses with bounded variance. Other results for bounded losses, sub-Gaussian losses and sub-exponential losses are included in Appendix C. Still, we briefly discuss our new results in Section 3.2.

3.1 Loss Function with Bounded Variance

In this section, we present our PAC-Bayesian bounds for the loss functions with bounded variance, i.e., any arbitrary distribution with bounded variance on the loss function ℓ (i.e., $\mathbb{V}\text{ar}_{(x,y)\sim D}[\ell(h(x), y)] \le \sigma^2$ for any $h \in \mathcal{H}$). The assumption of bounded variance is similar to the assumption of bounded second moment used by Holland (2019).

Suppose that we have a convex function $\Delta : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, that measures the discrepancy between the observed empirical Gibbs risk $R_S(G_Q)$ and the true Gibbs risk $R_D(G_Q)$ on distribution Q. Given that, the purpose of the PAC-Bayesian theorem is to upper-bound the discrepancy $t\Delta(R_D(G_Q), R_S(G_Q))$ for any $t > 0$. Let $\phi_D(h) := t\Delta(R_D(h), R_S(h))$, where the subscript of ϕ_D shows the dependency on the data distribution D. Let $\Delta(q, p) = (q - p)^2$.

Proposition 1 (The PAC-Bayesian bounds for loss function with bounded variance)**.** *Let P be a fixed prior distribution over a hypothesis space any $h \in \mathcal{H}$. For a given posterior distribution Q over an infinite hypothesis space $\mathcal{H}$, let $R_D(G_Q)$ and $R_S(G_Q)$ be the Gibbs risk and the empirical Gibbs risk as in Equation* (2) *and* (3) *respectively. For the sample size m and $\alpha > 1$, with probability at least $1 - \delta$, simultaneously for all posterior distributions Q, we have*

$$R_D(G_Q) \le R_S(G_Q) + \sqrt{\frac{\sigma^2}{m\delta}\big(\alpha(\alpha-1)D_\alpha(Q\|P) + 1\big)^{\frac{1}{\alpha}}}$$

$$R_D(G_Q) \le R_S(G_Q) + \sqrt{\frac{1}{m}\left(D_\alpha(Q\|P) + \frac{1}{\alpha(\alpha-1)}\right) + \frac{1}{m\alpha}\left(\frac{\sigma^2(\alpha-1)}{\delta}\right)^{\frac{\alpha}{\alpha-1}}} \tag{4}$$

Proofs are given in Appendix. By setting $\alpha = 2$ in Proposition 2, we have the following claim.

Corollary 2 (The PAC-Bayesian bounds with χ^2-divergence for bounded variance loss function)**.** *Let P be any prior distribution over an infinite hypothesis space $\mathcal{H}$. For a given posterior distribution Q over an infinite hypothesis space $\mathcal{H}$, let $R_D(G_Q)$ and $R_S(G_Q)$ be the Gibbs risk and the empirical Gibbs risk as in Equation* (2) *and* (3) *respectively. For the sample size $m > 0$, with probability at least $1 - \delta$, simultaneously*

Table 3: Our novel PAC-Bayesian bounds with α-divergence and χ^2-divergence.

