Simpler PAC-Bayesian Bounds for Hostile Data

Pierre Alquier, Benjamin Guedj

Simpler PAC-Bayesian Bounds for Hostile Data

Pierre Alquier* & Benjamin Guedj†

May 23, 2019

Abstract

PAC-Bayesian learning bounds are of the utmost interest to the learning community. Their role is to connect the generalization ability of an aggregation distribution ρ to its empirical risk and to its Kullback-Leibler divergence with respect to some prior distribution π. Unfortunately, most of the available bounds typically rely on heavy assumptions such as boundedness and independence of the observations. This paper aims at relaxing these constraints and provides PAC-Bayesian learning bounds that hold for dependent, heavy-tailed observations (hereafter referred to as *hostile data*). In these bounds the Kullack-Leibler divergence is replaced with a general version of Csiszár's f-divergence. We prove a general PAC-Bayesian bound, and show how to use it in various hostile settings.

1 Introduction

Learning theory can be traced back to the late 60s and has attracted a great attention since. We refer to the monographs Devroye et al. (1996) and Vapnik (2000) for a survey. Most of the literature addresses the simplified case of i.i.d observations coupled with bounded loss functions. Many bounds on the excess risk holding with large probability were provided - these bounds are refered to as PAC learning bounds since Valiant (1984).[1]

In the late 90s, the PAC-Bayesian approach was pioneered by Shawe-Taylor and Williamson (1997) and McAllester (1998, 1999). It consists of producing PAC bounds for a specific class of Bayesian-flavored estimators. Similar to classical PAC results, most PAC-Bayesian bounds have been obtained with bounded loss functions (see Catoni, 2007, for some of the most accurate results). Note that Catoni (2004) provides bounds for unbouded loss, but still under very strong exponential moment assumptions. Different types of PAC-Bayesian bounds were proved in very various models Seeger (2002); Langford and Shawe-Taylor (2002); Seldin and Tishby (2010); Seldin et al. (2012, 2011); Guedj and Alquier (2013);

*CREST, ENSAE, Université Paris Saclay, pierre.alquier@ensae.fr. This author gratefully acknowledges financial support from the research programme *New Challenges for New Data* from LCL and GENES, hosted by the *Fondation du Risque*, from Labex ECODEC (ANR-11-LABEX-0047) and from Labex CEMPI (ANR-11-LABX-0007-01).

†Modal project-team, Inria, benjamin.guedj@inria.fr.

[1] PAC stands for Probably Approximately Correct.

Bégin et al. (2016); Alquier et al. (2016); Oneto et al. (2016) but the boundedness or exponential moment assumptions were essentially not improved in these papers.

The relaxation of the exponential moment assumption is however a theoretical challenge, with huge practical implications: in many applications of regression, there is no reason to believe that the noise is bounded or sub-exponential. Actually, the belief that the noise is sub-exponential leads to an overconfidence in the prediction that is actually very harmful in practice, see for example the discussion in Taleb (2007) on finance. Still, thanks to the aforementionned works, the road to obtain PAC bounds for bounded observations has now become so nice and comfortable that it might refrain inclination to explore different settings.

Regarding PAC bounds for heavy-tailed random variables, let us mention three recent approaches.

- Using the so-called *small-ball property*, Mendelson and several co-authors developed in a striking series of papers tools to study the Empirical Risk Minimizer (ERM) and penalized variants without an exponential moment assumption: we refer to their most recent works (Mendelson, 2015; Lecué and Mendelson, 2016). Under a quite similar assumption, Grünwald and Mehta (2016) derived PAC-Bayesian learning bounds ("empirical witness of badness" assumption). Other assumptions were introduced in order to derive fast rates for unbounded losses, like the multiscale Bernstein assumption Dinh et al. (2016).

- Another idea consists in using robust loss functions. This leads to better confidence bounds than the previous approach, but at the price of replacing the ERM by a more complex estimator, usually building on PAC-Bayesian approaches (Audibert and Catoni, 2011; Catoni, 2012; Oliveira, 2013; Giulini, 2015; Catoni, 2016).

- Finally, Devroye et al. (2015), using median-of-means, provide bounds in probability for the estimation of the mean without exponential moment assumption. It is possible to extend this technique to more general learning problems Minsker (2015); Hsu and Sabato (2016); Lugosi and Mendelson (2016); Guillaume and Matthieu (2017); Lugosi and Mendelson (2017).

Leaving the well-marked path of bounded variables led the authors to sophisticated and technical mathematics, but in the end they obtained rates of convergence similar to the ones in bounded cases: this is highly valuable for the statistical and machine learning community.

Regarding dependent observations, like time series or random fields, PAC and/or PAC-Bayesian bounds were provided in various settings (Modha and Masry, 1998; Steinwart and Christmann, 2009; Mohri and Rostamizadeh, 2010; Ralaivola et al., 2010; Seldin et al., 2012; Alquier and Wintenberger, 2012; Alquier and Li, 2012; Agarwal and Duchi, 2013; Alquier et al., 2013; Kuznetsov and Mohri, 2014; Giraud et al., 2015; Zimin and Lampert, 2015; London et al., 2016). However these works massively relied on concentration inequalities for or limit theorems for time series Yu (1994); Doukhan (1994); Rio (2000); Kontorovich et al. (2008), for which boundedness or exponential moments are crucial.

This paper shows that a proof scheme of PAC-Bayesian bounds proposed by Bégin et al. (2016) can be extended to a very general setting, without independence nor exponential moments assumptions. We would like to stress that this approach is not comparable to the aforementionned work, and in particular it is technically far less sophisticated. However, while it leads to sub-optimal rates in many cases, it allows to derive PAC-Bayesian bounds in settings where no PAC learning bounds were available before: for example heavy-tailed time series.

Given the simplicity of the main result, we state it in the remainder of this section. The other sections are devoted to refinements and applications. Let ℓ denote a generic loss function. The observations are denoted $(X_1,Y_1),\dots,(X_n,Y_n)$. Note that we do not require the observations to be independent, nor indentically distributed. We assume that a family of predictors $(f_\theta,\theta\in\Theta)$ is chosen. Let $\ell_i(\theta)=\ell[f_\theta(X_i),Y_i]$, and define the (empirical) risk as

$$r_n(\theta)=\frac{1}{n}\sum_{i=1}^{n}\ell_i(\theta),$$
$$R(\theta)=\mathbb{E}\big[r_n(\theta)\big].$$

Based on the observations, the objective is to build procedures with a small risk R. While PAC bounds focus on estimators $\hat{\theta}_n$ that are obtained as functionals of the sample, the PAC-Bayesian approach studies an aggregation distribution $\hat{\rho}_n$ that depends on the sample. In this case, the objective is to choose $\hat{\rho}_n$ such that $\int R(\theta)\hat{\rho}_n(\mathrm{d}\theta)$ is small. In order to do so, a crucial point is to choose a reference probability measure π, often referred to as the *prior*. In Catoni (2007), the role of π is discussed in depth: rather than reflecting a prior knowledge on the parameter space Θ, it should serve as a tool to measure the complexity of Θ.

