Prediction of time series by statistical learning: general losses and fast rates

Pierre Alquier, Xiaoyin Li, Olivier Wintenberger

Prediction of time series by statistical learning: general losses and fast rates

Pierre Alquier[1] **, Xiaoyin Li**[2] **and Olivier Wintenberger**[3,4]

(1) University College Dublin
School of Mathematical Sciences
Belfield
Dublin 4 - Ireland
e-mail: pierre.alquier@ucd.ie
url: http://alquier.ensae.net/

(2) Université de Cergy-Pontoise site Saint-Martin
Laboratoire de Mathématiques
2, boulevard Adolphe Chauvin
95000 Cergy-Pontoise, France
e-mail: xiaoyin.li@u-cergy.fr

(3) Université Paris Dauphine - CEREMADE
Place du Maréchal de Lattre de Tassigny
75775 Paris CEDEX 16, France
e-mail: wintenberger@ceremade.dauphine.fr
url: http://wintenberger.fr/

(4) CREST-LFA
15, boulevard Gabriel Péri
92245 Malakoff CEDEX, France

Abstract: We establish rates of convergences in time series forecasting using the statistical learning approach based on oracle inequalities. A series of papers (e.g. [MM98, Mei00, BCV01, AW12]) extends the oracle inequalities obtained for iid observations to time series under weak dependence conditions. Given a family of predictors and n observations, oracle inequalities state that a predictor forecasts the series as well as the best predictor in the family up to a remainder term Δ_n. Using the PAC-Bayesian approach, we establish under weak dependence conditions oracle inequalities with optimal rates of convergence Δ_n. We extend results given in [AW12] for the absolute loss function to any Lipschitz loss function with rates $\Delta_n \sim \sqrt{c(\Theta)/n}$ where $c(\Theta)$ measures the complexity of the model. We apply the method for quantile loss functions to forecast the french GDP. Under additional conditions on the loss functions (satisfied by the quadratic loss function) and on the time series, we refine the rates of convergence to $\Delta_n \sim c(\Theta)/n$. We achieve for the first time these fast rates for uniformly mixing processes. These rates are known to be optimal in the iid case, see [Tsy03], and for individual sequences, see [CBL06]. In particular, we generalize the results of [DT08] on sparse regression estimation to the case of autoregression.

*We deeply thank Matthieu Cornec (INSEE) for useful discussions, and for providing the data with detailed explanations. We would like to thank Prs. Olivier Catoni, Paul Doukhan, Pascal Massart and Gilles Stoltz for useful comments. We want to mention that a preliminary version of Theorem 9 appeared in proceedings of DS'12 [AL12]. Research partially supported by the "Agence Nationale pour la Recherche", grant ANR-09-BLAN-0128 "PARCIMONIE" and grant ANR-Bandhits.

AMS 2000 subject classifications: Primary 62M20; secondary 68T05, 62M10, 62M45, 62P20.
Keywords and phrases: Statistical learning theory, Time series prediction, PAC-Bayesian bounds, weak-dependence, mixing, oracle inequalities, fast rates, GDP Forecasting.

1. Introduction

Time series forecasting is a fundamental subject in the mathematical statistics literature. The parametric approach contains a wide range of models associated with efficient estimation and prediction methods, see e.g. [Ham94]. Classical parametric models include linear processes such as ARMA models [BD09]. More recently, non-linear processes such as stochastic volatility and ARCH models received a lot of attention in financial applications - see, e.g., the seminal paper by Nobel prize winner [Eng82], and [FZ10] for a more recent introduction. However, parametric assumptions rarely hold on data. Assuming that the data satisfy a model can biased the prediction and underevaluate the risks, see among others the the polemical but highly informative discussion in [Tal07].

In the last few years, several universal approaches emerged from various fields such as non-parametric statistics, machine learning, computer science and game theory. These approaches share some common features: the aim is to build a procedure that predicts the time series as well as the best predictor in a given set of initial predictors Θ, without any parametric assumption on the distribution of the observed time series. However, the set of predictors can be inspired by different parametric or non-parametric statistical models. We can distinguish two classes in these approaches, with different quantification of the objective, and different terminologies:

- in the "prediction of individual sequences" approach, predictors are usually called "experts". The objective is online prediction: at each date t, a prediction of the future realization x_{t+1} is based on the previous observations x_1, ..., x_t, the objective being to minimize the cumulative prediction loss. See for example [CBL06, Sto10] for an introduction.
- in the statistical learning approach, the given predictors are sometimes referred as "models" or "concepts". The batch setting is more classical in this approach. A prediction procedure is built on a complete sample X_1, ..., X_n. The performance of the procedure is compared on the expected loss, called the risk, with the best predictor, called the "oracle". The environment is not deterministic and some hypotheses like mixing or weak dependence are required: see [Mei00, MM98, AW12].

In both settings, one is usually able to predict a time series as well as the best model or expert, up to an error term that decreases with the number of observations n. This type of results is referred in statistical theory as oracle inequalities. In other words, one builds on the basis of the observations a predictor $\hat{\theta}$ such that

$$R(\hat{\theta}) \leq \inf_{\theta \in \Theta} R(\theta) + \Delta(n, \Theta) \tag{1.1}$$

where $R(\theta)$ is a measure of the prediction risk of the predictor $\theta \in \Theta$. In general, the remainder term is of the order $\Delta(n,\Theta) \sim \sqrt{c(\Theta)/n}$ in both approaches, where $c(\Theta)$ measures the complexity of Θ. See, e.g., [CBL06] for the "individual sequences" approach; for the "statistical learning approach" the rate $\sqrt{c(\Theta)/n}$ is reached in [AW12] with the absolute loss function and under a weak dependence assumption. Different procedures are used to reach these rates. Let us mention the empirical risk minimization [Vap99] and aggregation procedures with exponential weights, usually referred as EWA [DT08, Ger11] or Gibbs estimator [Cat04, Cat07] in the batch approach, linked to the weighted majority algorithm of the online approach [LW94], see also [Vov90]. Note that results from the "individual sequences" approach can sometimes be extended to the batch setting, see e.g. [Ger11] for the iid case, and [AD11, DAJJ12] for mixing time series.

In this paper, we extend the results of [AW12] to the case of a general loss function. Another improvement with respect to [AW12] is to study both the ERM and the Gibbs estimator under various hypotheses. We achieve here inequalities of the form of (1.1) that hold with large probability ($1-\varepsilon$ for any arbitratily small confidence level $\varepsilon > 0$) with $\Delta(n,\Theta) \sim \sqrt{c(\Theta)/n}$. We assume to do so that the observations are taken from a bounded stationary process (X_t) (see [AW12] however for some possible extensions to unbounded observations). We also assume weak dependence conditions on the process process (X_t). Then we prove that the fast rate $\Delta(n,\Theta) \sim c(\Theta)/n$ can be reached for some loss functions including the quadratic loss. Note that [Mei00, MM98] deal with the quadratic loss, their rate can be better than $\sqrt{c(\Theta)/n}$ but cannot reach $c(\Theta)/n$.