Loss & Divergence		**Upper bound for** $R_D(G_Q) - R_S(G_Q)$	**Reference**
Bounded Loss	α	$O\big(R[\frac{1}{m}\log(\frac{1}{\delta})]^{\frac{1}{2}} D_\alpha(Q\|P)^{\frac{1}{2\alpha}}\big)$	Proposition 5 [**New**]
	α	$O\big((\frac{1}{m})^{\frac{1}{2}}[D_\alpha(Q\|P) + (R^2\log(\frac{1}{\delta}))^{\frac{\alpha}{\alpha-1}}]^{\frac{1}{2}}\big)$	Proposition 5 [**New**]
		There are no bounds to compare with in the literature	
	χ^2	$O\big(R[\frac{1}{m}\log(\frac{1}{\delta})]^{\frac{1}{2}} \chi^2(Q\|P)^{\frac{1}{4}}\big)$	Corollary 3 [**New**]
	χ^2	$O\big((\frac{1}{m})^{\frac{1}{2}}[\chi^2(Q\|P) + (R^2\log(\frac{1}{\delta}))^2]^{\frac{1}{2}}\big)$	Corollary 3 [**New**]
0-1 loss	χ^2	$O\big((\frac{1}{m\delta})^{\frac{1}{2}} \chi^2(Q\|P)^{\frac{1}{2}}\big)$	Honorio and Jaakkola (2014); Bégin et al. (2016)
		Our multiplicative bound is $O((\chi^2)^{1/4})$ *and* $O((\log(1/\delta))^{1/2})$ *while the comparison bound is* $O((\chi^2)^{1/2})$ *and* $O((1/\delta)^{1/2})$	
Sub-Gaussian	α	$O\big(\sigma[\frac{1}{m}\log(\frac{1}{\delta})]^{\frac{1}{2}} D_\alpha(Q\|P)^{\frac{1}{2\alpha}}\big)$	Proposition 6 [**New**]
	α	$O\big((\frac{1}{m})^{\frac{1}{2}}[D_\alpha(Q\|P) + (\sigma^2\log(\frac{1}{\delta}))^{\frac{\alpha}{\alpha-1}}]^{\frac{1}{2}}\big)$	Proposition 6 [**New**]
	α	$O\big(\sigma(\frac{1}{m})^{\frac{1}{2}}(\frac{1}{\delta})^{\frac{\alpha-1}{\alpha}} D_\alpha(Q\|P)^{\frac{1}{\alpha}}\big)$	Theorem 1 & Proposition 6 in Alquier and Guedj (2018)
		Our multiplicative bound is $O((D_\alpha)^{1/2\alpha})$ *and* $O((\log(1/\delta))^{1/2})$ *while the comparison bound is* $O((D_\alpha)^{1/\alpha})$ *and* $O((1/\delta)^{\alpha/(\alpha-1)})$	
	χ^2	$O\big(\sigma[\frac{1}{m}\log(\frac{1}{\delta})]^{\frac{1}{2}} \chi^2(Q\|P)^{\frac{1}{4}}\big)$	Corollary 4 [**New**]
	χ^2	$O\big((\frac{1}{m})^{\frac{1}{2}}[\chi^2(Q\|P) + (\sigma^2\log(\frac{1}{\delta}))^2]^{\frac{1}{2}}\big)$	Corollary 4 [**New**]
	χ^2	$O\big(\sigma(\frac{1}{m\delta})^{\frac{1}{2}} \chi^2(Q\|P)^{\frac{1}{2}}\big)$	Theorem 1 & Proposition 6 in Alquier and Guedj (2018)