Let us now introduce the two following key quantities.

Definition 1. *For any function g, let*

$$\mathcal{M}_{g,n}=\int\mathbb{E}\big[g(|r_n(\theta)-R(\theta)|)\big]\pi(\mathrm{d}\theta).$$

Definition 2. *Let f be a convex function with $f(1)=0$. The f-divergence between two distributions ρ and π is defined by*

$$D_f(\rho,\pi)=\int f\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}\pi$$

when ρ is absolutely continous with respect to π, and $D_f(\rho,\pi)=+\infty$ otherwise.

Csiszár introduced f-divergences in the 60s, see his recent monograph Csiszár and Shields (2004, Chapter 4) for a survey.

We use the following notation for recurring functions: $\phi_p(x)=x^p$.Consequently $\mathcal{M}_{\phi_p,n}=\int\mathbb{E}(|r_n(\theta)-R(\theta)|^p)\pi(\mathrm{d}\theta)$. Thus $\mathcal{M}_{\phi_p,n}$ is a moment of order p. As for divergences, we denote the Kullback-Leibler divergence by $\mathcal{K}(\rho,\pi)=D_f(\rho,\pi)$ when $f(x)=x\log(x)$, and the chi-square divergence $\chi^2(\rho,\pi)=D_{\phi_2-1}(\rho,\pi)$.

Theorem 1. *Fix $p>1$, put $q=\frac{p}{p-1}$ and fix $\delta\in(0,1)$. With probability at least $1-\delta$ we have for any aggregation distribution ρ*

$$\left|\int R\mathrm{d}\rho-\int r_n\mathrm{d}\rho\right|\le\left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}}\big(D_{\phi_p-1}(\rho,\pi)+1\big)^{\frac{1}{p}}. \tag{1}$$

The main message of Theorem 1 is that we can compare $\int r_n \mathrm{d}\rho$ (observable) to $\int R \mathrm{d}\rho$ (unknown, the objective) in terms of two quantities: the moment $\mathcal{M}_{\phi_q,n}$ (which depends on the distribution of the data) and the divergence $D_{\phi_p{-}1}(\rho,\pi)$ (which will reveal itself as a measure of the complexity of the set Θ). The most important practical consequence is that we have, with probability at least $1-\delta$, for any probability measure ρ,

$$\int R \mathrm{d}\rho \leq \int r_n \mathrm{d}\rho + \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}} \left(D_{\phi_p{-}1}(\rho,\pi)+1\right)^{\frac{1}{p}}. \tag{2}$$

This is a strong incitement to define our aggregation distribution $\hat{\rho}_n$ as the minimizer of the right-hand side of (2). The core of the paper will discuss in details this strategy and other consequences of Theorem 1.

Proof of Theorem 1. Introduce $\Delta_n(\theta) := |r_n(\theta) - R(\theta)|$. We follow a scheme of proof introduced by Bégin et al. (2016) in the bounded setting. We adapt the proof to the general case:

$$\begin{aligned}
\left|\int R \mathrm{d}\rho - \int r_n \mathrm{d}\rho\right| &\leq \int \Delta_n \mathrm{d}\rho = \int \Delta_n \frac{\mathrm{d}\rho}{\mathrm{d}\pi}\mathrm{d}\pi \\
&\leq \left(\int \Delta_n^q \mathrm{d}\pi\right)^{\frac{1}{q}} \left(\int \left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)^p \mathrm{d}\pi\right)^{\frac{1}{p}} \text{ (Hölder ineq.)} \\
&\leq \left(\frac{\mathbb{E}\int \Delta_n^q \mathrm{d}\pi}{\delta}\right)^{\frac{1}{q}} \left(\int \left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)^p \mathrm{d}\pi\right)^{\frac{1}{p}} \text{ (Markov ineq., w. prob. } 1-\delta) \\
&= \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}} \left(D_{\phi_p{-}1}(\rho,\pi)+1\right)^{\frac{1}{p}}.
\end{aligned}$$

□

In Section 2 we discuss the divergence term $D_{\phi_p{-}1}(\rho,\pi)$. In particular, we derive an explicit bound on this term when ρ is chosen in order to concentrate around the ERM (empirical risk minimizer) $\hat{\theta}_{\mathrm{ERM}} = \arg\min_{\theta\in\Theta} r_n(\theta)$. This is meant to provide the reader some intuition on the order of magnitude of the divergence term. In Section 3 we discuss how to control the moment $\mathcal{M}_{\phi_q,n}$. We derive explicit bounds in various examples: bounded and unbounded losses, i.i.d and dependent observations. The most important result of the section is a risk bound for auto-regression with heavy-tailed time series, something new up to our knowledge. In Section 4 we come back to the general case. We show that it is possible to explicitely minimize the right-hand side in (2). We then show that Theorem 1 leads to powerful oracle inequalities in the various statistical settings discussed above, exhibiting explicit rates of convergence.

2 Calculation of the divergence term

The aim of this section is to provide some hints on the order of magnitude of the divergence term $D_{\phi_p{-}1}(\rho,\pi)$. We start with the example of a finite parameter space Θ. The following proposition results from straightforward calculations.

Proposition 1. *Assume that* $\mathrm{Card}(\Theta) = K < \infty$ *and that* π *is uniform on* Θ. *Then*

$$D_{\phi_p{-}1}(\rho,\pi) + 1 = K^{p-1} \sum_{\theta\in\Theta} \rho(\theta)^p.$$

A special case of interest is when $\rho = \delta_{\hat{\theta}_{\mathrm{ERM}}}$, the Dirac mass concentrated on the ERM. Then

$$D_{\phi_p - 1}(\delta_{\hat{\theta}_{\mathrm{ERM}}}, \pi) + 1 = K^{p-1}.$$

Then (1) in Theorem 1 yields the following result.