Our main results are based on PAC-Bayesian oracle inequalities. The PAC-Bayesian point of view emerged in statistical learning in supervised classification using the 0/1-loss, see the seminal papers [STW97, McA99]. These results were then extended to general loss functions and more accurate bounds were given, see for example [Cat04, Cat07, Alq08, Aud10, AL11, SLCB+12, DS12]. In PAC-Bayesian inequalities the complexity term $c(\Theta)$ is defined thanks to a prior distribution on the set Θ.

The paper is organized as follows: Section 2 provides notations used in the whole paper. We give a definition of the Gibbs estimator and of the ERM in Section 3. The main hypotheses necessary to prove theoretical results on these estimators are provided in Section 4. We give examples of inequalities of the form (1.1) for classical set of predictors Θ in Section 5. When possible, we also prove some results on the ERM in these settings. These results only require a general weak-dependence type assumption on the time series to forecast. We then study fast rates under a stronger $\phi-$mixing assumptions of [Ibr62] in Section 6. Note that the ϕ-mixing setting coincides with the one of [AD11, DAJJ12] when (X_t) is stationary. In particular, we are able to generalize the results of [DT08, Ger11, AL11] on sparse regression estimation to the case of autoregression. In Section 7 we provide an application to French GDP forecasting. A short simulation study is provided in Section 8. Finally, the proofs of all the theorems are given in Appendices A and B.

2. Notations

Let $X_1, \dots, X_n$ denote the observations at time $t \in \{1, \dots, n\}$ of a time series $X = (X_t)_{t\in\mathbb{Z}}$ defined on $(\Omega, \mathcal{A}, \mathbb{P})$. We assume that this series is stationary and take values in $\mathbb{R}^p$ equipped with the Euclidean norm $\|\cdot\|$. We fix an integer k, that might depend on n, $k = k(n)$, and assume that family of predictors is available: $\left\{f_\theta : (\mathbb{R}^p)^k \to \mathbb{R}^p, \theta \in \Theta\right\}$. For any parameter θ and any time t, $f_\theta\left(X_{t-1}, \dots, X_{t-k}\right)$ is the prediction of X_t returned by the predictor θ when given $(X_{t-1}, \dots, X_{t-k})$. For the sake of shortness, we use the notation:

$$\hat{X}_t^\theta = f_\theta(X_{t-1}, \dots, X_{t-k}).$$

We assume that $\theta \mapsto f_\theta$ is a linear function. Let us fix a loss function ℓ that measures a distance between the forecast and the actual realization of the series. Assumptions on ℓ will be given in Section 4.

Definition 1. *For any $\theta \in \Theta$ we define the prediction risk as*

$$R\left(\theta\right) = \mathbb{E}\left[\ell\left(\hat{X}_t^\theta, X_t\right)\right]$$

($R(\theta)$ does not depend on t thanks to the stationarity assumption).

Using the statistics terminology, note that we may want to include parametric set of predictors as well as non-parametric ones (i.e. respectively finite dimensional and infinite dimensional Θ). Let us mention classical parametric and non-parametric families of predictors:

Example 1. *Define the set of linear autoregressive predictors as*

$$f_\theta(X_{t-1}, \dots, X_{t-k}) = \theta_0 + \sum_{j=1}^{k} \theta_j X_{t-j}$$

for $\theta = (\theta_0, \theta_1, \dots, \theta_k) \in \Theta \subset \mathbb{R}^{k+1}$.

In order to deal with non-parametric settings, we will also use a model-selection type notation: $\Theta = \cup_{j=1}^{M} \Theta_j$.

Example 2. *Consider non-parametric auto-regressive predictors*

$$f_\theta(X_{t-1}, \dots, X_{t-k}) = \sum_{i=1}^{j} \theta_i \varphi_i(X_{t-1}, \dots, X_{t-k})$$

where $\theta = (\theta_1, \dots, \theta_j) \in \Theta_j \subset \mathbb{R}^j$ and $(\varphi_i)_{i=0}^{\infty}$ is a dictionnary of functions $(\mathbb{R}^p)^k \to \mathbb{R}^p$ (e.g. Fourier basis, wavelets, splines...).

3. ERM and Gibbs estimator

3.1. The estimators

As the objective is to minimize the risk $R(\cdot)$, we use the empirical risk $r_n(\cdot)$ as an estimator of $R(\cdot)$.

Definition 2. *For any* $\theta \in \Theta$, $r_n(\theta) = \frac{1}{n-k} \sum_{i=k+1}^{n} \ell\left(\hat{X}_i^{\theta}, X_i\right)$.

Definition 3 (ERM estimator [Vap99])**.** *We define the Empirical Risk Minimizer estimator (ERM) by*

$$\hat{\theta}^{ERM} \in \arg\min_{\theta \in \Theta} r_n(\theta).$$

Let $\mathcal{T}$ be a σ-algebra on Θ and $\mathcal{M}_+^1(\Theta)$ denote the set of all probability measures on $(\Theta, \mathcal{T})$. The Gibbs estimator depends on a fixed probability measure $\pi \in \mathcal{M}_+^1(\Theta)$ called the *prior* that will be involved when measuring the complexity of Θ.

Definition 4 (Gibbs estimator or EWA)**.** *Define the Gibbs estimator with inverse temperature* $\lambda > 0$ *as*

$$\hat{\theta}_\lambda = \int_\Theta \theta \hat{\rho}_\lambda(\mathrm{d}\theta), \text{ where } \hat{\rho}_\lambda(\mathrm{d}\theta) = \frac{e^{-\lambda r_n(\theta)}\pi(\mathrm{d}\theta)}{\int e^{-\lambda r_n(\theta')}\pi(\mathrm{d}\theta')}.$$

The choice of π and λ in practice is discussed in Section 5.

3.2. Overview of the results

Our results assert that the risk of the ERM or Gibbs estimator is close to $\inf_\theta R(\theta)$ up to a remainder term $\Delta(n, \Theta)$ called the rate of convergence. For the sake of simplicity, let $\overline{\theta} \in \Theta$ be such that

$$R(\overline{\theta}) = \inf_\theta R(\theta).$$

If $\overline{\theta}$ does not exist, it is replaced by an approximative minimizer $\overline{\theta}_\alpha$ satisfying $R(\overline{\theta}_\alpha) \leq \inf_\theta R(\theta) + \alpha$ where α is negligible w.r.t. $\Delta(n, \Theta)$ (e.g. $\alpha < 1/n^2$). We want to prove that the ERM satisfies, for any $\varepsilon > 0$,

$$\mathbb{P}\left(R\left(\hat{\theta}^{ERM}\right) \leq R(\overline{\theta}) + \Delta(n, \Theta, \varepsilon)\right) \geq 1 - \varepsilon \tag{3.1}$$

where $\Delta(n, \Theta, \varepsilon) \to 0$ as $n \to \infty$. We also want to prove that and that the Gibbs estimator satisfies, for any $\varepsilon > 0$,

$$\mathbb{P}\left(R\left(\hat{\theta}_\lambda\right) \leq R(\overline{\theta}) + \Delta(n, \lambda, \pi, \varepsilon)\right) \geq 1 - \varepsilon \tag{3.2}$$

where $\Delta(n, \lambda, \pi, \varepsilon) \to 0$ as $n \to \infty$ for some $\lambda = \lambda(n)$. To obtain such results called *oracle inequalities*, we require some assumptions discussed in the next section.