		Our multiplicative bound is $O((\chi^2)^{1/4})$ *and* $O((\log(1/\delta))^{1/2})$ *while the comparison bound is* $O((\chi^2)^{1/2})$ *and* $O(((1/\delta))^{1/2})$	
Sub-exponential	α	$O\big(\frac{\beta}{m}\log(\frac{1}{\delta})D_\alpha(Q\|P)^{\frac{1}{2\alpha}}\big)$	Proposition 7 [**New**]
For	α	$O\big((\frac{1}{m})^{\frac{1}{2}}[D_\alpha(Q\|P) + m^{\frac{-\alpha}{\alpha-1}}(\beta\log(\frac{1}{\delta}))^{\frac{2\alpha}{\alpha-1}}]^{\frac{1}{2}}\big)$	Proposition 7 [**New**]
$m < \frac{2\beta^2\log(\frac{2}{\delta})}{\sigma^2}$	χ^2	$O\big(\frac{\beta}{m}\log(\frac{1}{\delta})\chi^2(Q\|P)^{\frac{1}{4}}\big)$	Corollary 5 [**New**]
	χ^2	$O\big((\frac{1}{m})^{\frac{1}{2}}[(\chi^2(Q\|P) + \frac{1}{m^2}(\beta\log(\frac{1}{\delta}))^4)]^{\frac{1}{2}}\big)$	Corollary 5 [**New**]
		There are no bounds to compare with in the literature	
Bounded Variance	α	$O\big(\sigma(\frac{1}{m\delta})^{\frac{1}{2}} D_\alpha(Q\|P)^{\frac{1}{2\alpha}}\big)$	Proposition 1 [**New**]
	α	$O\big((\frac{1}{m})^{\frac{1}{2}}[D_\alpha(Q\|P) + (\frac{\sigma^2}{\delta})^{\frac{\alpha}{\alpha-1}}]^{\frac{1}{2}}\big)$	Proposition 1 [**New**]
	α	$O\big((\frac{1}{m})^{\frac{1}{2}}(\frac{\sigma^2}{\delta})^{\frac{\alpha-1}{\alpha}}(D_\alpha(Q\|P))^{\frac{1}{\alpha}}\big)$	Proposition 4 in Alquier and Guedj (2018)
		Our multiplicative bound is $O((D_\alpha)^{1/2\alpha})$ *and* $O((1/\delta)^{1/2})$ *while the comparison bound is* $O((D_\alpha)^{1/\alpha})$ *and* $O((1/\delta)^{\alpha/(\alpha-1)})$	
	χ^2	$O\big(\sigma(\frac{1}{m\delta})^{\frac{1}{2}} \chi^2(Q\|P)^{\frac{1}{4}}\big)$	Corollary 2 [**New**]
	χ^2	$O\big((\frac{1}{m})^{\frac{1}{2}}[\chi^2(Q\|P) + (\frac{\sigma^2}{\delta})^2]^{\frac{1}{2}}\big)$	Corollary 2 [**New**]
	χ^2	$O\big(\sigma(\frac{1}{m\delta})^{\frac{1}{2}} \chi^2(Q\|P)^{\frac{1}{2}}\big)$	Corollary 1 in Alquier and Guedj (2018)
		Our multiplicative bound is $O((\chi^2)^{1/4})$ *while the comparison bound is* $O((\chi^2)^{1/2})$	

