HAL
open science

On change of measure inequalities for f-divergences

Antoine Picard-Weibel, Benjamin Guedj

To cite this version:

Antoine Picard-Weibel, Benjamin Guedj. On change of measure inequalities for f-divergences. 2022. hal-03573465

HAL Id: hal-03573465
https://inria.hal.science/hal-03573465v1

Preprint submitted on 14 Feb 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Preprint 1–17

On change of measure inequalities for f-divergences

Antoine Picard-Weibel APICARD.W@GMAIL.COM
SUEZ, CIRSEE, France
Benjamin Guedj BENJAMIN.GUEDJ@INRIA.FR
Inria and University College London, France and United Kingdom

arXiv:2202.05568v1 [stat.ML] 11 Feb 2022

Abstract

We propose new change of measure inequalities based on f-divergences (of which the Kullback-Leibler divergence is a particular case). Our strategy relies on combining the Legendre transform of f-divergences and the Young-Fenchel inequality. By exploiting these new change of measure inequalities, we derive new PAC-Bayesian generalisation bounds with a complexity involving f-divergences, and holding in mostly unchartered settings (such as heavy-tailed losses). We instantiate our results for the most popular f-divergences.

Keywords: Change of measure, f-divergence, Legendre transform, PAC-Bayesian theory, Kullback–Leibler divergence.

1. Introduction

The past few years have witnessed a surge of interest in PAC-Bayesian generalisation bounds, largely due to a number of papers introducing non-vacuous generalisation bounds for deep neural networks (starting with the pioneering work of Dziugaite and Roy, 2017 and more recently by Neyshabur et al., 2018, Zhou et al., 2019, Letarte et al., 2019, Perez-Ortiz et al., 2021a,b, Biggs and Guedj, 2021, 2022a,b, to name but a few). The PAC-Bayes theory originates in the seminal work of McAllester (1998, 1999) and was further developed by Catoni (2003, 2004, 2007), among other authors – we refer to the recent surveys Guedj (2019) and Alquier (2021) for an introduction to the field. This renewed interest in PAC-Bayes is calling for further development of the theory, and in particular its extension to new settings: notable recent contributions have investigated learning with unbounded or heavy-tailed losses (Alquier and Guedj, 2018; Holland, 2019; Guedj and Pujol, 2021; Haddouche et al., 2021) or with novel versions of the risk (contrastive losses in Nozawa et al., 2020, conditional value at risk in Mhammedi et al., 2020, with structure in Cantelobre et al., 2020).

A cornerstone of PAC-Bayesian generalisation bounds is the celebrated change of measure formula, established concomitantly by Csiszár (1975) and Donsker and Varadhan (1975) (see also Catoni, 2004, Equation 5.2.1 for a more recent write-up, or Guedj, 2019, Lemma 1 – this is Equation (1) below). A celebrated scheme of proof for PAC-Bayes generalisation bounds, established by Catoni (2004, 2007), consists in (i) using a deviation inequality (such as Bernstein's or Hoeffding's) to upper bound the distance between the risk of a chosen predictor and its empirical counterpart, (ii) making this deviation inequality valid for all predictors simultaneously, through the change of measure inequality (also known as the Legendre transform of the Kullback-Leibler divergence). This second step is the one we target in this paper: we propose as key contribution a novel change of measure inequality based on f-divergences (of which the Kullback-Leibler is a special case), which can in turn be plugged in any proof following the two-steps Catoni strategy, delivering new PAC-Bayesian generalisation bounds as byproducts.

© A. Picard-Weibel & B. Guedj.

Let $\mathcal{X}$ (respectively $\mathcal{Y}$) be a set of input data (respectively output data), and let $\Phi \subseteq \mathcal{Y}^{\mathcal{X}}$ be a set of predictors, *i.e.*, functions $\mathcal{X} \to \mathcal{Y}$. We equip Φ with a reference measure π, and we let $\mu \ll \nu$ denote that the measure μ is absolutely continuous with respect to ν. The Kullback-Leibler divergence between μ and ν is noted $\mathrm{KL}(\mu, \nu)$. For any function $h\colon \Phi \to \mathbb{R}$, the change of measure is given by

$$\log \int (\exp \circ h) \mathrm{d}\pi = \sup_{\nu \in \Pi} \left\{ \int h \mathrm{d}\nu - \mathrm{KL}(\nu, \pi) \right\}. \tag{1}$$

In Section 4, we will particularise the above with h being a functional of the risk, to yield PAC-Bayes generalisation bounds. The supremum in the right hand side of (1) is taken over all measures ν which are absolutely continuous with respect to π, *e.g.*, for which the Kullback-Leibler divergence is non-trivial. Unfortunately, as Equation (1) involves a Kullback-Leibler divergence penalisation between ν and π, this implies that we must control an exponential moment of the function h (as discussed in Remark 2): we argue in this paper that this stringent condition can be relaxed.

To the best of our knowledge, addressing the second step of the Catoni route to PAC-Bayes generalisation bounds has been mostly studied by Bégin et al. (2016), Alquier and Guedj (2018) and Ohnishi and Honorio (2021). Bégin et al. (2016) propose a unifying framework to derive PAC-Bayes bounds and extend previously known results to the case of the Rényi α-divergence; building on these results, Alquier and Guedj (2018) adapt the proof to be valid for f-divergences. We adopt a different scheme of proof and deliver new change of measure inequalities which result in generalisation bounds – of which the results of Alquier and Guedj (2018) are particular cases. We also acknowledge the work of Ohnishi and Honorio (2021): they produces change of measure inequalities similar to ours. Nonetheless, we believe we offer a different perspective on such inequalities, centered on the notion of Legendre transform (not discussed in the work of Ohnishi and Honorio, 2021), which offers direct implications on the tightness of the bounds, ways to optimise them and on the assumptions required on the underlying data generating distribution. We moreover introduce tighter bounds for a subclass of f-divergences (see our Lemma 4), and introduce new change of measure inequalities for some typical f-divergences: to the best of our knowledge, no previous such inequalities existed.

Contributions. We establish change of measure inequalities with f-divergence penalisation, for any convex function f. Our strategy is to upper-bound the Legendre transform of $\mathcal{D}_f(\cdot, \pi)$, the f-divergence where the second argument is fixed, then optimise on degrees of freedom. Doing so, we recapture the original change of measure from Csiszár (1975) and Donsker and Varadhan (1975) recalled in Equation (1) (we also refer to van Erven and Harremoës, 2014 for additional insights). A byproduct of our approach is to derive a unifying framework, and improve on the generalisation bounds obtained by Alquier and Guedj (2018) through a different technique. The resulting generalisation bounds hold for what Alquier and Guedj (2018) call "hostile data" (*i.e.*, not complying with classical and comforting statistics assumptions such as i.i.d. observations or bounded exponential moments of the loss – as expected from real-world data), therefore contributing to an extension of PAC-Bayes towards previously unchartered territories.

Our framework is suited for all generalised Bayesian predictors, regardless of whether the task be classification or regression, whether the data is independent and identically distributed or has some dependence structure. We stress here that we contribute a unified building block towards new generalisation bounds: more refined bounds are possible at the price of a more *ad hoc* treatment.