4. Main assumptions

We prove oracle inequalities under assumptions of two different types. On the one hand, assumptions **LipLoss**(K) and **Lip**(L) hold respectively on the loss function ℓ and the set of predictors Θ. In some extent, we choose the loss function and the predictors, so these assumptions can always be satisfied. Assumption **Margin**$(\mathcal{K})$ also holds on ℓ.

On the other hand, assumptions **Bound**$(\mathcal{B})$, **WeakDep**$(\mathcal{C})$, **PhiMix**$(\mathcal{C})$ hold on the dependence and boundedness of the time series. In practice, we cannot know whether these assumptions are satisfied on data. However, remark that these assumptions are not parametric and are satisfied for many classical models, see [Dou94, DDL+07].

Assumption LipLoss(K)**,** $K > 0$: the loss function ℓ is given by $\ell(x, x') = g(x - x')$ for some convex K-Lipschitz function g such that $g(0) = 0$ and $g \geq 0$.

Example 3. *A classical example in statistics is given by* $\ell(x, x') = \|x - x'\|$*, see [AW12]. It satisfies* **LipLoss**(K) *with* $K = 1$*. In [MM98, Mei00], the loss function used is the quadratic loss* $\ell(x, x') = \|x - x'\|^2$*. It satisfies* **LipLoss**$(4\mathcal{B})$ *for time series bounded by a constant* $\mathcal{B} > 0$*.*

Example 4. *The class of quantile loss functions introduced in [KB78] is given by*

$$\ell_\tau(x, y) = \begin{cases} \tau\,(x - y)\,, & \text{if } x - y > 0 \\ -\,(1 - \tau)\,(x - y)\,, & \text{otherwise} \end{cases}$$

where $\tau \in (0, 1)$ *and* x*,* $y \in \mathbb{R}$*. The risk minimizer of* $t \mapsto \mathbb{E}(\ell_\tau(V - t))$ *is the quantile of order* τ *of the random variable* V*. Choosing this loss function one can deal with rare events and build confidence intervals, see [Koe05, BC11, BP11]. In this case,* **LipLoss**(K) *is satisfied with* $K = \max(\tau, 1 - \tau) \leq 1$*.*

Assumption Lip(L)**,** $L > 0$: for any $\theta \in \Theta$ there are coefficients $a_j(\theta)$ for $1 \leq j \leq k$ such that, for any x_1, ..., x_k and y_1, ..., y_k,

$$\|f_\theta(x_1, \ldots, x_k) - f_\theta(y_1, \ldots, y_k)\| \leq \sum_{j=1}^{k} a_j(\theta)\,\|x_j - y_j\|\,,$$

with $\sum_{j=1}^{k} a_j(\theta) \leq L$.

Assumption Bound$(\mathcal{B})$**,** $\mathcal{B} > 0$: we assume that $\|X_0\| \leq \mathcal{B}$ almost surely.

Remark that under Assumptions **LipLoss**(K), **Lip**(L) and **Bound**(B), the empirical risk is a bounded random variable. Such a condition is required in the approach of individual sequences. We assume it here for simplicity but it is possible to extend the slow rates oracles inequalities to unbounded cases see [AW12].

Assumption **WeakDep**$(\mathcal{C})$ is about the $\theta_{\infty,n}(1)$-weak dependence coefficients of [Rio00, DDL+07].

Definition 5. *For any $k > 0$, define the $\theta_{\infty,k}(1)$-weak dependence coefficients of a bounded stationary sequence (X_t) by the relation*

$$\theta_{\infty,k}(1) := \sup_{f\in\Lambda_1^k, 0<j_1<\cdots<j_k} \left\| \mathbb{E}\left[f(X_{j_1},\dots,X_{j_\ell})|X_t, t\leq 0\right] - \mathbb{E}\left[f(X_{j_1},\dots,X_{j_\ell})\right] \right\|_\infty$$

where Λ_1^k is the set of 1-Lipshitz functions of k variables

$$\Lambda_1^k = \left\{ f : (\mathbb{R}^p)^k \to \mathbb{R}, \quad \frac{|f(u_1,\dots,u_k) - f(u'_1,\dots,u'_k)|}{\sum_{j=1}^k \|u_j - u'_j\|} \leq 1 \right\}.$$

The sequence $(\theta_{\infty,k}(1))_{k>0}$ is non decreasing with k. The idea is that as soon as X_k behaves"almost independently" from X_0, X_{-1}, ... then $\theta_{\infty,k}(1) - \theta_{\infty,k-1}(1)$ becomes negligible. Actually, it is known that for many classical models of stationary time series, the sequence is upper bounded, see [DDL+07] for details.

Assumption WeakDep($\mathcal{C}$), $\mathcal{C} > 0$: $\theta_{\infty,k}(1) \leq \mathcal{C}$ for any $k > 0$.

Example 5. *Examples of processes satisfying* **WeakDep**($\mathcal{C}$) *are provided in [AW12, DDL+07]. It includes Bernoulli shifts $X_t = H(\xi_t, \xi_{t-1}, \dots)$ where the ξ_t are iid, $\|\xi_0\| \leq b$ and H satisfies a Lipschitz condition:*

$$\|H(v_1, v_2, ...) - H(v'_1, v'_2, ...)\| \leq \sum_{j=0}^{\infty} a_j \|v_j - v'_j\| \text{ with } \sum_{j=0}^{\infty} j a_j < \infty.$$

Then (X_t) is bounded by $\mathcal{B} = H(0, 0, ...) + b\mathcal{C}$ and satisfies **WeakDep**($\mathcal{C}$) *with $\mathcal{C} = \sum_{j=0}^{\infty} j a_j$. In particular, solutions of linear* ARMA *models with bounded innovations satisfy* **WeakDep**($\mathcal{C}$).

In order to prove the fast rates oracle inequalities, a more restrictive dependence condition is assumed. It holds on the uniform mixing coefficients introduced by [Ibr62].

Definition 6. *The ϕ-mixing coefficients of the stationary sequence (X_t) with distribution $\mathbb{P}$ are defined as*

$$\phi_r = \sup_{(A,B)\in\sigma(X_t,t\leq 0)\times\sigma(X_t,t\geq r)} |\mathbb{P}(B/A) - \mathbb{P}(B)|.$$

Assumption PhiMix($\mathcal{C}'$), $\mathcal{C}' > 0$: $1 + \sum_{r=1}^{\infty} \sqrt{\phi_r} \leq \mathcal{C}'$.