Proposition 2. *Fix $p > 1$, $q = \frac{p}{p-1}$ and $\delta \in (0,1)$. With probability at least $1-\delta$ we have*

$$R(\hat{\theta}_{\mathrm{ERM}}) \leq \inf_{\theta \in \Theta} \left\{ r_n(\theta) \right\} + K^{1-\frac{1}{p}} \left(\frac{\mathcal{M}_{\phi_q, n}}{\delta} \right)^{\frac{1}{q}}.$$

Remark that $D_{\phi_p - 1}(\rho, \pi)$ seems to be related to the complexity K of the parameter space Θ. This intuition can be extended to an infinite parameter space, for example using the empirical complexity parameter introduced in Catoni (2007).

Assumption 1. *There exists $d > 0$ such that, for any $\gamma > 0$,*

$$\pi\left\{ \theta \in \Theta : \left\{ r_n(\theta) \right\} \leq \inf_{\theta' \in \Theta} r_n(\theta') + \gamma \right\} \geq \gamma^d.$$

In many examples, d corresponds to the ambient dimension (see Catoni (2007) for a thorough discussion). In this case, a sensible choice for ρ, as suggested by Catoni, is $\pi_\gamma(\mathrm{d}\theta) \propto \pi(\mathrm{d}\theta) \mathbf{1}\left[r(\theta) - r_n(\hat{\theta}_{\mathrm{ERM}}) \leq \gamma \right]$ for γ small enough (in Section 4, we derive the consequences of Assumption 1 for other aggregation distributions). We have

$$D_{\phi_p - 1}(\pi_\gamma, \pi) + 1 \leq \gamma^{-d(p-1)}$$

and

$$\int r_n(\theta) \mathrm{d}\pi_\gamma \leq r_n(\hat{\theta}_{\mathrm{ERM}}) + \gamma$$

so Theorem 1 leads to

$$\int R \mathrm{d}\pi_\gamma \leq r_n(\hat{\theta}_{\mathrm{ERM}}) + \gamma + \gamma^{-\frac{d}{q}} \left(\frac{\mathcal{M}_{\phi_q, n}}{\delta} \right)^{\frac{1}{q}}.$$

An explicit optimization with respect to γ leads to the choice

$$\gamma = \left(\frac{d}{q} \frac{\mathcal{M}_{\phi_q, n}}{\delta} \right)^{\frac{1}{1+\frac{d}{q}}}$$

and consequently to the following result.

Proposition 3. *Fix $p > 1$, $q = \frac{p}{p-1}$ and $\delta \in (0,1)$. Under Assumption 1, with probability at least $1-\delta$ we have,*

$$\int R \mathrm{d}\pi_\gamma \leq \inf_{\theta \in \Theta} \left\{ r_n(\theta) \right\} + \left(\frac{\mathcal{M}_{\phi_q, n}}{\delta} \right)^{\frac{1}{1+\frac{d}{q}}} \left\{ \left(\frac{d}{q} \right)^{\frac{1}{1+\frac{d}{q}}} + \left(\frac{d}{q} \right)^{\frac{-\frac{d}{q}}{1+\frac{d}{q}}} \right\}.$$

So the bound is in $\mathcal{O}((\mathcal{M}_{\phi_q, n}/\delta)^{1/(1+d/q)})$. In order to understand the order of magnitude of the bound, it is now crucial to understand the moment term $\mathcal{M}_{\phi_q, n}$. This is the object of the next section.

3 Bounding the moments

In this section, we show how to control $\mathcal{M}_{\phi_q,n}$ depending on the assumptions on the data.

3.1 The i.i.d setting

First, let us assume that the observations (X_i, Y_i) are independent and identically distributed. In general, when the observations are possibly heavy-tailed, we recommend to use Theorem 1 with $q \leq 2$ (which implies $p \geq 2$).

Proposition 4. *Assume that*

$$s^2 = \int \mathrm{Var}[\ell_1(\theta)]\pi(\mathrm{d}\theta) < +\infty$$

then

$$\mathcal{M}_{\phi_q,n} \leq \left(\frac{s^2}{n}\right)^{\frac{q}{2}}.$$

As a conclusion for the case $q \leq 2 \leq p$, (1) in Theorem 1 becomes:

$$\int R\,\mathrm{d}\rho \leq \int r_n\,\mathrm{d}\rho + \frac{\left(D_{\phi_p^{-1}}(\rho,\pi)+1\right)^{\frac{1}{p}}}{\delta^{\frac{1}{q}}}\sqrt{\frac{s^2}{n}}.$$

Without further assumptions, this bound can not be improved as a function of n (as can be seen in the simplest case where $\mathrm{card}(\Theta) = 1$, by using the CLT).

Proof of Proposition 4.

$$\begin{aligned}
\mathcal{M}_{\phi_q,n} &= \int \mathbb{E}\left(|r_n(\theta) - \mathbb{E}[r_n(\theta)]|^2 \frac{q}{2}\right)\pi(\mathrm{d}\theta) \\
&\leq \left(\int \mathbb{E}\left(|r_n(\theta) - \mathbb{E}[r_n(\theta)]|^2\right)\pi(\mathrm{d}\theta)\right)^{\frac{q}{2}} \\
&\leq \left(\int \frac{1}{n}\mathrm{Var}[\ell_1(\theta)]\pi(\mathrm{d}\theta)\right)^{\frac{q}{2}} = \left(\frac{s^2}{n}\right)^{\frac{q}{2}}.
\end{aligned}$$

□

As an example, consider the regression setting with quadratic loss, where we use linear predictors: $X_i \in \mathbb{R}^k$, $\Theta = \mathbb{R}^k$ and $f_\theta(\cdot) = \langle \cdot, \theta \rangle$. Define a prior π on Θ such that

$$\tau := \int \|\theta\|^4 \pi(\mathrm{d}\theta) < \infty \tag{3}$$

and assume that

$$\kappa := 8[\mathbb{E}(Y_i^4) + \tau\mathbb{E}(\|X_i\|^4)] < \infty. \tag{4}$$

Then

$$\ell_i(\theta) = (Y_i - \langle \theta, X_i \rangle)^2 \leq 2\left[Y_i^2 + \|\theta\|^2\|X_i\|^2\right]$$

and so

$$\mathrm{Var}(\ell_i(\theta)) \leq \mathbb{E}(\ell_i(\theta)^2) \leq 8\mathbb{E}\left[Y_i^4 + \|\theta\|^4\|X_i\|^4\right].$$

Finally,

$$s^2 = \int \mathrm{Var}(\ell_i(\theta))\pi(\mathrm{d}\theta) \leq \kappa < +\infty.$$

We obtain the following corollary of (1) in Theorem 1 with $p = q = 2$.