We consider a measurable space (Φ, Σ_Φ) equipped with a reference probability measure π. We let Π denote the set of all probability measures on (Φ, Σ_Φ) that are absolutely continuous with respect to π, and for $\nu \in \Pi$, we let $\frac{\mathrm{d}\nu}{\mathrm{d}\pi}$ be the Radon–Nykodym derivative of ν with respect to π. We consider the set

$$\mathcal{F} = \{f \mid f \in \overline{\mathbb{R}}^{\mathbb{R}},\ f \text{ convex},\ f(1) = 0,\ f(\mathbb{R}_+^*) \subset \mathbb{R},\ f(\mathbb{R}_-^*) = \{+\infty\}\}.$$

For all $f \in \mathcal{F}$, we define the f-divergence between two probability measures on (Φ, Σ_Φ) as

$$\mathcal{D}_f(\nu, \mu) := \begin{cases} \int f\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right) \mathrm{d}\mu & \text{if } \nu \ll \mu, \\ +\infty & \text{else.} \end{cases}$$

The Kullback-Leibler divergence is given by the generator $f(x) = x\log(x)$. The bounds we derive are of the following form: $\forall \pi, \forall \nu \ll \pi, \forall h : \Phi \mapsto \mathbb{R}$,

$$\int h \mathrm{d}\nu \leq \inf_{\gamma > 0,\ c} \left\{ c + \gamma \left(B_{f,\pi}(\gamma^{-1}(h - c)) + \mathcal{D}_f(\nu, \pi) \right) \right\} \tag{2}$$

where the functional $B_{f,\pi}$ should be thought of as a moment term with respect to the measure π and as such does not depend on ν. We also study the minimisation problem on γ and c for standard choices of f. We show that playing on the degree of freedom γ in (2) can actually yield different types of bounds, where the trade off between the f-divergence term and the moment term is not additive but rather multiplicative. In particular, we recapture the original bound from McAllester (1998, 1999) in the special case where $\mathcal{D}_f$ is the Kullback–Leibler divergence, but we also recapture (and improve on) bounds obtained by Alquier and Guedj (2018) for the power generator $f_p(x) \simeq x^p$.

Outline. We introduce our notation in Section 2 and our main results in Section 3 (Theorem 1 and Theorem 3): we then discuss how these can be plugged in the Catoni route to PAC-Bayes generalisation bounds, and instantiate those for the most popular f-divergences in Section 4. We highlight future avenues for work in Section 5 and gather additional proofs and comments in Appendix A.

2. Notation

For $f \in \mathcal{F}$, we denote f^* the Legendre transform of f given by

$$f^*(t) = \sup_{x \in \mathbb{R}^+} \{xt - f(x)\}.$$

If f is not C^1, f' denotes f's sub-derivative (respectively, $f^{*\prime}$ denotes the subderivative of f^* whenever f^* is not C^1), and in such a case evaluation of $f'(x)$ is achieved by taking any value in the subderivative of f at x. Even if f' does not admit a derivative to the right at 0, we note $f'(0) := \inf \cup_{x>0} f'(x)$ and $f'(\infty) := \sup \cup_{x>0} f'(x)$. Therefore, $\forall t \leq f'(0)$, $f^*(t) = -f(0)$, and $\forall t \geq f'(\infty)$, $f^*(t) = \infty$. In particular, we remark that $f^*(\max(t, f'(0)) = f^*(t)$.

We remind here some useful properties of Legendre transform which will be used throughout:

$$f^{**} = f, \qquad f^{*\prime} = f'^{-1}, \qquad f(x) = xf' - f^* \circ f'(c)$$

(these properties hold true *mutatis mutandis* using subgradient if f does not admit a derivative) and

$$\forall a, b, \qquad ab \leq f(a) + f^*(b).$$

This last property, which is a direct consequence of the definition of the Legendre transform, is the Young-Fenchel inequality, and plays an essential role in our arguments.

We consider the Legendre transform of the convex operator $\mathcal{D}_{f,\pi} = \mathcal{D}_f(\cdot, \pi)$, defined on all π-measurable function h as

$$\mathcal{D}^*_{f,\pi}(h) := \sup_{\mu \in \Pi} \left(\int h \mathrm{d}\mu - \mathcal{D}_f(\mu, \pi) \right). \tag{3}$$

Note that $\mathcal{D}^*_{f,\pi}(h)$ can take infinite values, which occurs whenever h is not π-integrable.

We use the standard notation $\mathbb{E}_{\mathbb{P}}$ for the expectation with respect to the probability measure $\mathbb{P}$ on the observation space, while $\mathbb{E}_\nu$, $\mathbb{E}_\pi$ denote expectations with respect to the measures defined on the predictor space Φ, that is $\mathbb{E}_\nu[\cdot] = \int \cdot \, \mathrm{d}\nu$.

3. Main results

3.1. Preliminary: bounding the Legendre transform of the f-divergence

We recall that the change of measure inequality established by Csiszár (1975) and Donsker and Varadhan (1975) states that the Kullback–Leibler divergence satisfies

$$\log \int (\exp \circ h) \mathrm{d}\pi = \sup_{\nu \in \Pi} \left\{ \int h \mathrm{d}\nu - \mathrm{KL}(\nu, \pi) \right\} \tag{4}$$

for any real valued Σ_Φ-measurable function h defined on Φ, where the supremum is taken on all probability measures ν such that $\mathrm{KL}(\nu, \pi) < \infty$. One can read (4) as stating that the left hand side of the equality is the Legendre transform of the Kullback–Leibler divergence with the second parameter π fixed. The variational formula implies that, $\forall \pi$, $\forall \nu \in \Pi$, $\forall h : \Phi \mapsto \mathbb{R}$,

$$\int h \mathrm{d}\nu \leq \log \int (\exp \circ h) \mathrm{d}\pi + \mathrm{KL}(\nu, \pi).$$

Such a decomposition upper-bounds the mean of a function with respect to the probability ν by two terms: a measure of the distance between ν and a reference probability π, and a moment of the function integrated with respect to the reference measure. Young-Fenchel's inequality implies that such a strategy can be used for all f-divergences, since

$$\forall h \in \Phi, \forall \nu \in \Pi, \qquad \int h \mathrm{d}\nu \leq \mathcal{D}^*_{f,\pi}(h) + \mathcal{D}_f(\nu, \pi), \tag{5}$$

Necessary assumptions on the generalisation gap, and consequently on the loss, could then be reversed-enginered from the specific form of the Legendre transform $\mathcal{D}^*_{f,\pi}$. Unfortunately, computing the Legendre transform is not always possible: indeed, we show that, under mild conditions on f, identifying the density reaching the supremum in the definition of the Legendre transform amounts to solving

$$\begin{cases} \frac{\mathrm{d}\nu^*}{\mathrm{d}\pi} = f^{*\prime}(h + c^*) \\ \int \frac{\mathrm{d}\nu^*}{\mathrm{d}\pi} \mathrm{d}\pi = 1 \end{cases} \tag{6}$$

where the constant c^* such that the second equation is satisfied is often not computable (see Theorem 1). Nonetheless, any upper bound of the Legendre transform of the f-divergence can be used to obtain Young-Fenchel inequalities. The "true" Legendre transform would be optimal in the sense that any function L such that $\mathbb{E}_\nu[h] \leq L(h) + \mathcal{D}_f(\nu, \pi)$ for any h and $\nu \in \Pi$ must satisfy $L(h) \geq \mathcal{D}^*_{f,\pi}(h)$. This is a strong incentive to search for tractable, yet tight, upper bounds of the Legendre transform.

A straightforward upper bound of the Legendre transform is given by

$$\mathcal{D}_{f,\pi}(h) \leq \mathcal{L}_{f,\pi}(h) := \mathbb{E}_\pi\left[f^*(h)\right] \tag{7}$$

with convention that $\mathbb{E}_\pi\left[f^*(h)\right] = +\infty$ if $f^*(h)$ is not π-integrable. This is a direct consequence of Young-Fenchel's inequality on f. Though simple, this upperbound is enough to recapture Theorem 1 in Alquier and Guedj (2018), and can be used on all f convex.