for all posterior distributions Q, we have

$$R_D(G_Q) \leq R_S(G_Q) + \sqrt{\frac{\sigma^2}{m\delta}\sqrt{\chi^2(Q\|P)+1}} \qquad (5)$$

$$R_D(G_Q) \leq R_S(G_Q) + \sqrt{\frac{1}{2m}\left(\chi^2(Q\|P) + 1 + \left(\frac{\sigma^2}{\delta}\right)^2\right)}$$

Please see Table 3 for detailed comparisons with existing results.

3.2 Discussion

Table 3 presents various PAC-Bayesian bounds based on our change of measure inequalities depending on different assumptions on the loss function ℓ, and compares them with existing results in the literature. The importance of the various types of PAC-Bayes bounds is justified by the connection between PAC-Bayes bounds and regularization in a learning problem. PAC-Bayesian theory provides a guarantee that upper-bounds the risk of Gibbs predictors si-

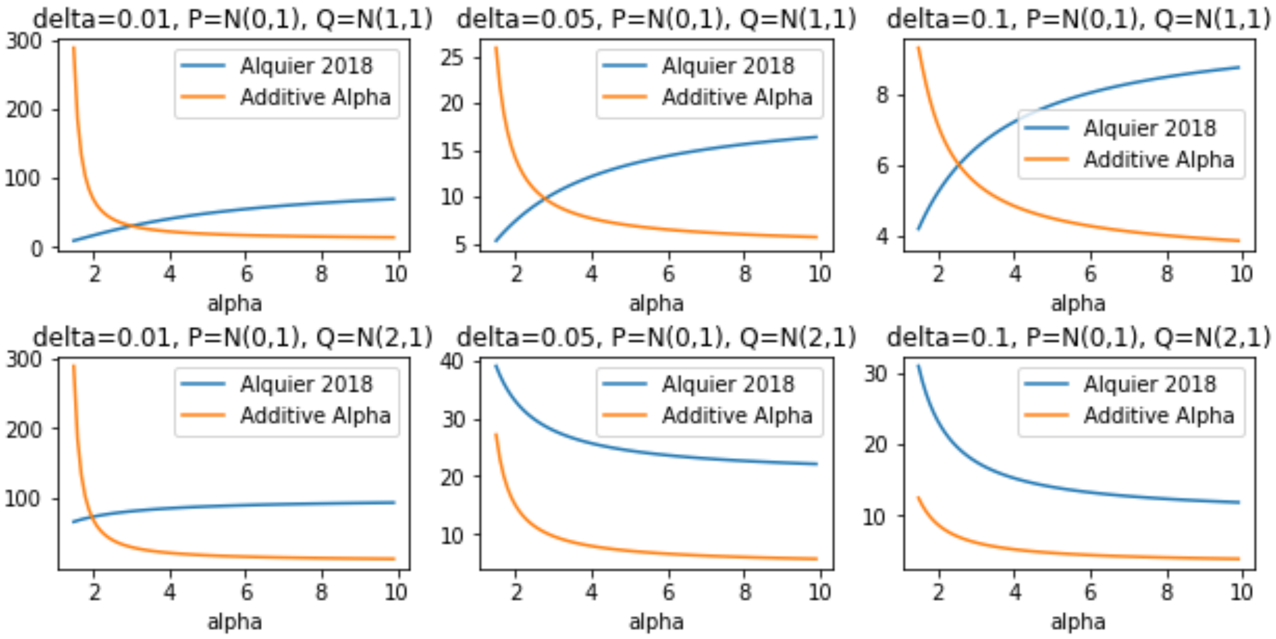

Figure 1: Bound comparisons for different values of α and δ.

multaneously for all posterior distributions Q. PAC-Bayes bounds enjoy this property due to change of measure inequalities. It is a well-known fact that KL-regularized objective functions are obtained from PAC-Bayes risk bounds, in which we can find the best posterior among all the possible posteriors Q. The complexity term, partially controlled by the divergence, serves as a regularizer (Germain et al., 2009, 2007; Bousquet and Elisseeff, 2002). Additionally, the link between Bayesian inference techniques and PAC-Bayesian risk bounds was shown by Germain et al. (2016), that is, the minimization of PAC-Bayesian risk bounds maximizes the Bayesian marginal likelihood when the loss is the negative log-likelihood. Our results pertain to important machine learning prediction problems, such as regression, classification and structured prediction and indicate which regularizer to use. All PAC-Bayes bounds presented here are either novel, or have a tighter complexity term than existing results in the literature. Since our bounds are based on either χ^2-divergence or α-divergence, we excluded the comparison with the bounds based on the KL-divergence (Seeger, 2003; Catoni, 2007; Holland, 2019; Mhammedi et al., 2019; Germain et al., 2016; Grünwald and Mehta, 2019; Sheth and Khardon, 2017). Our results for sub-exponential losses are entirely novel. For the other cases, our bounds are tighter than exisiting bounds in terms of the complexity term. For instance, our bound for bounded losses is tighter than those of Honorio and Jaakkola (2014); Bégin et al. (2016) since our bound has the complexity term $\chi^2(Q\|P)^{1/4}$ and $\log(1/\delta)$, while Honorio and Jaakkola (2014); Bégin et al. (2016) have $\chi^2(Q\|P)^{1/2}$ and $1/\delta$. For sub-Gaussian loss functions, our bound has the complexity term $D_\alpha(Q\|P)^{1/2\alpha}$ and $\log(1/\delta)$, whereas Alquier and Guedj (2018) has $D_\alpha(Q\|P)^{1/\alpha}$ and $1/\delta$ respectively. In addition, our additive bounds, such as Equation (4), have better rates than the existing bounds in Alquier and Guedj (2018), since in our bound, $D_\alpha(Q\|P)$ and $1/\delta$ are added, while in Alquier and Guedj (2018), $D_\alpha(Q\|P)$ and $1/\delta$ are multiplied.

Figure 3 compares the convergence rate of the additive bound in Proposition 1 with respect to α and δ. We consider Gaussian distributions for the prior and posterior distribution with different mean. The upper panels portray the comparisons of the bounds with the prior $N(0,1)$ and the optimal posterior distribution $N(1,1)$. The lower panels depict those with the optimal posterior $N(2,1)$. As might be readily apparent, our additive bound works better with large α. We can also see that the additive bound is robust since it works similarly regardless of the choice of prior distribution, α and δ. Even when we choose a "bad" prior distribution as in the lower panels in which the prior is far way from the optimal posterior, the bounds behave nicely.