Corollary 1. *Fix $\delta \in (0,1)$. Assume that π is chosen such that (3) holds, and assume that (4) also holds. With probability at least $1-\delta$ we have for any ρ*

$$\int R \mathrm{d}\rho \leq \int r_n \mathrm{d}\rho + \sqrt{\frac{\kappa[1+\chi^2(\rho,\pi)]}{n\delta}}.$$

Note that a similar upper bound was proved in Honorio and Jaakkola (2014), yet only in the case of the 0-1 loss (which is bounded). Also, note that the assumption on the moments of order 4 is comparable to the one in Audibert and Catoni (2011) and allow heavy-tailed distributions. Still, in our result, the dependence in δ is less good than in Audibert and Catoni (2011). So, we end this subsection with a study of the sub-Gaussian case (wich also includes the bounded case). In this case, we can use any $q \geq 2$ in Theorem 1. The larger q, the better will be the dependence with respect to δ.

Definition 3. *A random variable U is said to be sub-Gaussian with parameter σ^2 if for any $\lambda > 0$,*

$$\mathbb{E}\Big\{\exp\big[\lambda(U-\mathbb{E}(U))\big]\Big\} \leq \exp\left[\frac{\lambda^2\sigma^2}{2}\right].$$

Proposition 5 (Theorem 2.1 page 25 in Boucheron et al. (2013))**.** *When U is sub-Gaussian with parameter σ^2 then for any $q \geq 2$,*

$$\mathbb{E}\big[(U-\mathbb{E}(U))^q\big] \leq 2\left(\frac{q}{2}\right)!(2\sigma^2)^{\frac{q}{2}} \leq 2(q\sigma^2)^{\frac{q}{2}}.$$

A straighforward consequence is the following result.

Proposition 6. *Assume that, for any θ, $\ell_i(\theta)$ is sub-Gaussian with parameter σ^2 (that does not depend on θ), then $\frac{1}{n}\sum_{i=1}^n \ell_i(\theta)$ is sub-Gaussian with parameter σ^2/n and then, for any $q \geq 2$,*

$$\mathcal{M}_{\phi_q,n} \leq 2\left(\frac{q\sigma^2}{n}\right)^{\frac{q}{2}}.$$

As an illustration, consider the case of a finite parameter space, that is $\mathrm{card}(\Theta) = K < +\infty$. Following Proposition 2 and Proposition 6, we obtain for any $q \geq 2$ and $\delta \in (0,1)$, with probability at least $1-\delta$,

$$R(\hat{\theta}_{\mathrm{ERM}}) \leq \inf_{\theta\in\Theta}\{r_n(\theta)\} + \sigma\sqrt{\frac{q}{n}}\left(\frac{2K}{\delta}\right)^{\frac{1}{q}}.$$

Optimization with respect to q leads to $q = 2\log(2K/\delta)$ and consequently

$$R(\hat{\theta}_{\mathrm{ERM}}) \leq \inf_{\theta\in\Theta}\{r_n(\theta)\} + \sqrt{\frac{2\mathrm{e}\sigma^2\log\left(\frac{2K}{\delta}\right)}{n}}.$$

Without any additional assumption on the loss ℓ, the rate on the right-hand side is optimal. This is for example proven by Audibert (2009) for the absolute loss.

3.2 Dependent observations

Here we propose to analyze the harder and more realistic case where the observations (X_i, Y_i) are possibly dependent. It includes the autoregressive case where $X_i = Y_{i-1}$ or $X_i = (Y_{i-1}, \dots, Y_{i-p})$. Note that in this setting, different notions of risks were used in the literature. The risk $R(\theta)$ considered in this paper is the same as the one used in many references given in the introduction, Modha and Masry (1998); Steinwart and Christmann (2009); Alquier and Wintenberger (2012); Alquier and Li (2012) among others. Alternative notions of risk were proposed, for example by Zimin and Lampert (2015).

We remind the following definition.

Definition 4. *The α-mixing coefficients between two σ-algebras $\mathcal{F}$ and $\mathcal{G}$ are defined by*

$$\alpha(\mathcal{F}, \mathcal{G}) = \sup_{A \in \mathcal{F}, B \in \mathcal{G}} \left| \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) \right|.$$

We refer the reader to Doukhan (1994) and Rio (2000) (among others) for more details. We still provide a basic interpretation of this definition. First, when $\mathcal{F}$ and $\mathcal{G}$ are independent, then for all $A \in \mathcal{F}$ and $B \in \mathcal{G}$, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ by definition of independence, and so $\alpha(\mathcal{F}, \mathcal{G}) = 0$. On the other hand, when $\mathcal{F} = \mathcal{G}$, as soon as these σ-algebras contain an event A with $\mathbb{P}(A) = 1/2$ then $\alpha(\mathcal{F}, \mathcal{G}) = |\mathbb{P}(A \cap A) - \mathbb{P}(A)\mathbb{P}(A)| = |1/2 - 1/4| = 1/4$. More generally, $\alpha(\mathcal{F}, \mathcal{G})$ is a measure of the dependence of the information provided by $\mathcal{F}$ and $\mathcal{G}$, ranging from 0 (independance) to 1/4 (maximal dependence). We provide another interpretation in terms of covariances.

Proposition 7 (Classical, see Doukhan (1994) for a proof)**.** *We have*

$$\alpha(\mathcal{F}, \mathcal{G}) = \sup \Big\{ \mathrm{Cov}(U, V), 0 \leq U \leq 1, 0 \leq V \leq 1, \\ U \text{ is } \mathcal{F}\text{-measurable}, V \text{ is } \mathcal{G}\text{-measurable} \Big\}.$$

For short, define

$$\alpha_j = \alpha[\sigma(X_0, Y_0), \sigma(X_j, Y_j)]$$

where we remind that for any random variable Z, $\sigma(Z)$ is the σ-algebra generated by Z. The idea is that, when the future of the series is strongly dependent of the past, α_j will remain constant, or decay very slowly. On the other hand, when the near future is almost independent of the past, then the α_j decay very fast to 0 (examples of both kind can be found in Doukhan (1994); Rio (2000)). And, indeed, we will see below that when the rate of convergence of the α_j's to 0 is fast enough it is possible to derive results rather similar to the ones the independent case.

Let us first consider the bounded case.