A tighter bound can be obtained on the further hypothesis that f is twice differentiable and that $1/f''$ is concave. Then, starting from Lemma 14.2 in Boucheron et al. (2013), we show that $\mathcal{D}^*_{f,\pi}(h)$ is bounded by

$$\tilde{\mathcal{D}}^*_{f,\pi}(h) := \mathbb{E}_\pi\left[f^*(h)\right] + f'\left(\mathbb{E}_\pi\left[f^{*\prime}(h)\right]\right) - f^* \circ f'\left(\mathbb{E}_\pi\left[f^{*\prime}(h)\right]\right) \tag{8}$$

with the convention that (i) whenever $f^*(h)$ is not π-integrable, $\tilde{\mathcal{D}}^*_{f,\pi}(h) = +\infty$, (ii) whenever $f^{*\prime}(h)$ is not π-integrable, $\tilde{\mathcal{D}}^*_{f,\pi}(h) = \mathbb{E}_\pi\left[f^*(h)\right]$. Since $f(1) = 0$, it follows that $\forall t, t - f^*(t) \leq 0$ which implies that (8) gives indeed a tighter bound than (7).

Note that $f \mapsto \mathcal{D}_f$ is a linear operator defined on the vector space of functions such that $f(1) = 0$, and has kernel $\mathrm{Ker} = \{cf_\delta \mid c \in \mathbb{R}\}$, where $f_\delta : t \mapsto t - 1$. As such, the "true" Legendre transform and the resulting bound do not depend on which representative f_c in $f + \mathrm{Ker}$ is chosen. However, the upperbounds $\tilde{\mathcal{D}}_{f_c,\pi}$ and $\mathcal{L}_{f_c,\pi}$ do vary with c. In more details,

$$\tilde{\mathcal{D}}^*_{f_c,\pi}(h) = \tilde{\mathcal{D}}^*_{f,\pi}(h + c) - c \qquad \text{and} \qquad \mathcal{L}_{f_c,\pi}(h) = \mathcal{L}_{f,\pi}(h + c) - c.$$

Therefore $\tilde{\mathcal{D}}^*_{f,\pi}$ and $\mathcal{L}_{f,\pi}$ are not (in general) the "true" Legendre transform. Besides, $\tilde{\mathcal{D}}^*_{f,\pi}$ is not even necessarily convex.

This implies that one can pick c in such a way as to minimise the upperbounds $\tilde{\mathcal{D}}_{f,c,\pi}$ and $\mathcal{L}_{f_c,\pi}$. Moreover, noticing that

$$\int f \mathrm{d}\nu \leq \mathcal{D}^*_{\lambda f,\pi}(h) + \mathcal{D}_{\lambda f}(\nu, \pi),$$

the upper-bounds can also be minimised with respect to a scale parameter λ. All in all, we optimise (5) on affine transformations $f \to \lambda f + cf_\delta$. Interestingly enough, the transforms $h \to \lambda^{-1}(h - c)$ and $f \to \lambda f + cf_\delta$ result in the same bounds, both when considering $\mathcal{L}_{f,\pi}$ and $\tilde{\mathcal{D}}^*_{f,\pi}$, and therefore there is no other degree of freedom to consider.

Optimising the bound on c actually recaptures the exact Legendre transform, for both loose and tight bounds, on mild conditions on f.

Theorem 1 *For all $f \in \mathcal{F}$ strictly convex, for all probability measure π on Φ and for all π-measurable function h such that $\mathcal{D}^*_{f,\pi}(h) < \infty$,*

$$\mathcal{D}^*_{f,\pi}(h) = \inf_{c \in \mathbb{R}} \mathbb{E}_\pi\left[f^*(h + c)\right] - c. \tag{9}$$

Moreover, any c^* such that $\mathbb{E}_\pi\left[f^{*\prime}(h + c^*)\right] = 1$ is a minimiser of the right hand side of (9), while the probability measure defined by $\frac{d\nu^*}{d\pi} = f^{*\prime}(h + c^*)$ maximises (3).

Assuming $1/f''$ is concave, as $\mathcal{D}^*_{f,\pi} \leq \tilde{\mathcal{D}}^*_{f,\pi} \leq \mathcal{L}_{f,\pi}$, Theorem 1 is also valid if one replaces $\mathcal{L}_{f,\pi}$ by $\tilde{\mathcal{D}}^*_{f,\pi}$ in Equation (9).

Remark 2 *A consequence of Theorem 1 is that, if one wishes to bound $\mathbb{E}_\pi[h]$ by something of the form $B_{f,\pi}(h)+\gamma\mathcal{D}_f(\nu,\pi)$ simultaneously for all ν, then there must exists c such that the f^* moment of $\gamma^{-1}(h-c)$ is upperbounded. This implies that in the case of Kullback–Leibler, the exponential moment assumption can not be weakened, since $f^*(t) = \exp(t-1)$.*

Proof [Theorem 1] Let us first remark that, since $\mathcal{D}_f$ is invariant on $f \to f + cf_\delta$,

$$\sup_{\nu\in\Pi}\left\{\mathbb{E}_\nu\left[h\right] - \mathcal{D}_f(\nu,\pi)\right\} = \mathcal{D}^*_{f,\pi}(h) \leq \inf_{c\in\mathbb{R}}\left\{\mathbb{E}_\pi\left[f^*(h+c)\right] - c\right\}. \tag{10}$$

Now assume that there exists $c^* \in \mathbb{R}$ such that $\mathbb{E}_\pi\left[f^{*\prime}(h+c^*)\right] = 1$. Since $f^{*\prime}$ takes non negative values, we can define the probability measure ν^* such that $\frac{\mathrm{d}\nu^*}{\mathrm{d}\pi} = f^{*\prime}(h+c^*)$. Therefore

$$\begin{aligned}
\mathcal{D}^*_{f,\pi}(h) &= \sup_{\nu\in\Pi}\left(\mathbb{E}_\nu\left[h\right] - \mathcal{D}_f(\nu,\pi)\right)(h) \\
&\geq \mathbb{E}_\pi\left[hf^{*\prime}(h+c^*) - f\circ f^{*\prime}(h+c^*)\right] \\
&\geq \mathbb{E}_\pi\left[f^*(h+c^*)\right] - c^* \\
&\geq \inf_c\left\{\mathbb{E}_\pi\left[f^*(h+c)\right] - c\right\}
\end{aligned}$$

which implies equality from (10). There only remains to prove the existence of c^*. Notice that the function $M : c \mapsto \mathbb{E}_\pi\left[f^{*\prime}(h+c)\right]$ is non decreasing, and goes to 0 as $c \to -\infty$, and goes to $+\infty$ as $c \to +\infty$. As f is strictly convex, $f^{*\prime}$ is continuous and therefore c^* exists. An interesting consequence of this observation is that the Legendre transform must satisfy

$$\mathcal{D}^*_{f,\pi}(h+c) = \mathcal{D}^*_{f,\pi}(h) + c, \qquad \forall c \in \mathbb{R}.$$

■

3.2. A fresh look at the change of measure

We let $\mathcal{F}_c$ denote the subset of $\mathcal{F}$ of continuous functions on $\mathbb{R}_+$ and twice differentiable on $\mathbb{R}^*_+$ such that $1/f''$ is concave.