This assumption appears to be more restrictive than **WeakDep**($\mathcal{C}$) for bounded time series:

Proposition 1 ([Rio00])**.**

$$\textbf{Bound}(\mathcal{B}) \textit{ and } \textbf{PhiMix}(\mathcal{C}) \Rightarrow \textbf{Bound}(\mathcal{B}) \textit{ and } \textbf{WeakDep}(\mathcal{C}\mathcal{B}).$$

(This result is not stated in [Rio00] but it is a direct consequence of the last inequality in the proof of Corollaire 1, p. 907 in [Rio00]).

Finally, for fast rates oracle inequalities, an additional assumption on the loss function ℓ is required. In the iid case, such a condition is also required. It is called Margin assumption, e.g. in [MT99, Alq08], or Bernstein hypothesis, [Lec11].

Assumption Margin($\mathcal{K}$), $\mathcal{K} > 0$:

$$\mathbb{E}\left\{\left[\ell\Big(X_{q+1}, f_\theta(X_q, ..., X_1)\Big) - \ell\Big(X_{q+1}, f_{\overline{\theta}}(X_q, ..., X_1)\Big)\right]^2\right\} \leq \mathcal{K}\left[R(\theta) - R(\overline{\theta})\right].$$

As assumptions **Margin**($\mathcal{K}$) and **PhiMix**($\mathcal{C}$) are used only to obtain fast rates, we give postpone examples to Section 6.

5. Slow rates oracle inequalities

In this section, we give oracle inequalities (3.1) and/or (3.2) with slow rates of convergence $\Delta(n, \Theta) \sim \sqrt{c(\Theta)/n}$. The proof of these results are given in Section B. Note that the results concerning the Gibbs estimator are actually corollaries of a general result, Theorem 9, stated in Section A. We introduce the following notation for the sake of shortness.

Definition 7. *When Assumptions* **Bound**($\mathcal{B}$), **LipLoss**(K), **Lip**(L) *and* **WeakDep**($\mathcal{C}$) *are satisfied, we say that we are under the set of Assumption* **SlowRates**(κ) *where* $\kappa = K(1 + L)(\mathcal{B} + \mathcal{C})/\sqrt{2}$.

5.1. Finite classes of predictors

Consider first the toy example where Θ is finite with $|\Theta| = M$, $M \geq 1$. In this case, the optimal rate in the iid case is known to be $\sqrt{\log(M)/n}$, see e.g. [Vap99].

Theorem 1. *Assume that* $|\Theta| = M$ *and that* **SlowRates**(κ) *is satisfied for* $\kappa > 0$. *Let* π *be the uniform probability distribution on* Θ. *Then the oracle inequality* (3.2) *is satisfied for any* $\lambda > 0$, $\varepsilon > 0$ *with*

$$\Delta(n, \lambda, \pi, \varepsilon) = \frac{2\lambda\kappa^2}{n\left(1 - k/n\right)^2} + \frac{2\log\left(2M/\varepsilon\right)}{\lambda}.$$

The choice of λ in practice in this toy example is already not trivial. The choice $\lambda = \sqrt{\log(M)n}$ yields the oracle inequality:

$$R(\hat{\theta}_\lambda) \leq R(\overline{\theta}) + 2\sqrt{\frac{\log(M)}{n}}\left(\frac{\kappa}{1 - k/n}\right)^2 + \frac{2\log\left(2/\varepsilon\right)}{\sqrt{n\log(M)}}.$$

However, this choice is not optimal and one would like to choose λ as the minimizer of the upper bound

$$\frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2}+\frac{2\log\left(M\right).}{\lambda}$$

However $\kappa=\kappa(K,L,\mathcal{B},\mathcal{C})$ and the constants $\mathcal{B}$ and $\mathcal{C}$ are, usually, unknown. In this context we will prefer the ERM predictor that performs as well as the Gibbs estimator with optimal λ:

Theorem 2. *Assume that* $|\Theta|=M$ *and that* **SlowRates**(κ) *is satisfied for* $\kappa>0$. *Then the oracle inequality* (3.1) *is satisfied for any* $\varepsilon>0$ *with*

$$\Delta(n,\Theta,\varepsilon)=\inf_{\lambda>0}\left[\frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2}+\frac{2\log\left(2M/\varepsilon\right)}{\lambda}\right]=\frac{4\kappa}{1-k/n}\sqrt{\frac{\log\left(2M/\varepsilon\right)}{n}}.$$

5.2. Linear autoregressive predictors

We focus on the linear predictors given in Example 1.

Theorem 3. *Consider the linear autoregressive model of* AR(k) *predictors*

$$f_\theta(x_{t-1},\dots,x_{t-k})=\theta_0+\sum_{j=1}^{k}\theta_j x_{t-j}$$

with $\theta\in\Theta=\{\theta\in\mathbb{R}^{k+1},\|\theta\|\leq L\}$ *such that* **Lip**(L) *is satisfied. Assume that Assumptions* **Bound**$(\mathcal{B})$, **LipLoss**(K) *and* **WeakDep**$(\mathcal{C})$ *are satisfied. Let* π *be the uniform probability distribution on the extended parameter set* $\{\theta\in\mathbb{R}^{k+1},\|\theta\|\leq L+1\}$. *Then the oracle inequality* (3.2) *is satisfied for any* $\lambda>0$, $\varepsilon>0$ *with*

$$\Delta(n,\lambda,\pi,\varepsilon)=\frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2}+2\frac{(k+1)\log\left(\frac{(K\mathcal{B}\vee K^2\mathcal{B}^2)(L+1)\sqrt{e}\lambda}{k+1}\right)+\log\left(2/\varepsilon\right)}{\lambda}.$$

In theory, λ can be chosen of the order $\sqrt{(k+1)n}$ to achieve the optimal rates $\sqrt{(k+1)/n}$ up to a logarithmic factor. But the choice of the optimal λ in practice is still a problem. The ERM predictor still performs as well as the Gibbs predictor with optimal λ.

Theorem 4. *Under the assumptions of Theorem 3, the oracle inequality* (3.1) *is satisfied for any* $\varepsilon>0$ *with*

$$\Delta(n,\Theta,\varepsilon)=\inf_{\lambda\geq 2K\mathcal{B}/(k+1)}\left[\frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2}+\frac{(k+1)\log\left(\frac{2eK\mathcal{B}(L+1)\lambda}{k+1}\right)+2\log\left(2/\varepsilon\right)}{\lambda}\right].$$

The additional constraint on λ does not depend on n. It is restrictive only when $k+1$, the complexity of the autoregressive model, has the same order than n. For n sufficiently large and $\lambda = ((1-k/n)/\kappa)\sqrt{((k+1)n/2)}$ satisfying the constraint $\lambda \geq 2K\mathcal{B}/(k+1)$ we obtain the oracle inequality

$$R(\hat{\theta}^{ERM}) \leq R(\overline{\theta}) + \sqrt{\frac{2(k+1)}{n}}\frac{\kappa}{1-k/n}\log\left(\frac{2e^2K\mathcal{B}(R+1)}{\kappa}\sqrt{\frac{n}{k+1}}\right) + \frac{2\sqrt{2}\kappa\log(2/\varepsilon)}{\sqrt{(k+1)n}\,(1-k/n)}.$$

Theorems 3 and 4 are both direct consequences of the following results about general classes of predictors.