Proposition 8. *Assume that $0 \leq \ell \leq 1$. Assume that $(X_i, Y_i)_{i \in \mathbb{Z}}$ is a stationary process, and that it satisfies $\sum_{j \in \mathbb{Z}} \alpha_j < \infty$. Then*

$$\mathcal{M}_{\phi_2, n} \leq \frac{1}{n} \sum_{j \in \mathbb{Z}} \alpha_j.$$

Examples of processes satisfying this assumption are discussed in Doukhan (1994); Rio (2000). For example, if the (X_i, Y_i)'s are actually a geometrically ergodic Markov chain then there exist some $c_1, c_2 > 0$ such that $\alpha_j \leq c_1 \mathrm{e}^{-c_2|j|}$. Thus

$$\mathcal{M}_{\phi_2,n} \leq \frac{1}{n} \frac{2c_1}{1-\mathrm{e}^{-c_2}}.$$

Proof of Proposition 8. We have:

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} \ell_i(\theta) - \mathbb{E}[\ell_i(\theta)]\right)^2\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}[\ell_i(\theta), \ell_j(\theta)] \\ &\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\in\mathbb{Z}} \alpha_{j-i} = \frac{\sum_{j\in\mathbb{Z}} \alpha_j}{n}\end{aligned}$$

that does not depend on θ, and so

$$\mathcal{M}_{\phi_2,n} = \int \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} \ell_i(\theta) - R(\theta)\right)^2\right] \pi(\mathrm{d}\theta) \leq \frac{\sum_{j\in\mathbb{Z}} \alpha_j}{n}.$$

□

Remark 1. *Other assumptions than α-mixing can be used. Actually, we see from the proof that the only requirement to get a bound on $\mathcal{M}_{\phi_2,n}$ is to control the covariance $\mathrm{Cov}[\ell_i(\theta), \ell_j(\theta)]$; α-mixing is very stringent as it imposes that we can control this for any function $\ell_i(\theta)$. In the case of a Lipschitz loss, we could actually consider more general conditions like the weak dependence conditions in Dedecker et al. (2007); Alquier and Wintenberger (2012).*

We now turn to the unbounded case.

Proposition 9. *Assume that $(X_i, Y_i)_{i\in\mathbb{Z}}$ is a stationary process. Let $r \geq 1$ and $s \geq 2$ be any numbers with $1/r + 2/s = 1$ and assume that*

$$\sum_{j\in\mathbb{Z}} \alpha_j^{1/r} < \infty$$

and

$$\int \left\{\mathbb{E}\left[\ell_i^s(\theta)\right]\right\}^{\frac{2}{s}} \pi(\mathrm{d}\theta) < \infty.$$

Then

$$\mathcal{M}_{\phi_2,n} \leq \frac{1}{n}\left(\int \left\{\mathbb{E}\left[\ell_i^s(\theta)\right]\right\}^{\frac{2}{s}} \pi(\mathrm{d}\theta)\right)\left(\sum_{j\in\mathbb{Z}} \alpha_j^{\frac{1}{r}}\right).$$

Proof of Proposition 9. The proof relies on the following property.

Proposition 10 (Doukhan (1994))**.** *For any random variables U and V, resp. $\mathcal{F}$ and $\mathcal{G}$-mesurable, we have*

$$|\mathrm{Cov}(U,V)| \leq 8\alpha^{\frac{1}{r}}(\mathcal{F},\mathcal{G})\|U\|_s\|V\|_t$$

where $1/r + 1/s + 1/t = 1$.

We use this with $U = \ell_i(\theta)$, $V = \ell_j(\theta)$ and $s = t$. Then

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\ell_i(\theta) - \mathbb{E}[\ell_i(\theta)]\right)^2\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}[\ell_i(\theta),\ell_j(\theta)] \\ &\leq \frac{8}{n^2}\sum_{i=1}^{n}\sum_{j\in\mathbb{Z}}\alpha_{j-i}^{\frac{1}{r}}\|\ell_i(\theta)\|_s\|\ell_j(\theta)\|_s \\ &\leq \frac{8\left\{\mathbb{E}\left[\ell_i^s(\theta)\right]\right\}^{\frac{2}{s}}\sum_{j\in\mathbb{Z}}\alpha_j^{\frac{1}{r}}}{n}.\end{aligned}$$

□

As an example, consider auto-regression with quadratic loss, where we use linear predictors: $X_i = (1, Y_{i-1}) \in \mathbb{R}^2$, $\Theta = \mathbb{R}^2$ and $f_\theta(\cdot) = \langle\theta,\cdot\rangle$. Then

$$|\ell_i(\theta)|^3 \leq 32[Y_i^6 + 4\|\theta\|^6(1 + Y_{i-1}^6)]$$

and so

$$\mathbb{E}\left(|\ell_i(\theta)|^3\right) \leq 32(1 + 4\|\theta\|^6)\mathbb{E}\left(Y_i^6\right).$$

Taking $s = r = 3$ in Proposition 9 leads to the following result.

Corollary 2. *Fix $\delta \in (0,1)$. Assume that π is chosen such that*

$$\int \|\theta\|^6\pi(\mathrm{d}\theta) < +\infty,$$

$\mathbb{E}\left(Y_i^6\right) < \infty$ and $\sum_{j\in\mathbb{Z}}\alpha_j^{\frac{1}{3}} < +\infty$. Put

$$\nu = 32\mathbb{E}\left(Y_i^6\right)^{\frac{2}{3}}\sum_{j\in\mathbb{Z}}\alpha_j^{\frac{1}{3}}\left(1 + 4\int\|\theta\|^6\pi(\mathrm{d}\theta)\right).$$

With probability at least $1-\delta$ we have for any ρ

$$\int R\mathrm{d}\rho \leq \int r_n\mathrm{d}\rho + \sqrt{\frac{\nu[1+\chi^2(\rho,\pi)]}{n\delta}}.$$

This is, up to our knowledge, the first PAC(-Bayesian) bound in the case of a time series without any boundedess nor exponential moment assumption.

4 Optimal aggregation distribution and oracle inequalities

We have now gone through the way to control the different terms in our PAC-Bayesian inequality (Theorem 1). We now come back to this result to derive which predictor minimizes the bound, and which statistical guarantees can be achieved by this predictor.

We start with a reminder of two consequences of Theorem 1: for $p > 1$, and $q = p/(p-1)$, with probability at least $1-\delta$ we have for any ρ

$$\int R\mathrm{d}\rho \leq \int r_n\mathrm{d}\rho + \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}}\left(D_{\phi_p-1}(\rho,\pi)+1\right)^{\frac{1}{p}} \tag{5}$$

and

$$\int r_n \mathrm{d}\rho \leq \int R \mathrm{d}\rho + \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}} \left(D_{\phi_p{}^{-1}}(\rho,\pi)+1\right)^{\frac{1}{p}}. \tag{6}$$

In this section we focus on the minimizer $\hat{\rho}_n$ of the right-hand side of (5) , and on its statistical properties.