Theorem 3 *For all $f \in \mathcal{F}_c$, $\lambda > 0$, c, $h : \Phi \mapsto \mathbb{R}$ π-measurable,*

$$\mathbb{E}_\nu\left[h\right] \leq \lambda^{-1}\tilde{\mathcal{D}}^*_{f,\pi}\left(\lambda(h-c)\right) + c + \lambda^{-1}\mathcal{D}_f\left(\nu,\pi\right). \tag{11}$$

For all f convex, such that $f(1) = 0$, for all $\lambda > 0$, c, $h : \Phi \mapsto \mathbb{R}$ π-measurable,

$$\mathbb{E}_\nu\left[h\right] \leq \lambda^{-1}\mathcal{L}_{f,\pi}\left(\lambda\left(h-c\right)\right) + c + \lambda^{-1}\mathcal{D}_f\left(\nu,\pi\right). \tag{12}$$

Before stating the proof of Theorem 3, we need the following intermediary result.

Lemma 4 *For $f \in \mathcal{F}_c$, for all ν such that $\mathcal{D}_f\left(\nu,\pi\right) < \infty$, $\mathcal{D}_f\left(\nu,\pi\right)$ is equal to*

$$\mathcal{D}_f\left(\nu,\pi\right) = \sup_h\left\{\mathbb{E}_\nu\left[h\right] - \tilde{\mathcal{D}}^*_{f,\pi}(h)\right\} \tag{13}$$

where the supremum is taken on all π-measurable functions h.

Proof [Lemma 4] The proof starts with Lemma 14.2 in Boucheron et al. (2013), which states that for any f convex, twice differentiable on $\mathbb{R}_+^*$ such that $\frac{1}{f''}$ is concave, for any Z such that $f(Z)$ is π-integrable, then

$$\mathbb{E}_\pi\left[f(Z)\right] - f\left(\mathbb{E}_\pi Z\right) = \sup_{T\neq 0}\left\{\mathbb{E}_\pi\left[\left(f'(T) - f'(\mathbb{E}_\pi T)\right)(Z-T) + f(T)\right] - f\left(\mathbb{E}_\pi T\right)\right\}$$

where the supremum is taken on all non negative π-integrable random variables T. The maximum is achieved for $T = Z$. For ν such that $\mathcal{D}_f(\nu,\pi) < \infty$, applying this result to $Z = \frac{\mathrm{d}\nu}{\mathrm{d}\pi}(T_0)$ for $T_0 \simeq \pi$ yields

$$\begin{aligned}\mathcal{D}_f\left(\nu,\pi\right) &= \sup_{T\neq 0}\left\{\mathbb{E}_\nu\left[f'(T)\right] + \mathbb{E}_\pi\left[-Tf'(T) + (T-1)f'(\mathbb{E}_\pi T) + f(T) - f(\mathbb{E}_\pi T)\right]\right\}\\ &= \sup_{T\neq 0}\left\{\mathbb{E}_\nu\left[f'(T)\right] - \left(\mathbb{E}_\pi\left[f^*\circ f'(T)\right] - f^*\circ f'\left(\mathbb{E}_\pi T\right) + f'(\mathbb{E}_\pi T)\right)\right\}\end{aligned}$$

where we used $xf'(x) - f(x) = f^*\circ f'(x)$ twice to obtain the second equality. Through the change of variable $h(\phi) = f'\circ T\left(\phi\right)$ which maps π-integrable random variables to the set of functions

$$\mathcal{T} := \left\{h \mid \forall\phi\in \mathrm{Supp}\left(\pi\right),\ f'(0)\leq h(\phi)\leq f'(\infty),\ \int |f^{*\prime}\circ h|\mathrm{d}\pi < \infty\right\},$$

we find that $\mathcal{D}_f(\nu,\pi)$ is equal to (13) taking the supremum on $h\in\mathcal{T}$. We can relax the condition that $h\geq f'(0)$ since $f^{*\prime}\circ h = f^{*\prime}(\max(f'(0),h))$ and $f^*\circ h = f^*(\max(f'(0),h))$. Whenever $f^{*\prime}(h)$ is not π-integrable, $\mathbb{E}_\nu\left[h\right] - \tilde{\mathcal{D}}^*_{f,\pi}(h)$ equals $\mathbb{E}_\nu[h] - \mathbb{E}_\pi[f^*(h)]$ which we can bound, using (14), by $\mathcal{D}_f(\nu,\pi)$. Thus

$$\mathcal{D}_f(\nu,\pi) = \sup_{h\in\mathcal{T}}\left\{\mathbb{E}_\nu\left[h\right] - \tilde{\mathcal{D}}^*_{f,\pi}(h)\right\} = \sup_{\substack{h<f'(\infty),\\ \mathbb{E}_\pi[h]<\infty}}\left\{\mathbb{E}_\nu\left[h\right] - \tilde{\mathcal{D}}^*_{f,\pi}(h)\right\} = \sup_{h}\left\{\mathbb{E}_\nu\left[h\right] - \tilde{\mathcal{D}}^*_{f,\pi}(h)\right\}.$$

■

Proof [Theorem 3] Let us first establish the upperbounds (7) and (8) of the Legendre transform. The proof of (7) is straightforward, using Young-Fenchel's inequality pointwise,

$$\int h\mathrm{d}\nu = \int h\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\mathrm{d}\pi \leq \int f^*(h)\mathrm{d}\pi + \int f\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right)\mathrm{d}\pi. \tag{14}$$

This would be sufficient for our purpose, but for completeness sake, we show that this implies that $\mathbb{E}_\pi\left[f^*(h)\right] \geq \mathcal{D}^*_{f,\pi}(h)$. Indeed, from the preceding inequality, it follows that:

$$\mathcal{D}^*_{f,\pi}(h) = \sup_{\nu\in\Pi}\left(\mathbb{E}_\nu\left[h\right] - \mathcal{D}_f(\nu,\pi)\right) \leq \mathbb{E}_\pi\left[f^*(h)\right].$$

The proof of (8) starts with lemma 4: we have that for all ν such that $\mathcal{D}_f(\nu,\pi) < \infty$, for all π-measurable h,

$$\mathbb{E}_\nu[h] \leq \tilde{\mathcal{D}}^*_{f,\pi}(h) + \mathcal{D}_f(\nu,\pi). \tag{15}$$

We can relax the condition $\mathcal{D}_f(\nu,\pi) < \infty$ since when it is not the case, the bound remains correct, though trivial.

It only remains to see how the bounds behave on the transform $f \to \lambda^{-1} f + c f_\delta$ for $\lambda \in \mathbb{R}_+^*$, $c \in \mathbb{R}$. Let us first remark that this transform does not change the concavity of $1/f''$ since $f_\delta'' = 0$ and $\lambda > 0$. Noting $g = \lambda^{-1} f + c f_\delta$, we have $g^*(t) = c + \lambda^{-1} f^*(\lambda(t-c))$, $g^{*\prime}(t) = f^{*\prime}(\lambda(t-c))$ and $g'(x) = \lambda^{-1} f'(x) + c$.

Therefore,

$$\tilde{\mathcal{D}}_{g,\pi}^*(h) = \lambda^{-1} \tilde{\mathcal{D}}_{f,\pi}^*(\lambda(h-c)) + c \quad \text{and} \quad \mathcal{L}_{g,\pi} = \lambda^{-1} \mathcal{L}_{g,\pi}(\lambda(h-c)) + c$$

and thus, since $\mathcal{D}_g = \lambda^{-1} \mathcal{D}_f$, both

$$\mathbb{E}_\nu[h] \leq \lambda^{-1} \tilde{\mathcal{D}}_{f,\pi}^*(\lambda(h-c)) + c + \lambda^{-1} \mathcal{D}_f(\nu, \pi)$$

and its analogue using $\mathcal{L}_{f,\pi}$ hold. One can easily see that the same bounds would also arise by evaluating (14) and (15) for $\tilde{h} = \lambda(h-c)$. ■

Theorem 3 states that we can control the mean of a random variable over a probability measure ν from two terms : a measure of the distance between ν and π, and what is morally a moment of the random variable with respect to π. The more the f-divergence discriminates between ν and π, the weaker is the moment needed. On the other hand, if strong moment assumptions can be made on the random variable, one can control its mean over ν for a larger set of probability measures.