5.3. General parametric classes of predictors

We state a general result about finite-dimensional families of predictors. The complexity $k+1$ of the autoregressive model is replaced by a more general measure of the dimension $d(\Theta,\pi)$. We also introduce some general measure $D(\Theta,\pi)$ of the diameter that will, for most compact models, be linked to the diameter of the model.

Theorem 5. *Assume that* **SlowRates**(κ) *is satisfied and the existence of* $d = d(\Theta,\pi) > 0$ *and* $D = D(\Theta,\pi) > 0$ *satisfying the relation*

$$\forall \delta > 0, \quad \log\frac{1}{\int_{\theta\in\Theta}\mathbf{1}\{R(\theta)-R(\overline{\theta})<\delta\}\pi(\mathrm{d}\theta)} \leq d\log\left(\frac{D}{\delta}\right).$$

Then the oracle inequality (3.2) *is satisfied for any* $\lambda > 0$, $\varepsilon > 0$ *with*

$$\Delta(n,\lambda,\pi,\varepsilon) = \frac{2\lambda\kappa^2}{n\,(1-k/n)^2} + 2\frac{d\log(D\sqrt{e}\lambda/d) + \log(2/\varepsilon)}{\lambda}.$$

A similar result holds for the ERM predictor under a more restrictive assumption on the structure of Θ, see Remark 1 below.

Theorem 6. *Assume that*

1. $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq D\}$,
2. $\|\hat{X}_1^{\theta_1} - \hat{X}_1^{\theta_2}\| \leq \psi.\|\theta_1-\theta_2\|_1$ *a.s. for some* $\psi > 0$ *and all* $(\theta_1,\theta_2) \in \Theta^2$.

Assume also that **Bound**$(\mathcal{B})$, **LipLoss**(K) *and* **WeakDep**$(\mathcal{C})$ *are satisfied and that* **Lip**(L) *holds on the extended model* $\Theta' = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq D+1\}$. *Then the oracle inequality* (3.1) *is satisfied for any* $\varepsilon > 0$ *with*

$$\Delta(n,\Theta,\varepsilon) = \inf_{\lambda \geq 2K\psi/d}\left[\frac{2\lambda\kappa^2}{n\,(1-k/n)^2} + \frac{d\log(2eK\psi(D+1)\lambda/d) + 2\log(2/\varepsilon)}{\lambda}\right].$$

This result yields to nearly optimal rates of convergence for the ERM predictors. Indeed, for n sufficiently large and $\lambda = ((1-k/n)/\kappa)\sqrt{(dn/2)} \geq 2K\psi/d$ we obtain the oracle inequality

$$R(\hat{\theta}^{ERM}) \leq R(\overline{\theta}) + \sqrt{\frac{2d}{n}}\frac{\kappa}{1-k/n}\log\left(\frac{2e^2K\psi(D+1)}{\kappa}\sqrt{\frac{n}{d}}\right) + \frac{2\sqrt{2}\kappa\log(2/\varepsilon)}{\sqrt{dn}\,(1-k/n)}.$$

Thus, the ERM procedure yields prediction that are close to the oracle with an optimal rate of convergence up to a logarithmic factor.

Example 6. *Consider the linear autoregressive model of* AR(k) *predictors studied in Theorems 3 and 4. Then* **Lip**(L) *is automatically satisfied with* $L = D+1$. *The assumptions of Theorem 6 are satisfied with* $d = k+1$ *and* $\psi = \mathcal{B}$. *Moreover, thanks to Remark 1, the assumptions of Theorem 5 are satisfied with* $D(\Theta,\pi) = (K\mathcal{B} \vee K^2\mathcal{B}^2)(R+1)$. *Then Theorems 3 and 4 are actually direct consequences of Theorems 5 and 6.*

Note that the context of Theorem 6 are less general than the one of Theorem 5:

Remark 1. *Under the assumptions of Theorem 6 we have for any* $\theta \in \Theta$

$$\begin{aligned} R(\theta) - R(\overline{\theta}) &= \mathbb{E}\left\{ g\left(\hat{X}_1^{\theta} - X_1\right) - g\left(\hat{X}_1^{\overline{\theta}} - X_1\right)\right\} \\ &\leq \mathbb{E}\left\{ K\left\|\hat{X}_1^{\theta} - \hat{X}_1^{\overline{\theta}}\right\|\right\} \\ &\leq K\psi\|\theta - \overline{\theta}\|_1. \end{aligned}$$

Define π *as the uniform distribution on* $\Theta' = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq D+1\}$. *We derive from simple computation the inequality*

$$\log\frac{1}{\int_{\theta\in\Theta}\mathbf{1}\{R(\theta)-R(\overline{\theta})<\delta\}\pi(\mathrm{d}\theta)} \leq \log\frac{1}{\int_{\theta\in\Theta}\mathbf{1}\{\|\theta-\overline{\theta}\|_1<\frac{\delta}{K\psi}\}\pi(\mathrm{d}\theta)}$$
$$\begin{cases} = d\log\left(\frac{K\psi(D+1)}{\delta}\right) & \text{when } \delta/K\psi \leq 1 \\ \leq d\log\left(K\psi(D+1)\right) & \text{otherwise.} \end{cases}$$

Thus, in any case,

$$\log\frac{1}{\int_{\theta\in\Theta}\mathbf{1}\{R(\theta)-R(\overline{\theta})<\delta\}\pi(\mathrm{d}\theta)} \leq d\log\left(\frac{(K\psi\vee K^2\psi^2)(D+1)}{\delta}\right)$$

and the assumptions of Theorem 5 are satisfied for $d(\Theta,\pi) = d$ *and* $D(\Theta,\pi) = (K\psi\vee K^2\psi^2)(D+1)$.

As a conclusion, for some predictors set with a non classical structure, the Gibbs estimator might be preferred to the ERM.