Definition 5. *We define $\overline{r}_n = \overline{r}_n(\delta, p)$ as*

$$\overline{r}_n = \min\left\{u \in \mathbb{R}, \int [u - r_n(\theta)]_+^q \, \pi(\mathrm{d}\theta) = \frac{\mathcal{M}_{\phi_q,n}}{\delta}\right\}.$$

Note that such a minimum always exists as the integral is a continuous function of u, is equal to 0 when $u = 0$ and $\to \infty$ when $u \to \infty$. We then define

$$\frac{\mathrm{d}\hat{\rho}_n}{\mathrm{d}\pi}(\theta) = \frac{[\overline{r}_n - r_n(\theta)]_+^{\frac{1}{p-1}}}{\int [\overline{r}_n - r_n]_+^{\frac{1}{p-1}} \, \mathrm{d}\pi}. \tag{7}$$

The following proposition states that $\hat{\rho}_n$ is actually the minimizer of the right-hand side in inequality (5).

Proposition 11. *Under the assumptions of Theorem 1, with probability at least $1-\delta$,*

$$\begin{aligned}\overline{r}_n &= \int r_n \mathrm{d}\hat{\rho}_n + \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}} \left(D_{\phi_p{}^{-1}}(\hat{\rho}_n,\pi)+1\right)^{\frac{1}{p}} \\ &= \min_{\rho}\left\{\int r_n \mathrm{d}\rho + \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}} \left(D_{\phi_p{}^{-1}}(\rho,\pi)+1\right)^{\frac{1}{p}}\right\}\end{aligned}$$

where the minimum holds for any probability distribution ρ over Θ.

Proof of Proposition 11. For any ρ we have

$$\begin{aligned}\overline{r}_n - \int r_n \mathrm{d}\rho &= \int [\overline{r}_n - r_n] \mathrm{d}\rho \\ &= \int [\overline{r}_n - r_n]_+ \mathrm{d}\rho - \int [\overline{r}_n - r_n]_- \mathrm{d}\rho \\ &\leq \int [\overline{r}_n - r_n]_+ \mathrm{d}\rho = \int [\overline{r}_n - r_n]_+ \frac{\mathrm{d}\rho}{\mathrm{d}\pi} \mathrm{d}\pi \\ &\leq \left(\int [\overline{r}_n - r_n]_+^q \mathrm{d}\pi\right)^{\frac{1}{q}} \left(\int \left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)^p \mathrm{d}\pi\right)^{\frac{1}{p}} \\ &\leq \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}} \left(D_{\phi_p{}^{-1}}(\rho,\pi)+1\right)^{\frac{1}{p}}\end{aligned}$$

where we used Hölder's inequality and then the definition of $\overline{r}_n$ in the last line. Moreover, we can check that the two inequalities above become equalities when $\rho = \hat{\rho}_n$: from (7),

$$\overline{r}_n - \int r_n \mathrm{d}\hat{\rho}_n = \int [\overline{r}_n - r_n] \mathrm{d}\hat{\rho}_n = \int [\overline{r}_n - r_n]_+ \mathrm{d}\hat{\rho}_n$$

$$= \frac{\int [\overline{r}_n - r_n]_+ [\overline{r}_n - r_n]_+^{\frac{1}{p-1}} \mathrm{d}\pi}{\int [\overline{r}_n - r_n]_+^{\frac{1}{p-1}} \mathrm{d}\pi} = \frac{\int [\overline{r}_n - r_n]_+^q \mathrm{d}\pi}{\int [\overline{r}_n - r_n]_+^{\frac{1}{p-1}} \mathrm{d}\pi}$$

$$= \frac{\left(\int [\overline{r}_n - r_n]_+^q \mathrm{d}\pi\right)^{\frac{1}{p}+\frac{1}{q}}}{\int [\overline{r}_n - r_n]_+^{\frac{1}{p-1}} \mathrm{d}\pi} = \left(\int [\overline{r}_n - r_n]_+^q \mathrm{d}\pi\right)^{\frac{1}{q}} \frac{\left(\int [\overline{r}_n - r_n]_+^{\frac{p}{p-1}} \mathrm{d}\pi\right)^{\frac{1}{p}}}{\int [\overline{r}_n - r_n]_+^{\frac{1}{p-1}} \mathrm{d}\pi}$$

$$= \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}} \left(\int \left(\frac{\mathrm{d}\hat{\rho}_n}{\mathrm{d}\pi}\right)^p \mathrm{d}\pi\right)^{\frac{1}{p}} = \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}} \left(D_{\phi_p{}^{-1}}(\hat{\rho}_n, \pi) + 1\right)^{\frac{1}{p}}.$$

□

A direct consequence of (5) and (6) is the following result, which provides theoretical guarantees for $\hat{\rho}_n$.

Proposition 12. *Under the assumptions of Theorem 1, with probability at least* $1-\delta$,

$$\int R \mathrm{d}\hat{\rho}_n \leq \overline{r}_n \leq \inf_{\rho} \left\{ \int R \mathrm{d}\rho + 2\left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}} \left(D_{\phi_p{}^{-1}}(\rho, \pi) + 1\right)^{\frac{1}{p}} \right\}. \tag{8}$$

Proof of Proposition 12. First, (5) brings:

$$\begin{aligned} \int R \mathrm{d}\hat{\rho}_n &\leq \int r_n \mathrm{d}\hat{\rho}_n + \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}} \left(D_{\phi_p{}^{-1}}(\hat{\rho}_n, \pi) + 1\right)^{\frac{1}{p}} \\ &= \inf_{\rho} \left\{ \int r_n \mathrm{d}\rho + \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}} \left(D_{\phi_p{}^{-1}}(\rho, \pi) + 1\right)^{\frac{1}{p}} \right\} \end{aligned} \tag{9}$$

by definition of $\hat{\rho}_n$, and Proposition 11 shows that the right-hand side is $\bar{r}_n$. Plug (6) into (9) to get the desired result. □