Table 4: Our novel non-asymptotic interval for Monte Carlo estimates

Divergence	**Non-asymptotic interval for Monte Carlo estimates**	**Reference**
	$\left\lvert\mathbb{E}_Q[\phi(X)] - \frac{1}{n}\sum_{i=1}^n \phi(X_i)\right\rvert \leq \frac{4L^2\log(\frac{2}{\delta})}{n\gamma} + \mathcal{K}$	
Pseudo α	$\mathcal{K} = \frac{2^{\frac{2\alpha-1}{\alpha}} L\widetilde{\mathcal{D}_\alpha}(Q\Vert P)^{\frac{1}{\alpha}}}{\sqrt{\gamma}}\Gamma(\frac{3\alpha-2}{2(\alpha-1)})^{\frac{\alpha-1}{\alpha}}$	Proposition 2 [**New**]
χ^2	$\mathcal{K} = \begin{cases} \sqrt{\chi^2(Q\Vert P)\{\frac{1}{n}\sum_{i=1}^n \phi^2(x_i) + \frac{16L^2}{\gamma}\sqrt{\frac{1}{n}\log\frac{2}{\delta}}\}} \text{ if } \log(\frac{2}{\delta}) \leq n \\ \sqrt{\chi^2(Q\Vert P)\{\frac{1}{n}\sum_{i=1}^n \phi^2(x_i) + \frac{16L^2}{n\gamma}\log\frac{2}{\delta}\}} \text{ if } n < \log(\frac{2}{\delta}) \end{cases}$	Proposition 3 [**New**]
KL	$\mathcal{K} = KL(Q\Vert P) + \frac{L^2}{n\gamma}$	Proposition 4 [**New**]
	There are no bounds to compare with in the literature	

4 Non-Asymptotic Interval for Monte Carlo Estimates

We now turn to another application of change of measure inequalities. We will introduce a methodology that enables us to find a non-asymptotic interval for Monte Calro (MC) estimate. All results shown in this section are entirely novel and summarized in Table 4. Under some circumstances, our methodology could be a promising alternative to existing MC methods, e.g., Importance Sampling and Rejection Sampling (Robert and Casella, 2005) because their non-asymptotic intervals are hard to analyze. In this section, we consider a problem to estimate an expectation of an Lipschitz function $\phi : \mathbb{R}^d \to \mathbb{R}$ with respect to a complicated distribution Q, namely $\mathbb{E}_Q[\phi(X)]$ by the sample mean $\frac{1}{n}\sum_{i=1}^n \phi(X_i)$ over a distribution P, where Q is any distribution we are not able to sample from. For a motivating application, consider Q being a probabilistic graphical model (see e.g., Honorio (2011)). Assume that we have a strongly log-concave distribution P with a parameter γ for a sampling distribution (see Appendix D for definition). Under the above conditions, we claim the following proposition.

Proposition 2 (A general expression of non-asymptotic interval for Monte Carlo estimates)**.** *Let X be a d-dimensional random vector. Let P and Q be the probability measures over R^d. Assume P is strongly log-concave with parameter $\gamma > 0$. Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be any L-Lipschitz function with respect to the Euclidean norm. Suppose we draw i.i.d. samples $X_1, X_2, ..., X_n \sim P$. For $\alpha > 1$, with probability at least $1-\delta$, we have*

$$\left|\mathbb{E}_Q[\phi(X)] - \frac{1}{n}\sum_{i=1}^n \phi(X_i)\right| \leq \frac{4L^2\log(\frac{2}{\delta})}{n\gamma} + \frac{2^{\frac{2\alpha-1}{\alpha}} L\widetilde{\mathcal{D}_\alpha}(Q\|P)^{\frac{1}{\alpha}}}{\sqrt{\gamma}}\Gamma\left(\frac{3\alpha-2}{2(\alpha-1)}\right)^{\frac{\alpha-1}{\alpha}}$$

The second term on the right hand side indicates a bias of an empirical mean $\frac{1}{n}\sum_{i=1}^n \phi(X_i)$ under the sampling distribution P. Next, we present a more informative bound.