The bounds can be optimised on two degrees of freedom, controlling the mean and fluctuation of the random variable. The first bound is only valid on the strong condition that $1/f''$ is concave while the second bound is valid for all f-divergence. Let us remark that the condition $1/f''$ concave implies that $\liminf f^*(t)/t^2 > 0$ (a proof of this claim is deferred to Appendix A.2), and therefore (11) is not trivial only when the random variable has at least a moment of order 2. A non trivial bound of form (12) can be found for any random variable h which admits a moment $\mathbb{E}_\pi[g \circ h_+]$ for any g such that $\lim_{t\to\infty} g(t)/t = +\infty$.

Although the bounds can be optimised on two degrees of freedom, it might not be possible to apply this double optimisation procedure. We do not have clear arguments to favour optimising with respect to c over optimising with respect to λ or vice-versa. We note however that most of the bounds we examined proved easier to optimise on the scale parameter rather than on the positional degree of freedom.

Note that for all $g \in \mathcal{F}$, the reverse f-divergence $\nu, \mu \mapsto \mathcal{D}_g(\mu, \nu)$ is a f-divergence for $f_g(t) = t \times g\left(\frac{1}{t}\right)$. Therefore, Theorem 3 can also be used to obtain change of measure inequalities for the reverse f-divergence.

Remark 5 *Embedded in the bounds of Theorem 3 is the condition that $\lambda(h-c) < f'(\infty)$ π-almost surely, since whenever that is not the case, the bounds become trivial. Plainly speaking, whenever $f'(+\infty) \neq +\infty$, h needs to be upperbounded for the bounds to be useable. Whenever this is the case, we find it good practice to choose f such that $f'(\infty) = 0$, and as such, we can apply the bound to $\lambda(h - h_{\max} - c)$ for $\lambda > 0, c > 0$.*

Remark 6 *One can reinterpret the minimisation on λ and c for every bound of the form (2) in term of Legendre transforms. Indeed, for (12), define $\Gamma_{f,h,c,\pi} : \gamma \mapsto \gamma \mathcal{L}_{f,\pi}(\gamma^{-1}(h-c))$ for $\gamma \geq 0$, $+\infty$ for $\gamma < 0$. $\Gamma_{f,h,c,\pi}$ is convex[1], and we can rewrite (12) as*

$$\mathbb{E}_\nu[h] \leq c - \Gamma_{f,h,c,\pi}^*(-\mathcal{D}_f(\nu, \pi)).$$

1. Since $\lambda \mapsto \mathcal{L}_{f,\pi}(\lambda(h-c))$ is convex and for all F convex defined on $\mathbb{R}^+$, $t \mapsto tF(t^{-1})$ is convex.

On the other hand, noting $\beta_{f,h,\gamma,\pi}(c) = \gamma\mathcal{L}_{f,\pi}(\gamma^{-1}(h-c))$*,* β *is also convex and we can rewrite* (12) *as*

$$\mathbb{E}_\nu[h] \leq -\beta^*_{f,h,\gamma,\pi}(-1) + \gamma\mathcal{D}_f(\nu,\pi).$$

The same argument can be used for (11)*, although in that case,* $\Gamma_{f,h,c,\pi}$ *and* $\beta_{f,h,\gamma,\pi}$ *might not be convex.*

Remark 7 *So far, we have assumed that, for all* $x < 0$*,* $f(x) = +\infty$*. This basically introduces a threshold at* $f'(0)$*, in the sense that the upperbounds evaluated at* h *and at* $\max(h, f'(0))$ *are identical. This threshold can, for some* f*-divergences, be problematic when trying to optimise the bounds on* λ *and* c*. A possible work around for* (12) *is to extend* f *to some part of* $\mathbb{R}^-$ *in such a way as to keep* $\tilde{f}$ *convex, and to use* $\tilde{f}^*$ *rather than* f^**. As* $f^* \geq \tilde{f}^*$*, the bound remains valid. Note that this is always possible to extend* f *whenever* $f'(0) > -\infty$ *and* $\infty > f''(0) > 0$ *through* $\tilde{f}(t) = f(0) + tf'(0) + \frac{t^2}{2}f''(0)$ *for* $t < 0$*. This gives some flexibility between tractability of expressions and tightness of the bound.*

4. Application to Learning Theory

4.1. Some more PAC-Bayesian bounds

We now explore how Theorem 3 can be leveraged in learning theory. We consider two random variables X, Y on a product space $\mathcal{X} \times \mathcal{Y}$, drawn from a joint (unknown) probability measure $\mathbb{P}$. We now set (Φ, Σ_Φ) to be a set of predictors ϕ, which are measurables functions from $\left(\mathcal{X}, \mathbb{P}_{|X}\right)$ to $\left(\mathcal{Y}, \mathbb{P}_{|Y}\right)$. The reference probability measure π is now called prior in analogy with the usual Bayesian prior (see Guedj, 2019, for a discussion on the terminology in PAC-Bayes). We choose a measurable loss function $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$. For an observation (x, y), the empirical risk of a predictor $\phi \in \Phi$ is defined as

$$r(\phi, (x, y)) := \ell(\phi(x), y),$$

and the risk of $\phi \in \Phi$ is given by

$$R(\phi) := \mathbb{E}_\mathbb{P}\left[r(\phi,\ (X, Y)\right].$$

The function $\phi \mapsto R(\phi) - r(\phi)$ is called the generalisation gap.

Applying Theorem 3 to $|R - r|$ yields elementary PAC-Bayesian bounds. Since we control $\mathbb{E}_\nu[|R - r|]$ in term of a moment of $|R - r|$ for π and the distance between ν and π, making use of the fact that the first term is independant of ν, we obtain that, with probability higher than $1 - \varepsilon$,

$$\mathbb{E}_\nu[R] \leq \mathbb{E}_\nu[r] + \varepsilon^{-1}\mathbb{E}_\mathbb{P}\mathbb{E}_\pi\left[f^* \circ |R - r|\right] + \mathcal{D}_f(\nu, \pi).$$

Note that we consider a setting with only one observation. This enables us to apply bounds to a wide variety of models, for instance n-i.i.d. observations (with $(\mathcal{X} \times \mathcal{Y}) = (\mathcal{X}_0 \times \mathcal{Y}_0)^n$ and $\mathbb{P} = \mathbb{P}_0^n$), but also Markov chains or other non independant data. To obtain tight PAC-Bayes bound, some concentration inequality is needed on $\mathbb{E}_\pi\left[f^* \circ |R - r|\right]$. Therefore the rate depends on the structure of the data.

It is possible to apply our proposed change of measure inequalities to other measure of the distance between the empirical risk and true risk. For instance, Bégin et al. (2016) introduced the

notion of Δ-function. For Δ convex, they give PAC-bounds for $\mathbb{E}_\nu[\Delta(R,r)]$ using Kullback–Leibler penalisation. Such bounds can be obtained in our framework by applying the bounds with $h = \Delta(R,r)$.