Example 1. *As an example of an application of Proposition 12, we come back to the setting of a possibly heavy-tailed time series. More precisely, we assume that we are under the assumptions of Corollary 2. In particular,* $X_i = (1, Y_{i-1}) \in \mathbb{R}^2$ *and* $f_\theta(\cdot) = \langle \theta, \cdot \rangle$ *and* $p = q = 2$. *For the sake of simplicity, assume that the parameter space is* $\Theta = [-1,1]^2$. *Let us fix* π *as uniform on* $[-1,1]^2$. *The empirical bound stated that, with probability at least* $1-\delta$, *for any* ρ,

$$\int R \mathrm{d}\rho \leq \int r_n \mathrm{d}\rho + \sqrt{\frac{\nu[1 + \chi^2(\rho, \pi)]}{n\delta}}$$

where we remind that

$$\begin{aligned} \nu &= 32 \mathbb{E}\left(Y_i^6\right)^{\frac{2}{3}} \sum_{j \in \mathbb{Z}} \alpha_j^{\frac{1}{3}} \left(1 + 4 \int \|\theta\|^6 \pi(\mathrm{d}\theta)\right) \\ &\leq 1056 \mathbb{E}\left(Y_i^6\right)^{\frac{2}{3}} \sum_{j \in \mathbb{Z}} \alpha_j^{\frac{1}{3}}. \end{aligned}$$

In this context the minimizer of the right-hand side is

$$\frac{\mathrm{d}\hat{\rho}_n}{\mathrm{d}\pi}(\theta) = \frac{[\overline{r}_n - r_n(\theta)]_+}{\int [\overline{r}_n - r_n]_+ \mathrm{d}\pi}$$

where

$$\bar{r}_n = \min\left\{u \in \mathbb{R}, \int [u - r_n(\theta)]_+ \pi(\mathrm{d}\theta) = \frac{\nu}{n\delta}\right\}.$$

The application of Proposition 12 leads to, with probability at least $1-\delta$*,*

$$\int R \mathrm{d}\hat{\rho}_n \leq \inf_{\rho}\left\{\int R \mathrm{d}\rho + 2\sqrt{\frac{\nu[1+\chi^2(\rho,\pi)]}{n\delta}}\right\}.$$

Note that it is possible to derive an oracle inequality from this. Let us denote by $\bar{\theta} = (\bar{\theta}_1, \bar{\theta}_2)$ *the minimizer of* R*. Consider the following posteriors, for* $1 \leq i, j \leq N$*:*

$$\rho_{(i,j),N} \text{ is uniform on } \left[-1+\frac{2(i-1)}{N}, -1+\frac{2i}{N}\right] \times \left[-1+\frac{2(j-1)}{N}, -1+\frac{2j}{N}\right].$$

For N *fixed, there is always a pair* (i,j) *such that* $\bar{\theta}$ *belongs to the support of* $\rho_{(i,j),N}$*. Elementary calculus shows that, for any* θ *in the support of* $\rho_{(i,j),N}$ *then*

$$R(\theta) - R(\bar{\theta}) \leq \frac{2}{N}\left(1 + 4\mathbb{E}(|Y_i|) + 3\mathbb{E}(Y_i^2)\right) =: \frac{\nu'}{N}.$$

Moreover,

$$1 + \chi^2(\rho_{(i,j),N}, \pi) = \frac{N^2}{2}.$$

So the bound becomes

$$\int R \mathrm{d}\hat{\rho}_n \leq \inf_{\theta \in [-1,1]^2} R(\theta) + \inf_{N \in \mathbb{N}^*}\left\{\frac{\nu'}{N} + N\sqrt{\frac{2\nu}{n\delta}}\right\}$$

and in particular, the choice $N = \left\lceil \sqrt{\nu' \sqrt{n\delta/(2\nu)}} \right\rceil$ *leads to*

$$\int R \mathrm{d}\hat{\rho}_n \leq \inf_{\theta \in [-1,1]^2} R(\theta) + 3\left(\frac{2\nu\nu'^2}{n\delta}\right)^{\frac{1}{4}}$$

at least for n *large enough to ensure* $\left(\frac{n\delta}{2\nu}\right)^{\frac{1}{4}} \geq \sqrt{\nu'}$*.*

The last example shows that it is possible in some cases to deduce from Proposition 12 an oracle inequality, that is, a comparison to the performance of the optimal parameter. The end of this section is devoted to a systematic derivation of such oracle inequalities, using the complexity parameter introduced in Section 3, first in its empirical version, and then in its theoretical form.

Theorem 2. *Under the assumptions of Theorem 1 together with Assumption 1, with probability at least* $1-\delta$*,*

$$\int R \mathrm{d}\hat{\rho}_n \leq \bar{r}_n \leq \inf_{\theta \in \Theta}\{r_n(\theta)\} + 2\left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q+d}}. \tag{10}$$

Proof of Theorem 2. Put

$$\gamma = \bar{r}_n - \inf_{\theta \in \Theta}\{r_n(\theta)\}.$$

Note that $\gamma \geq 0$. Then:

$$\left(\frac{\gamma}{2}\right)^q \pi\left\{r_n(\theta) \leq \frac{\gamma}{2} + \inf r_n\right\} \leq \underbrace{\int [\overline{r}_n - r_n]_+^q \, \mathrm{d}\pi}_{= \frac{\mathcal{M}_{\phi_q,n}}{\delta}} \leq \gamma^q \pi\{r_n(\theta) \leq \gamma + \inf r_n\}.$$

So:

$$\left(\frac{\gamma}{2}\right)^q \pi\left\{r_n(\theta) \leq \frac{\gamma}{2} + \inf r_n\right\} \leq \frac{\mathcal{M}_{\phi_q,n}}{\delta}$$

and, using Assumption 1,

$$\left(\frac{\gamma}{2}\right)^q \left(\frac{\gamma}{2}\right)^d \leq \frac{\mathcal{M}_{\phi_q,n}}{\delta}$$

which yields:

$$\gamma \leq 2\left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q+d}}.$$

□

We can also perform an explicit minimization of the oracle-type bound (8), which leads to a variant of Theorem 2 under a non-empirical complexity assumption.

Definition 6. *Put*

$$\overline{R}_n = \min\left\{u \in \mathbb{R} : \int [u - R(\theta)]_+^q \, \pi(\mathrm{d}\theta) = \frac{2^q \mathcal{M}_{\phi_q}}{\delta}\right\}.$$

Assumption 2. *There exists $d > 0$ such that, for any $\gamma > 0$,*

$$\pi\left\{\theta \in \Theta : R(\theta) \leq \inf_{\theta' \in \Theta} \{R(\theta')\} + \gamma\right\} \geq \gamma^d.$$

Theorem 3. *Under the assumptions of Theorem 1 together with Assumption 2, with probability at least $1 - \delta$,*

$$\int R \mathrm{d}\hat{\rho}_n \leq \overline{R}_n \leq \inf_{\theta \in \Theta} R(\theta) + 2^{\frac{q}{q+d}} \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q+d}}.$$

The proof is a direct adaptation of the proofs of Proposition 11 and Theorem 2.