5.4. Aggregation in the model-selection setting

Consider now several models of predictors Θ_1, ..., Θ_M and consider $\Theta = \bigsqcup_{i=1}^M \Theta_i$ (disjoint union). Our aim is to predict as well as the best predictors among all Θ_j's, but paying only the price for learning in the Θ_j that contains the oracle. In order to get such a result, let us choose M priors π_j on each models such that $\pi_j(\Theta_j) = 1$ for all $j \in \{1, ..., M\}$. Let $\pi = \sum_{j=1}^M p_j \pi_j$ be a mixture of these priors with prior weights $p_j \geq 0$ satisfying $\sum_{j=1}^M p_j = 1$. Denote

$$\overline{\theta}_j \in \arg\min_{\theta \in \Theta_j} R(\theta)$$

the oracle of the model Θ_j for any $1 \leq j \leq M$. For any $\lambda > 0$, denote $\hat{\rho}_{\lambda,j}$ the Gibbs distribution on Θ_j and $\hat{\theta}_{\lambda,j} = \int_{\Theta_j} \theta \hat{\rho}_{\lambda,j}(\mathrm{d}\theta)$ the corresponding Gibbs estimator. A Gibbs predictor based on a model selection procedure satisfies an oracle inequality with slow rate of convergence:

Theorem 7. *Assume that:*

1. **Bound**$(\mathcal{B})$ *is satisfied for some* $\mathcal{B} > 0$*;*
2. **LipLoss**(K) *is satisfied for some* $K > 0$*;*
3. **WeakDep**$(\mathcal{C})$ *is satisfied for some* $\mathcal{C} > 0$*;*
4. *for any* $j \in \{1, ..., M\}$ *we have*
   (a) **Lip**(L_j) *is satisfied by the model* Θ_j *for some* $L_j > 0$*,*
   (b) *there are constants* $d_j = d(\Theta_j, \pi)$ *and* $D_j = c(\Theta_j, \pi_j)$ *are such that*

$$\forall \delta > 0, \quad \log \frac{1}{\int_{\theta \in \Theta_j} \mathbf{1}\{R(\theta) - R(\overline{\theta}_j) < \delta\} \pi_j(\mathrm{d}\theta)} \leq d_j \log\left(\frac{D_j}{\delta}\right)$$

Denote $\kappa_j = \kappa(K, L_j, \mathcal{B}, \mathcal{C}) = K(1 + L_j)(\mathcal{B} + \mathcal{C})/\sqrt{2}$ *and define* $\hat{\theta} = \hat{\theta}_{\lambda_{\hat{j}}, \hat{j}}$ *where* $\hat{j}$ *minimizes the function of* j

$$\int_{\Theta_j} r_n(\theta) \hat{\rho}_{\lambda_j, j}(\mathrm{d}\theta) + \frac{\lambda_j \kappa_j}{n(1 - k/n)^2} + \frac{\mathcal{K}(\hat{\rho}_{\lambda_j, j}, \pi_j) + \log\left(2/(\varepsilon p_j)\right)}{\lambda_j}$$

with

$$\lambda_j = \arg\min_{\lambda > 0} \left[\frac{2\lambda \kappa_j^2}{n\left(1 - k/n\right)^2} + 2 \frac{d_j \log\left(D_j e \lambda / d_j\right) + \log\left(2/(\varepsilon p_j)\right)}{\lambda} \right].$$

Then, with probability at least $1 - \varepsilon$*, the following oracle inequality holds*

$$R(\hat{\theta}) \leq \inf_{1 \leq j \leq M} \left[R(\overline{\theta}_j) + 2 \frac{\kappa_j}{1 - k/n} \left\{ \sqrt{\frac{d_j}{n}} \log\left(\frac{D_j e^2}{\kappa_j} \sqrt{\frac{n}{d_j}} \right) + \frac{\log\left(2/(\varepsilon p_j)\right)}{\sqrt{n d_j}} \right\} \right].$$

The proof is given in Appendix B. A similar result can be obtained if we replace the Gibbs predictor in each model by the ERM predictor in each model. The resulting procedure is known in the iid case under the name SRM (Structural Risk Minimization), see [Vap99], or penalized risk minimization, [BM01]. However, as it was already the case for a fixed model, additional assumptions are required to deal with ERM predictors. In the model-selection context, the procedure to choose among all the ERM predictors also depends on the unknown κ_j's. Thus the model-selection procedure based on Gibbs predictors outperforms the one based on the ERM predictors.

6. Fast rates oracle inequalities

6.1. Discussion on the assumptions

In this section, we study conditions under which the rate $1/n$ can be achieved. These conditions are restrictive:

- now $p = 1$, i.e. the process $(X_t)_{t\in\mathbb{Z}}$ is real-valued;
- the dependence condition **WeakDep**$(\mathcal{C})$ is replaced by **PhiMix**$(\mathcal{C})$;
- we assume additionally **Margin**$(\mathcal{K})$ for some $\mathcal{K} > 0$.

Let us provide some examples of processes satisfying the uniform mixing assumption **PhiMix**$(\mathcal{C})$. In the three following examples (ϵ_t) denotes an iid sequence (called the innovations).

Example 7 (AR(p) process)**.** *Consider the stationary solution (X_t) of an AR(p) model: $\forall t \in \mathbb{Z}$, $X_t = \sum_{j=1}^{p} a_j X_{t-j} + \epsilon_t$. Assume that (ϵ_t) is bounded with a distribution possessing an absolutely continuous component. If $\mathcal{A}(z) = \sum_{j=1}^{p} a_j z^j$ has no root inside the unit disk in $\mathbb{C}$ then (X_t) is a geometrically ϕ-mixing processe, see [AP86] and* **PhiMix**$(\mathcal{C})$ *is satisfied for some $\mathcal{C}$.*

Example 8 (MA(p) process)**.** *Consider the stationary process (X_t) such that $X_t = \sum_{j=1}^{p} b_j \epsilon_{t-j}$ for all $t \in \mathbb{Z}$. By definition, the process (X_t) is stationary and ϕ-dependent - it is even p-dependent, in the sense that $\phi_r = 0$ for $r > p$. Thus* **PhiMix**$(\mathcal{C})$ *is satisfied for some $\mathcal{C} > 0$.*

Example 9 (Non linear processes)**.** *For extensions of the AR(p) model of the form $X_t = F(X_{t-1}, \ldots, X_{t-p}; \epsilon_t)$, Φ-mixing coefficients can also be computed and satisfy* **PhiMix**$(\mathcal{C})$*. See e.g. [MT93].*

We now provide an example of predictive model satisfying all the assumptions required to obtain fast rates oracle inequalities, in particular **Margin**$(\mathcal{K})$, when the loss function ℓ is quadratic, i.e. $\ell(x, x') = (x - x')^2$:

Example 10. *Consider Example 2 where*

$$f_\theta(X_{t-1}, \ldots, X_{t-k}) = \sum_{i=1}^{N} \theta_i \varphi_i(X_{t-1}, \ldots, X_{t-k}),$$

for functions $(\varphi_i)_{i=0}^{\infty}$ *of* $(\mathbb{R}^p)^k$ *to* $\mathbb{R}^p$*, and* $\theta = (\theta_1, \dots, \theta_{\mathbb{N}}) \in \mathbb{R}^N$*. Assume the* φ_i *upper bounded by* 1 *and* $\Theta = \{\theta \in \mathbb{R}^N, \|\theta\|_1 \leq L\}$ *such that* $\mathbf{Lip}(L)$*. Moreover* $\mathbf{LipLoss}(K)$ *is satisfied with* $K = 2\mathcal{B}$*. Assume that* $\overline{\theta} = \arg\min_{\theta \in \mathbb{R}^N} R(\theta) \in \Theta$ *in order to have:*

$$\begin{aligned}
&\mathbb{E}\left\{\left[\Big(X_{q+1} - f_\theta(X_q, ..., X_1)\Big)^2 - \Big(X_{q+1} - f_{\overline{\theta}}(X_q, ..., X_1)\Big)^2\right]^2\right\} \\
&= \mathbb{E}\Big\{\big[f_\theta(X_q, ..., X_1) - f_{\overline{\theta}}(X_q, ..., X_1)\big]^2 \\
&\qquad\qquad \big[2X_{q+1} - f_\theta(X_q, ..., X_1) - f_{\overline{\theta}}(X_q, ..., X_1)\big]^2\Big\} \\
&\leq \mathbb{E}\left\{\big[f_\theta(X_q, ..., X_1) - f_{\overline{\theta}}(X_q, ..., X_1)\big]^2 4\mathcal{B}^2(1+R)^2\right\} \\
&\leq 4\mathcal{B}^2(1+R)^2\left[R(\theta) - R(\overline{\theta})\right] \text{ by Pythagorean theorem.}
\end{aligned}$$