Note that Remark 2 all but imply that the generalisation gap, and therefore the loss, should have f^* moments to obtain non trivial PAC bounds. Notably, whenever $f'(\infty) < \infty$, the loss should be bounded. On the other hand, we aim to help design tailored-made bounds for a diversity of applications. The Legendre transform approach enables one to choose a penalisation suited for the assumptions one is willing to make on the data-generating distribution. Indeed, the choice of the convex function f defining the f-divergence implies a moment assumption on the generalisation gap $R - r$. This moment assumption can afterwards be reversered engineered to a choice of adequate penalisation. As a rule of thumb, the flatter the chosen f, the stronger the assumptions required, with $f(x) = x\log(x)$ requiring exponential moment bounds, and $f(x) \simeq x^p$ only requiring finite moment bounds.

4.2. Change of measure inequalities for standard f-divergence

We apply Theorem 1 to the most popular f-divergences found in the literature. Table 1 presents a summary of all the resulting change of measure inequalities. Note that the optimal values of λ and the Legendre transform f^* are gathered, when available, in Table 2 which is deferred to Appendix A.

Equation 11 recovers the exact Legendre transform of the Kullback–Leibler divergence. The bound is also quite tight for Pearson's χ^2- divergence. Indeed, (1) implies that for h such that $h > 0$ and $\mathbb{E}_\pi[h] \leq 1$, $\mathcal{D}^*_{f,\pi}(h) = \tilde{\mathcal{D}}^*_{f,\pi}(h)$.

The bounds presented in Table 1 are coherent with those obtained independently by Ohnishi and Honorio (2021). The last three are, to the best of our knowledge, the first change of measure inequalities obtained for these f-divergences.

For Kullback–Leibler, we recover the change of measure inequality established by Csiszár (1975) and Donsker and Varadhan (1975). That bound is the starting point of the proof of the general PAC bound established by Bégin et al. (2016), which recovers bounds obtained in Langford and Seeger (2001), McAllester (2003), Catoni (2007) and Alquier et al. (2016). We leave for future work the optimisation on λ (Dalalyan and Tsybakov, 2012 provide some preliminary heuristics).

For power p divergences with $p > 1$, only moments of order $\frac{p}{p-1} = q$ for h are needed rather than exponential moments, considerably lessening the assumptions needed on the loss l and the underlying data distribution. When $1 < p \leq 2$, the bounds we propose improve on those obtained in Alquier and Guedj (2018). Indeed, these last bounds exactly match those we obtain through (12). The bounds for $1 < p \leq 2$ can be slightly simplified, noticing that

$$\mathbb{E}_\pi\left[(h_+)^q\right] - \mathbb{E}_\pi\left[(h_+)^{\frac{q}{p}}\right]^p \leq \mathbb{E}_\pi\left[h^q\right] - \mathbb{E}_\pi\left[h^{\frac{q}{p}}\right]^p.$$

For all the remaining f-divergences, $f'(\infty) < \infty$. Therefore the Legendre transform of these f-divergences only take real values on bounded functions h. The bounds are of the form $h_{\max}$ minus a term controlling how much probability mass can be put on or close to $h^{-1}(h_{\max})$, knowing the distance between ν and π and some moment of h on π.

For the power divergences with $0 < p < 1$, let us remark that when $\mathcal{D}_{f_p}(\nu,\pi) \to 0$, the bound is optimised for $c \to \infty$, while when $\mathcal{D}_{f_p}(\nu,\pi) \to 1$, the bound is optimised for $c \to 0$. A similar behavior is observed for power divergences with $p < 0$. It seems important to pick adequatly $c\left(\mathcal{D}_{f_p}(\nu,\pi)\right)$ if one wishes to obtain tight bounds for all ν.

Table 1: Bounds for typical f-divergence. We denote $\max(h, 0) =: h_+$. For power-divergences, q is such that $\frac{1}{q} + \frac{1}{p} = 1$. For Lin's measure, f_θ is given by $f_\theta(t) = (\theta t \log(t\theta) - (\theta t + 1 - \theta) \log(\theta t + 1 - \theta) - \theta \log(\theta)$.

f-div	$f(t)$	$\mathbb{E}_\nu [h] \leq \dots$	c, λ, γ
KL	$t \log(t)$	$\lambda^{-1} \log \mathbb{E}_\pi [\exp(\lambda h)] + \lambda^{-1}\mathrm{KL}(\nu, \pi)$	$\lambda > 0$
Power-p, $1 < p \leq 2$	$t^p - 1$	$\mathbb{E}_\pi \left[h_+^{q-1}\right]^{p-1} + \left(\mathbb{E}_\pi \left[h_+^q\right] - \mathbb{E}_\pi \left[h_+^{\frac{q}{p}}\right]^p\right)^{\frac{1}{q}} \mathcal{D}_{f_p}(\nu, \pi)^{\frac{1}{p}}$	
Power-p, $1 < p$	$t^p - 1$	$c + \mathbb{E}_\pi \left[(h - c)_+^q\right]^{\frac{1}{q}} \left(1 + \mathcal{D}_{f_p}(\nu, \pi)\right)^{\frac{1}{p}}$	$c \in \mathbb{R}$
Pearson χ^2	$t^2 - 1$	$\mathbb{E}_\pi [h_+] + \mathbf{Var}_\pi [h]^{\frac{1}{2}} \chi^2(\nu, \pi)^{\frac{1}{2}}$	
Power-p, $0 < 1 < p$	$1 - t^p$	$h_{\max} + c - \mathbb{E}_\pi \left[(h_{\max} - h + c)^q\right]^{\frac{1}{q}} \left(1 - \mathcal{D}_{f_p}(\nu, \pi)\right)^{\frac{1}{p}}$	$c > 0$
Power-p, $p < 0$	$t^p - 1$	$h_{\max} + c - \mathbb{E}_\pi \left[(h_{\max} - h + c)^q\right]^{\frac{1}{q}} \left(1 + \mathcal{D}_{f_p}(\nu, \pi)\right)^{\frac{1}{p}}$	$c \geq 0$
TV	$\frac{\lvert t-1 \rvert}{2}$	$h_{\max} + \mathbb{E}_\pi [\max(h - h_{\max}, -\gamma)] + \gamma TV(\nu, \pi)$	$\gamma > 0$
Squared Hellinger	$1 - \sqrt{t}$	$h_{\max} + c - \left(1 - H^2(\nu, \pi)\right)^2 \mathbb{E}_\pi \left[\frac{1}{h_{\max} - h + c}\right]^{-1}$	$c > 0$
Reverse Pearson	$t^{-1} - 1$	$h_{\max} + c - \frac{\mathbb{E}_\pi \left[\sqrt{c + h_{\max} - h}\right]^2}{1 + \chi^2(\pi, \nu)}$	$c > 0$
Reverse KL	$-\log(t)$	$h_{\max} + c - \exp\left(\mathbb{E}\left[\log\left(h_{\max} - h + c\right)\right] - \mathrm{KL}(\pi, \nu)\right)$	$c > 0$
Lin's measure, $0 < \theta < 1$	$f_\theta(t)$	$h_{\max} + c - \lambda^{-1}(1 - \theta)\mathbb{E}_\pi \left[\log\left(1 - \exp\left(\frac{\lambda(h - h_{\max} - c)}{\theta}\right)\right)\right]$ $+\lambda^{-1}\left[L_\theta(\nu, \pi) + (1 - \theta)\log(1 - \theta) - \theta \log(\theta)\right]$	$\lambda > 0$, $c > 0$
Jensen–Shannon	$f_{\theta=\frac{1}{2}}(t)$	$h_{\max} + c - \lambda^{-1}\mathbb{E}_\pi \left[\frac{1}{2} \log\left(1 - \exp\left(2\lambda(h - h_{\max} - c)\right)\right)\right]$ $+\lambda^{-1}\mathrm{JS}(\nu, \pi)$	$\lambda > 0$, $c > 0$
Vincze–Le Cam	$\frac{2-2t}{t+1}$	$2\left(h_{\max} + c\right) + \mathbb{E}_\pi[-h] - \frac{4\mathbb{E}_\pi \left[\sqrt{c + h_{\max} - h}\right]^2}{2 + \mathrm{VC}(\nu, \pi)}$	$c > 0$

For total variation, let us first remark that since the generator $f(x) = \frac{\lvert x-1 \rvert}{2}$ is not differentiable at $x = 1$, it can not be approximated by a sequence of convex functions such that $1/f_n''$ is concave[2]. It is possible to minimise the bound on c, but we could not compute the optimal scale parameter.