5 Discussion and perspectives

We proposed a new type of PAC-Bayesian bounds, which makes use of Csiszár's f-divergence to generalize the Kullback-Leibler divergence. This is an extension of the results in Bégin et al. (2016). In favourable contexts, there exists sophisticated approaches to get better bounds, as discussed in the introduction. However, the major contribution of our work is that our bounds hold in hostile situations where no PAC bounds at all were available, such as heavy-tailed time series. We plan to study the connections between our PAC-Bayesian bounds and aforementionned approaches by Mendelson (2015) and Grünwald and Mehta (2016) in future works.

Acknowledgements

We would like to thank Pascal Germain for fruitful discussions, along with two anonymous Referees and the Editor for insightful comments.

References

A. Agarwal and J. C. Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2013. 2

P. Alquier and X. Li. Prediction of quantiles by statistical learning and application to gdp forecasting. In *15th International Conference on Discovery Science 2012*, pages 23–36. Springer, 2012. 2, 8

P. Alquier and O. Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883–913, 2012. 2, 8, 9

P. Alquier, X. Li, and O. Wintenberger. Prediction of time series by statistical learning: General losses and fast rates. *Dependence Modeling*, 1:65–93, 2013. 2

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of gibbs posteriors. *Journal of Machine Learning Research*, 17(239): 1–41, 2016. URL http://jmlr.org/papers/v17/15-290.html. 2

J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009. 7

J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *The Annals of Statistics*, pages 2766–2794, 2011. 2, 7

L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy. PAC-Bayesian bounds based on the Rényi divergence. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 435–444, 2016. 2, 3, 4, 14

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013. 7

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics. Springer, 2004. 1

O. Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007. 1, 3, 5

O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185. Institut Henri Poincaré, 2012. 2

O. Catoni. PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*, 2016. 2

I. Csiszár and P. C. Shields. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004. 3

J. Dedecker, P. Doukhan, G. Lang, L. R. J. Rafael, S. Louhichi, and C. Prieur. Weak dependence. In *Weak Dependence: With Examples and Applications*, pages 9–20. Springer, 2007. 9

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996. 1

L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian mean estimators. *arXiv preprint arXiv:1509.05845*, 2015. 2

V. C. Dinh, L. S. Ho, B. Nguyen, and D. Nguyen. Fast learning rates with heavy-tailed losses. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 505–513. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6104-fast-learning-rates-with-heavy-tailed-losses.pdf. 2

P. Doukhan. *Mixing: Properties and Examples*. Lecture Notes in Statistics. Springer, New York, 1994. 2, 8, 9

C. Giraud, F. Roueff, and A. Sanchez-Pérez. Aggregation of predictors for nonstationary sub-linear processes and online adaptive forecasting of time varying autoregressive processes. *The Annals of Statistics*, 43(6):2412–2450, 2015. 2

I. Giulini. PAC-Bayesian bounds for Principal Component Analysis in Hilbert spaces. Preprint arXiv:1511.06263, 2015. 2

P. D. Grünwald and N. A. Mehta. Fast rates with unbounded losses. *arXiv preprint arXiv:1605.00252*, 2016. 2, 14

B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, 7:264–291, 2013. 1

L. Guillaume and L. Matthieu. Learning from mom's principles. *arXiv preprint arXiv:1701.01961*, 2017. 2

J. Honorio and T. Jaakkola. Tight bounds for the expected risk of linear classifiers and PAC-Bayes finite-sample guarantees. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 384–392, 2014. 7

D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016. 2

L. A. Kontorovich, K. Ramanan, et al. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6): 2126–2158, 2008. 2

V. Kuznetsov and M. Mohri. Generalization bounds for time series prediction with non-stationary processes. In *International Conference on Algorithmic Learning Theory*, pages 260–274. Springer, 2014. 2

J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, pages 439–446. MIT Press, 2002. 1

G. Lecué and S. Mendelson. Regularization and the small-ball method I: sparse recovery. *arXiv preprint arXiv:1601.05584*, 2016. 2

B. London, B. Huang, and L. Getoor. Stability and generalization in structured prediction. *Journal of Machine Learning Research*, 17(222):1–52, 2016. 2

G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*, 2016. 2

G. Lugosi and S. Mendelson. Regularization, sparse recovery, and median-of-means tournaments. *arXiv preprint arXiv:1701.04112*, 2017. 2

D. A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234, New York, 1998. ACM. 1

D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170. ACM, 1999. 1

S. Mendelson. Learning without concentration. *J. ACM*, 62(3):21:1–21:25, June 2015. ISSN 0004-5411. doi: 10.1145/2699439. URL http://doi.acm.org/10.1145/2699439. 2, 14

S. Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015. 2

D. S. Modha and E. Masry. Memory-universal prediction of stationary random processes. *IEEE transactions on information theory*, 44(1):117–133, 1998. 2, 8

M. Mohri and A. Rostamizadeh. Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(Feb):789–814, 2010. 2

R. I. Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. *arXiv preprint arXiv:1312.2903, to appear in Probability Theory and Related Fields*, 2013. 2

L. Oneto, D. Anguita, and S. Ridella. PAC-Bayesian analysis of distribution dependent priors: Tighter risk bounds and stability analysis. *Pattern Recognition Letters*, 80:200–207, 2016. 2

L. Ralaivola, M. Szafranski, and G. Stempfel. Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research*, 11(Jul):1927–1956, 2010. 2

E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants*, volume 31 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2000. 2, 8, 9

M. Seeger. PAC-Bayesian generalisation error bounds for gaussian process classification. *Journal of machine learning research*, 3(Oct):233–269, 2002. 1

Y. Seldin and N. Tishby. PAC-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11(Dec):3595–3646, 2010. 1

Y. Seldin, P. Auer, J. Shawe-Taylor, R. Ortner, and F. Laviolette. PAC-Bayesian analysis of contextual bandits. In *Advances in Neural Information Processing Systems*, pages 1683–1691, 2011. 1

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012. 1, 2

J. Shawe-Taylor and R. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 2–9, New York, 1997. ACM. 1

I. Steinwart and A. Christmann. Fast learning from non-iid observations. In *Advances in Neural Information Processing Systems*, pages 1768–1776, 2009. 2, 8

N. N. Taleb. *The black swan: The impact of the highly improbable*. Random house, 2007. 2

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11): 1134–1142, 1984. 1

V. N. Vapnik. *The nature of Statistical Learning Theory*. Springer, 2000. 1

B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994. 2

A. Zimin and C. H. Lampert. Conditional risk minimization for stochastic processes. *arXiv preprint arXiv:1510.02706*, 2015. 2, 8