Assumption $\mathbf{Margin}(\mathcal{K})$ *is satisfied with* $\mathcal{K} = 4\mathcal{B}^2(1+D)^2$*. According to Theorem 8 below, the oracle inequality with fast rates holds as soon as Assumption* $\mathbf{PhiMix}(\mathcal{C})$ *is satisfied.*

6.2. General result

We only give oracle inequalities for the Gibbs predictor in the model-selection setting. In the case of one single model, this result can be extended to the ERM predictor. For several models, the approach based on the ERM predictors requires a penalized risk minimization procedure as in the slow rates case. In the fast rates case, the Gibbs predictor itself directly have nice properties. Let $\Theta = \bigsqcup_{i=1}^{M} \Theta_i$ (disjoint union), choose $\pi = \sum_{j=1}^{M} p_j \pi_j$ and denote $\overline{\theta}_j \in \arg\min_{\theta \in \Theta_j} R(\theta)$ as previously.

Theorem 8. *Assume that:*

1. $\mathbf{Margin}(\mathcal{K})$ *and* $\mathbf{LipLoss}(K)$ *are satisfied for some* K, $\mathcal{K} > 0$*;*
2. $\mathbf{Bound}(\mathcal{B})$ *is satisfied for some* $\mathcal{B} > 0$*;*
3. $\mathbf{PhiMix}(\mathcal{B})$ *is satisfied for some* $\mathcal{C} > 0$*;*
4. $\mathbf{Lip}(L)$ *is satisfied for some* $L > 0$*;*
5. *for any* $j \in \{1, ..., M\}$*, there exist* $d_j = d(\Theta_j, \pi)$ *and* $D_j = D(\Theta_j, \pi_j)$ *satisfying the relation*

$$\forall \delta > 0, \quad \log \frac{1}{\int_{\theta \in \Theta_j} \mathbf{1}\{R(\theta) - R(\overline{\theta}_j) < \delta\} \pi_j(\mathrm{d}\theta)} \leq d_j \log\left(\frac{D_j}{\delta}\right).$$

Then for

$$\lambda = \frac{n-k}{4kKL\mathcal{B}\mathcal{C}} \wedge \frac{n-k}{16k\mathcal{C}}$$

the oracle inequality (3.2) *for any* $\varepsilon > 0$ *with*

$$\Delta(n,\lambda,\pi,\varepsilon)$$
$$= 4\inf_j \left\{ R(\overline{\theta}_j) - R(\overline{\theta}) + 4k\mathcal{C}\left(4 \vee KL\mathcal{B}\right) \frac{d_j \log\left(\frac{D_j e(n-k)}{16k\mathcal{C}d_j}\right) + \log\left(\frac{2}{\varepsilon p_j}\right)}{n-k} \right\}.$$

Compare with the slow rates case, we don't have to optimize with respect to λ as the optimal order for λ is independent of j. In practice, the value of λ provided by Theorem 8 is too conservative. In the iid case, it is shown in [DT08] that the value $\lambda = n/(4\sigma^2)$, where σ^2 is the variance of the noise of the regression yields good results. In our simulations results, we will use $\lambda = n/\hat{\text{var}}(X)$, where $\hat{\text{var}}(X)$ is the empirical variance of the observed time series.

Notice that for the index j_0 such that $R(\overline{\theta}_{j_0}) = R(\overline{\theta})$ we obtain:

$$R\left(\hat{\theta}_\lambda\right) \leq$$
$$R(\overline{\theta}) + 4k\mathcal{C}\left(4 \vee KL\mathcal{B}\right) \frac{d_{j_0} \log\left(c_{j_0} e(n-k)/(16k\mathcal{C}d_{j_0})\right) + \log\left(2/(\varepsilon p_{j_0})\right)}{n-k}.$$

So, the oracle inequality achieves the fast rate $d_{j_0}/n \log(n/d_{j_0})$ where j_0 is the model of the oracle. However, note that the choice $j = j_0$ does not necessarily reach the infimum in Theorem 8.

Let us compare the rates in Theorem 8 to the ones in [Mei00, MM98, AD11, DAJJ12]. In [Mei00, MM98], the optimal rate $1/n$ is never obtained. The paper [AD11] proves fast rates for online algorithms that are also computationally efficient, see also [DAJJ12]. The fast rate $1/n$ is reached when the coefficients (ϕ_r) are geometrically decreasing. In other cases, the rate is slower. Note that we do not suffer such a restriction. The Gibbs estimator of Theorem 8 can also be computed efficiently thanks to MCMC procedures, see [AL11, DT08].

6.3. Corollary: sparse autoregression

Let the predictors be the linear autoregressive predictors

$$\hat{X}_p^\theta = \sum_{j=1}^p X_{p-j}\theta_j.$$

For any $J \subset \{1,\dots,p\}$, define the model:

$$\Theta_J = \{\theta \in \mathbb{R}^p : \|\theta\|_1 \leq L \text{ and } \theta_j \neq 0 \Leftrightarrow j \in J\}.$$

Let us remark that we have the disjoint union $\Theta = \bigsqcup_{J\subset\{1,\dots,p\}} \Theta_J = \{\theta \in \mathbb{R}^p : \|\theta\|_1 \leq 1\}$. We choose π_J as the uniform probability measure on Θ_J and $p_j = 2^{-|J|-1}\binom{p}{|J|}^{-1}$.

Corollary 1. *Assume that* $\overline{\theta} = \arg\min_{\theta\in\mathbb{R}^N} R(\theta) \in \Theta$ *and* **PhiMix**$(\mathcal{C})$ *is satisfied for some* $\mathcal{C} > 0$ *as well as* **Bound**$(\mathcal{B})$. *Then the oracle inequality* (3.2) *is satisfied for any* $\varepsilon > 0$ *with*

$$\Delta(n,\lambda,\pi,\varepsilon) = 4\inf_J \left\{ R(\overline{\theta}_J) - R(\overline{\theta}) + \text{cst.}\frac{|J|\log\left((n-k)p/|J|\right) + \log\left(\frac{2}{\varepsilon}\right)}{n-k} \right\}$$

for some constant $\text{cst} = \text{cst}(\mathcal{B},\mathcal{C},L)$.