Proposition 3 (χ^2-based expression of non-asymptotic interval for Monte Carlo estimates)**.** *Let X, P, Q, L, γ and ϕ be as in Proposition 2 . Suppose we have i.i.d. samples $X_1, X_2, ..., X_n \sim P$. Then, with probability at least $(1-\delta)^2$, we have*

$$\left|\mathbb{E}_Q[\phi(X)] - \frac{1}{n}\sum_{i=1}^n \phi(X_i)\right| \leq \frac{4L^2\log(\frac{2}{\delta})}{n\gamma} + \mathcal{K}$$

where

$$\mathcal{K} = \begin{cases} \sqrt{\chi^2(Q\|P)\{\frac{1}{n}\sum_{i=1}^n \phi^2(X_i) + \frac{16L^2}{\gamma}\sqrt{\frac{1}{n}\log\frac{2}{\delta}}\}} \\ \hfill \textit{if } \log(\frac{2}{\delta}) \leq n \\ \sqrt{\chi^2(Q\|P)\{\frac{1}{n}\sum_{i=1}^n \phi^2(X_i) + \frac{16L^2}{n\gamma}\log\frac{2}{\delta}\}} \\ \hfill \textit{if } n < \log(\frac{2}{\delta}) \end{cases}$$

This result provides an insight into how good a proposal distribution P is, meaning that the effect of the deviation between P and Q, namely $\chi^2(Q\|P)$, is inflated by the empirical variance. This implication might support some results in the literature (Cornebise et al. (2008); Dieng et al. (2017)). So far we have considered the pseudo α-divergence as well as χ^2 divergence. Duembgen et al. (2010) presented an approximation for an arbitrary distribution Q by distributions with log-concave density with respect to the KL divergence, which motivates the following result.

Proposition 4 (KL-based expression of non-asymptotic interval for Monte Carlo estimates)**.** *Let X, P, Q, L, γ and ϕ be as in Proposition 2 . Suppose we have i.i.d. samples $X_1, X_2, ..., X_n \sim P$. Then, with probability at least $1-\delta$, we have*

$$\left|\mathbb{E}_Q[\phi(X)] - \frac{1}{n}\sum_{i=1}^n \phi(X_i)\right| \leq \frac{4L^2\log(\frac{2}{\delta})}{n\gamma} + KL(Q\|P) + \frac{L^2}{n\gamma}$$

This result shows that, under the assumption of Proposition 4, the empirical mean $\frac{1}{n}\sum_{i=1}^n \phi(X_i)$ over the sampling distribution P aymptotically differs from $\mathbb{E}_Q[\phi(X)]$ by $KL(Q\|P)$ at most.

References

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

Ambroladze, A., Parrado-hernández, E., and Shawe-taylor, J. S. (2007). Tighter PAC-Bayes Bounds. In Schölkopf, B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 9–16. MIT Press.

Bousquet, O. and Elisseeff, A. (2002). Stability and Generalization. *Journal of Machine Learning Research*, 2:499–526.

Bégin, L., Germain, P., Laviolette, F., and Roy., J.-F. (2016). *PAC-Bayesian Bounds based on the Rényi Divergence.* International Conference on Artificial Intelligence and Statistics.

Catoni, O. (2007). Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning.

Cornebise, J., Moulines, É., and Olsson, J. (2008). Adaptive Methods for Sequential Importance Sampling with Application to State Space Models. *Statistics and Computing*, 18(4):461–480.

Csiszar, I. (1975). *I*-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability*, 3(1):146–158.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. (2017). Variational Inference via chi Upper Bound Minimization. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 2732–2741. Curran Associates, Inc.

Donsker, M. D. and Varadhan, S. R. S. (1975). Asymptotic Evaluation of Certain Markov Process Expectations for Large Time. *Communications on Pure and Applied Mathematics*, 28(1):1–47.

Duembgen, L., Samworth, R., and Schuhmacher, D. (2010). Approximation by Log-Concave Distributions, with Applications to Regression. *The Annals of Statistics*, 39.

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

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. (2007). A PAC-Bayes Risk Bound for General Loss Functions. In Schölkopf, B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 449–456. MIT Press.

Germain, P., Lacasse, A., Marchand, M., Shanian, S., and Laviolette, F. (2009). From PAC-Bayes Bounds to KL Regularization. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 603–610. Curran Associates, Inc.