2. Whenever f' is not continuous at $x_0 > 0$, then f'' is a Dirac mass at x_0 and therefore $1/f''(x_0) = 0$. It follows that $1/f''$ has a local minima at x_0 since $f'' \geq 0$, and therefore it can not be concave for any reasonable approximation.

Vincze-Le Cam's bound somewhat stands out as it involves $2h_{\max}$ rather than $h_{\max}$. This is explained by the fact that the bound is not derived directly from Theorem 3, but results from remark 7, extending f to $t \in (-1, 0)$ by $f(t) = \frac{2-2t}{t+1}$.

5. Perspectives

A new look at hostile data. Alquier and Guedj (2018) went beyond the usual bounded or bounded exponential moment requirement on the generalisation gap and obtained bounds on the more reasonable assumption of bounded moments. The Legendre transform framework we propose requires at least bounded order 1 moment on h. We give here bounds on the assumption of bounded $\mathbb{E}_\pi[h \log(h)]$. For $h \geq 0$,

$$\mathbb{E}_\nu [h] \leq \int h \log(h) \mathrm{d}\pi - \int h \mathrm{d}\pi \log \left(\int h \mathrm{d}\pi \right) + \log \left(\int \exp \left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi} \right) \mathrm{d}\pi \right) \int h \mathrm{d}\pi.$$

This limits the exploration of Π to probabilities with bounded exponential divergence from the prior.

Legendre transform of the entropy. We might be willing to tradeoff some tighteness on the bound for more tractable expressions. One could for instance decide to explore the Legendre transform of the f-entropy, which is defined as

$$\mathcal{E}_{f,\pi} (h) = \mathbb{E}_\pi [f \circ h] - f (\mathbb{E}_\pi h) .$$

The f-entropy collapses to the f-divergence between μ and ν if $h = \frac{\mathrm{d}\nu}{\mathrm{d}\pi}$, since $f(1) = 0$. We can not find any reason why the f-entropy should be convex on h. Nevermind: upperbounding $\mathcal{E}^*_{f,\pi}$ amounts to upperbounding $\mathcal{D}^*_{f,\pi}$. More generally, any extension of the f-divergence to a larger space can be used to upperbound $\mathcal{D}^*_{f,\pi}$.

Finding approximately optimal λ and c. As discussed in Section 4, an appropriate choice of c and λ is necessary to obtain tight inequalities. In most cases, we could not explicity compute which values are optimal, especially for c. Getting some theoretical or pratical insight on how to pick these parameters in such a way as to obtain nearly optimal bounds is an exciting future avenue.

Acknowledgments

The authors are grateful to Roman Moscoviz, Gilles Faÿ and Marc-Antoine Giuliani for helpful insights and discussions. Antoine Picard acknowledges support from SUEZ. Benjamin Guedj acknowledges partial support by the U.S. Army Research Laboratory and the U.S. Army Research Office, and by the U.K. Ministry of Defence and the U.K. Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/R013616/1. Benjamin Guedj acknowledges partial support from the French National Agency for Research, grants ANR-18-CE40-0016-01 and ANR-18-CE23-0015-02.

References

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. 2021. URL https://www.arxiv.org/abs/2110.11216.

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

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

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian Bounds based on the Rényi Divergence. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 435–444, Cadiz, Spain, 09–11 May 2016. PMLR. URL https://proceedings.mlr.press/v51/begin16.html.

Felix Biggs and Benjamin Guedj. Differentiable PAC-Bayes objectives with partially aggregated neural networks. *Entropy*, 23(10):1280, 2021. doi: 10.3390/e23101280. URL https://doi.org/10.3390/e23101280.

Felix Biggs and Benjamin Guedj. On margins and derandomisation in PAC-Bayes. In *AISTATS*, 2022a. URL https://www.arxiv.org/abs/2107.03955.

Felix Biggs and Benjamin Guedj. Non-vacuous generalisation bounds for shallow neural networks. 2022b. URL https://arxiv.org/abs/2202.01627.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Théophile Cantelobre, Benjamin Guedj, María Pérez-Ortiz, and John Shawe-Taylor. A PAC-Bayesian perspective on structured prediction with implicit loss embeddings. 2020. URL https://arxiv.org/abs/2012.03780.

Olivier Catoni. A PAC-Bayesian approach to adaptive classification. *preprint*, 840, 2003.

Olivier Catoni. *Statistical Learning Theory and Stochastic Optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001*. Springer, 2004.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes - Monograph Series*. Institute of Mathematical Statistics, 2007.

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

Arnak S. Dalalyan and Alexandre B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5):1423–1443, 2012. doi: 10.1016/j.jcss.2011.12.023. URL http://hal.archives-ouvertes.fr/docs/00/45/68/06/PDF/HAL_EWA_LMC.pdf.

MD Donsker and SRS 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.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *Conference on Uncertainty in Artificial Intelligence 33.*, 2017.

Benjamin Guedj. A Primer on PAC-Bayesian Learning. In *Proceedings of the second congress of the French Mathematical Society*, 2019. URL https://arxiv.org/abs/1901.05353.

Benjamin Guedj and Louis Pujol. Still no free lunches: the price to pay for tighter PAC-Bayes bounds. *Entropy*, 23(11), 2021. ISSN 1099-4300. doi: 10.3390/e23111529. URL https://www.mdpi.com/1099-4300/23/11/1529.

Maxime Haddouche, Benjamin Guedj, Omar Rivasplata, and John Shawe-Taylor. PAC-Bayes unleashed: generalisation bounds with unbounded losses. *Entropy*, 23(10):1330, 2021.

Matthew J. Holland. PAC-Bayes under potentially heavy tails. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2711–2720, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3a20f62a0af1aa152670bab3c602fee

John Langford and Matthias Seeger. Bounds for averaging classifiers. 2001. URL http://www.cs.cmu.edu/~jcl/papers/averaging/averaging_tech.pdf.

Gaël Letarte, Pascal Germain, Benjamin Guedj, and Francois Laviolette. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6872–6882. Curran Associates, Inc., 2019.

David A McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational Learning Theory*, pages 230–234. ACM, 1998.

David A McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational Learning Theory*, pages 164–170. ACM, 1999.

David A McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

Zakaria Mhammedi, Benjamin Guedj, and Robert C. Williamson. PAC-Bayesian bound for the conditional value at risk. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems [NeurIPS] 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d02e9bdc27a894e882fa0c9055c9972

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

Kento Nozawa, Pascal Germain, and Benjamin Guedj. PAC-Bayesian contrastive unsupervised representation learning. In *Conference on Uncertainty in Artificial Intelligence [UAI]*, 2020. URL https://proceedings.mlr.press/v124/nozawa20a.html.

Yuki Ohnishi and Jean Honorio. Novel change of measure inequalities with applications to PAC-Bayesian bounds and Monte Carlo estimation. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 1711–1719. PMLR, 2021. URL http://proceedings.mlr.press/v130/ohnishi21a.html.