Grünwald, P. D. and Mehta, N. A. (2019). A Tight Excess Risk Bound via a Unified PAC-Bayesian–Rademacher–Shtarkov–MDL Complexity. In Garivier, A. and Kale, S., editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 433–465, Chicago, Illinois. PMLR.

Guedj, B. (2019). A Primer on PAC-Bayesian Learning. *Preprint arXiv:1901.05353.*

Holland, M. (2019). PAC-Bayes Under Potentially Heavy Tails. In *Advances in Neural Information Processing Systems 32*, pages 2715–2724. Curran Associates, Inc.

Honorio, J. (2011). Lipschitz Parametrization of Probabilistic Graphical Models. pages 347–354. Conference on Uncertainty in Artificial Intelligence.

Honorio, J. and Jaakkola, T. (2014). Tight Bounds for the Expected Risk of Linear Classifiers and PAC-Bayes Finite-Sample Guarantees. In *International Conference on Artificial Intelligence and Statistics*, page 384–392.

Jiao, J., Han, Y., and Weissman, T. (2017). Dependence Measures Bounding the Exploration Bias for General Measurements. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1475–1479.

Katsoulakis, M. A., Rey-Bellet, L., and Wang, J. (2017). Scalable Information Inequalities for Uncertainty Quantification. *Journal of Computational Physics*, 336:513 – 545.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation.* Springer Verlag.

Lever, G., Laviolette, F., and Shawe-Taylor, J. (2013). Tighter PAC-Bayes Bounds Through Distribution-Dependent Priors. *Theor. Comput. Sci.*, 473:4–28.

McAllester, D. (2003a). Simplified PAC-Bayesian Margin Bounds. In Schölkopf, B. and Warmuth, M. K., editors, *Learning Theory and Kernel Machines*, pages 203–215, Berlin, Heidelberg. Springer Berlin Heidelberg.

McAllester, D. A. (1998). Some PAC-Bayesian Theorems. In *Machine Learning*, pages 230–234. ACM Press.

McAllester, D. A. (1999). PAC-Bayesian Model Averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, COLT '99, pages 164–170, New York, NY, USA. ACM.

McAllester, D. A. (2003b). PAC-Bayesian Stochastic Model Selection. *Machine Learning*, 51(1):5–21.

Mhammedi, Z., Grünwald, P., and Guedj, B. (2019). PAC-Bayes Un-Expected Bernstein Inequality. In *Advances in Neural Information Processing Systems 32*, pages 12202–12213. Curran Associates, Inc.

Nguyen, X., Wainwright, M., and Jordan, M. (2010a). Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization. *Information Theory, IEEE Transactions on*, 56:5847 – 5861.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010b). Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization. *IEEE Trans. Inf. Theor.*, 56(11):5847–5861.

Rivasplata, O., Kuzborskij, I., Szepesvári, C., and Shawe-Taylor, J. (2020). PAC-Bayes Analysis Beyond the Usual Bounds. In *Advances in Neural Information Processing Systems 34*.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Ruderman, A., Reid, M. D., García-García, D., and Petterson, J. (2012). Tighter Variational Representations of F-divergences via Restriction to Probability Measures. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1155–1162, USA. Omnipress.

Seeger, M. (2003). Pac-bayesian Generalisation Error Bounds for Gaussian Process Classification. *J. Mach. Learn. Res.*, 3:233–269.

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

Shalaeva, V., Esfahani, A. F., Germain, P., and Petreczky, M. (2020). Improved PAC-Bayesian Bounds for Linear Regression. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 5660–5667. AAAI Press.

Shawe-Taylor, J. and Williamson, R. C. (1997). A PAC Analysis of a Bayesian Estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, COLT '97, pages 2–9, New York, NY, USA. ACM.

Sheth, R. and Khardon, R. (2017). Excess Risk Bounds for the Bayes Risk using Variational Inference in Latent Gaussian Models. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5151–5161. Curran Associates, Inc.

Tolstikhin, I. O. and Seldin, Y. (2013). PAC-Bayes-Empirical-Bernstein Inequality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 109–117. Curran Associates, Inc.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.