Maria Perez-Ortiz, Omar Rivasplata, Benjamin Guedj, Matthew Gleeson, Jingyu Zhang, John Shawe-Taylor, Miroslaw Bober, and Josef Kittler. Learning PAC-Bayes priors for probabilistic neural networks. 2021a. URL https://arxiv.org/abs/2109.10304.

Maria Perez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvari. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227):1–40, 2021b. URL http://jmlr.org/papers/v22/20-879.html.

Tim van Erven and Peter Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500. URL https://doi.org/10.1109/TIT.2014.2320500.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a PAC-Bayesian compression approach. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=BJgqqsAct7.

Appendix A. Proofs

A.1. Proofs for Section 4

The proof for each bound follows the same pattern: for each f-divergence, compute f^*, check whether $(1/f'')$ is concave, then apply accordingly either (11) or (12) to obtain:

$$\int h \mathrm{d}\nu \leq c + \lambda^{-1}\mathrm{B}(\lambda(h-c)) + \lambda^{-1}\mathcal{D}_f(\nu,\pi).$$

Then optimise on λ and c whenever feasible. We therefore sum up the proofs in Table 2 which gives the information listed above.

The Kullback–Leibler, power divergence for $1 < p \leq 2$ and Pearson χ^2 satisfy $1/f''$ concave, and we therefore use (11). All the other bounds use (12). For the total variation, it is simple to see that to get non trivial bounds, we need to pick $c \leq \frac{1}{2} - h_{\max}$. Diminishing c to $c - \delta c$ decreases the integral by at most δc (the threshold can only dampen the decrease), while the other term increases by δc. This implies that $c^* = \frac{1}{2} - h_{\max}$.

For the Vincze–Le Cam divergence, we are in the situation described in Theorem 7. The upper-bound obtained through $\tilde{f}$ is much more tractable than the one obtained through f, and in particular,

Table 2: f^*, λ^* optimising change of measure inequalities

f-div	$f^*(t) = \dots$	λ^*
KL	$\exp(t-1)$	
Power-p, $1 < p \leq 2$	$\frac{p^{1-q}}{q}\max(t,0)^q + 1$	$p\left(\frac{\mathcal{D}_{f_p}(\nu,\pi)}{\mathbb{E}_\pi\left[h_+^q\right]-\mathbb{E}_\pi\left[h_+^{\frac{q}{p}}\right]^p}\right)^{\frac{1}{q}}$
Power-p, $2 < p$	$\frac{p^{1-q}}{q}\max(t,0)^q + 1$	$p\left(\frac{1+\mathcal{D}_{f_p}(\nu,\pi)}{\mathbb{E}_\pi[h^q]}\right)^{\frac{1}{q}}$
Pearson χ^2	$\sqrt{2}\max(t,0)^2 + 1$	$2\left(\frac{\mathcal{D}_{f_2}(\nu,\pi)}{\mathbf{Var}_\pi[h_+]}\right)^{\frac{1}{2}}$
Power-p, $0 < p < 1$	$\begin{cases} -1 + \frac{p^{1-q}}{-q}(-t)^q & t < 0 \\ +\infty & \text{else} \end{cases}$	$p\left(\frac{1-\mathcal{D}_{f_p}(\nu,\pi)}{\mathbb{E}_\pi[(-h)^q]}\right)^{\frac{1}{q}}$
Squared Hellinger	$\begin{cases} -1 + \frac{1}{4}(-t)^{-1} & t < 0 \\ +\infty & t \geq 0 \end{cases}$	$\frac{\mathbb{E}_\pi\left[(-h)^{-1}\right]}{2-2H^2(\nu,\pi)}$
Power-p, $p < 0$	$\begin{cases} 1 - \frac{(-p)^{1-q}}{q}(-t)^q & t \leq 0 \\ +\infty & \text{else} \end{cases}$	$(-p)\left(\frac{1+\mathcal{D}_{f_p}(\nu,\pi)}{\mathbb{E}_\pi[(-h)^q]}\right)^{\frac{1}{q}}$
Reverse Pearson	$\begin{cases} 1 - 2(-t)^{1/2} & t \leq 0 \\ +\infty & \text{else} \end{cases}$	$\left(\frac{1+\chi^2(\pi,\nu)}{\mathbb{E}_\pi\left[\sqrt{(-h)}\right]}\right)^2$
Total Variation	$\begin{cases} -\frac{1}{2} & t < -\frac{1}{2} \\ t & -\frac{1}{2} \leq t \leq \frac{1}{2} \\ +\infty & t > \frac{1}{2} \end{cases}$	
Reverse KL	$\begin{cases} -(1+\log(-t) & t < 0 \\ +\infty & \text{else} \end{cases}$	$\exp\left(\mathbf{KL}(\pi,\nu) - \mathbb{E}_\pi\left[\log(-h)\right]\right)$
Lin's measure, $0 < \theta < 1$	$(1-\theta)\log\left(\frac{1-\theta}{1-\theta\exp(t\theta^{-1})}\right)$	
Jensen–Shannon	$-\frac{1}{2}\log\left(2 - e^{2t}\right)$	
Vincze–Le Cam	$\begin{cases} -2 & t \leq -4 \\ -4\sqrt{-t} - t + 2 & -4 \leq t \leq 0 \\ +\infty & \text{else} \end{cases}$	$\left(\frac{1+\frac{1}{2}VC(\nu,\pi)}{\mathbb{E}_\pi[\sqrt{-h}]}\right)^2$

it can be optimised on the scale parameter λ. It is this bound through $\tilde{f}$ which we use to obtain the final bound. Using the Legendre transform of $f_{|\mathbb{R}_+}$ yields this tighter, though somewhat unusable in practice, inequality for all $C \geq h_{\max}$, $\lambda > 0$

$$\mathbb{E}_\nu[h] \leq C + \lambda^{-1}\mathbb{E}_\pi\left[\mathbf{1}\left[h \geq C - \frac{4}{\lambda}\right]\left(4 - 4\sqrt{\lambda(C-h)} + \lambda(C-h)\right)\right] + \lambda^{-1}\left(\mathbf{VC}(\nu,\pi) - 2\right).$$

A.2. Other proofs

$1/f''$ concave implies $\liminf f^*(t)/t^2 > 0$. Since $1/f''$ is concave and positive on $\mathbb{R}_+^*$, it follows that $1/f''$ is non decreasing. There thus exists $\alpha > 0$, t_0 such that for all $t > t_0$, $1/f''(t) > \alpha$. Therefore $\exists\, \beta_0 > 0$ such that for all $t > t_0$, $f'(t) \leq \alpha^{-1}(t - t_0) + \beta_0$, which implies that all $t > t_1$, $f^{*\prime}(t) \geq \alpha^{-1}t + \beta_1$. Integrating ends the proof.

Non trivial bounds with weak moment assumption. We assume that

$$\mathbb{E}_\pi[(g \circ h)_+] < +\infty, \tag{16}$$

with g such that $\lim_{t\to+\infty} g(t)/t = +\infty$. Denote g_c the lower convex enveloppe of g. Then define $\tilde{g}_c : t \mapsto g_c(t) - t\max(0, \inf g_c')$. $\tilde{g}$ still satisfies $\lim_{t\to+\infty} \tilde{g}(t)/t = +\infty$ and (16). Then apply (12) using $f = \tilde{g}_c^* - \tilde{g}_c^*(1)$, which is non infinite on $\mathbb{R}_+^*$. This bound is not trivial for h on the condition that $\mathcal{D}_f(\nu, \pi) < \infty$. Since f is real valued on $\mathbb{R}_+^*$, the set $\{\nu \mid \nu \in \Pi, \mathcal{D}_f(\nu, \pi) < \infty\}$ is not reduced to $\{\pi\}$.