This extends the results of [AL11, DT08, Ger11] to the case of autoregression.

Proof. The proof follows the computations of Example 10 that we do not reproduce here: we check the conditions **LipLoss**(K) with $K = 2\mathcal{B}$, **Lip**(L) and **Margin**$(\mathcal{K})$ with $\mathcal{K} = 4\mathcal{B}^2(1+L)^2$. We can apply Theorem 8 with $d_J = |J|$ and $D_j = L$. □

7. Application to French GDP forecasting

7.1. Uncertainty in GDP forecasting

Every quarter $t \geq 1$, the French national bureau of statistics, INSEE[1], publishes the growth rate of the French GDP (Gross Domestic Product). Since it involves a huge amount of data that take months to be collected and processed, the computation of the GDP growth rate $\log(\text{GDP}_t/\text{GDP}_{t-1})$ takes a long time (two years). This means that at time t, the value $\log(\text{GDP}_t/\text{GDP}_{t-1})$ is actually not known. However, a preliminary value of the growth rate is published 45 days only after the end of the current quarter t. This value is called a *flash estimate* and is the quantity that INSEE forecasters actually try to predict, at least in a first time. As we want to work under the same constraint as the INSEE, we will now focus on the prediction on the flash estimate and let ΔGDP_t denote this quantity. To forecast at time t, we will use:

1. the past forecastings[2] ΔGDP_j, $0 < j < t$;
2. past *climate indicators* I_j, $0 < j < t$, based on *business surveys*.

Business surveys are questionnaires of about ten questions sent monthly to a representative panel of French companies (see [Dev84] for more details). As a consequence, these surveys provide informations from the economic decision makers. Moreover, they are available each end of months and thus can be used to forecast the french GDP. INSEE publishes a composite indicator, the *French business climate indicator* that summarizes information of the whole business survey, see [CM09, DM06]. Following [Cor10], let I_t be the mean of the last three (monthly based) climate indicators available for each quarter $t > 0$ at the date of publication of ΔGDP_t. All these values (GDP, climate indicator) are

[1] *Institut National de la Statistique et des Etudes Economiques*http://www.insee.fr/

[2] It has been checked that to replace past flash estimates by the actual GDP growth rate when it becomes available do not improve the quality of the forecasting [Min10].

available from the INSEE website. Note that a similar approach is used in other countries, see e.g. [BBR08] on forecasting the European Union GDP growth thanks to EUROSTATS data.

In order to provide a quantification of the uncertainty of the forecasting, associated interval confidences are usually provided. The ASA and the NBER started using density forecasts in 1968, while the Central Bank of England and INSEE provide their prediction with a *fan chart*, see ee [DTW97, TW00] for surveys on density forecasting and [BFW98] for fan charts. However, the statistical methodology used is often crude and, until 2012, the fan charts provided by the INSEE was based on the homoscedasticity of the Gaussian forecasting errors, see [Cor10, Dow04]. However, empirical evidences are

1. the GDP forecasting is more uncertain in a period of crisis or recession;
2. the forecasting errors are not symmetrically distributed.

7.2. Application of Theorem 6 for the GDP forecasting

Define X_t as the data observed at time t: $X_t = (\Delta\text{GDP}_t, I_t)' \in \mathbb{R}^2$. We use the quantile loss function (see Example 4 page 6) for some $0 < \tau < 1$ of the quantity of interested ΔGDP_t:

$$\ell_\tau((\Delta\text{GDP}_t, I_t), (\Delta'\text{GDP}_t, I'_t)) = \begin{cases} \tau\left(\Delta\text{GDP}_t - \Delta'\text{GDP}_t\right), & \text{if } \Delta\text{GDP}_t - \Delta'\text{GDP}_t > 0 \\ -(1-\tau)\left(\Delta\text{GDP}_t - \Delta'\text{GDP}_t\right), & \text{otherwise.} \end{cases}$$

We use the family of forecasters proposed by [Cor10] given by the relation

$$f_\theta(X_{t-1}, X_{t-2}) = \theta_0 + \theta_1 \Delta\text{GDP}_{t-1} + \theta_2 I_{t-1} + \theta_3 (I_{t-1} - I_{t-2}) |I_{t-1} - I_{t-2}| \quad (7.1)$$

where $\theta = (\theta_0, \theta_1, \theta_2, \theta_3) \in \Theta(B)$. Fix $D > 0$ and

$$\Theta = \left\{ \theta = (\theta_0, \theta_1, \theta_2, \theta_3) \in \mathbb{R}^4, \|\theta\|_1 = \sum_{i=0}^{3} |\theta_i| \leq D \right\}.$$

Let us denote $R^\tau(\theta) := \mathbb{E}\left[\ell_\tau\left(\Delta\text{GDP}_t, f_\theta(X_{t-1}, X_{t-2})\right)\right]$ the risk of the forecaster f_θ and let r_n^τ denote the associated empirical risk. We let $\hat{\theta}^{ERM,\tau}$ denote the ERM with quantile loss ℓ_τ:

$$\hat{\theta}^{ERM,\tau} \in \arg\min_{\theta \in \Theta} r_n^\tau(\theta).$$

We apply Theorem 6 as $\mathbf{Lip}(L)$ is satisfied Θ' with $L = D+1$ and $\mathbf{LipLoss}(K)$ with $K = 1$. If the observations are bounded, stationary such that $\mathbf{WeakDep}(\mathcal{C})$ holds for some $\mathcal{C} > 0$, the assumptions of Theorem 6 are satisfied with $\psi = \mathcal{B}$ and $d = 4$:

Corollary 2. *Let us fix $\tau \in (0,1)$. If the observations are bounded, stationary such that* **WeakDep**$(\mathcal{C})$ *holds for some $\mathcal{C} > 0$ then for any $\varepsilon > 0$ and n large enough, we have*

$$\mathbb{P}\left\{ R^\tau(\hat{\theta}^{ERM,\tau}) \leq \inf_{\theta\in\Theta} R^\tau(\theta) + \frac{2\kappa\sqrt{2}}{\sqrt{n}\,(1-4/n)} \log\left(\frac{2e^2\mathcal{B}(D+1)\sqrt{n}}{\kappa\varepsilon}\right)\right\} \geq 1-\varepsilon.$$

In practice the choice of D has little importance as soon as D is large enough (only the theoretical bound is influenced). As a consequence we take $D = 100$ in our experiments.

7.3. Results

The results are shown in Figure 1 for forecasting corresponding to $\tau = 0.5$. Figure 2 represents the confidence intervals of order 50%, i.e. $\tau = 0.25$ and $\tau = 0.75$ (left) and for confidence interval of order 90%, i.e. $\tau = 0.05$ and $\tau = 0.95$ (right). We report only the results for the period 2000-Q1 to 2011-Q3 (using the period 1988-Q1 to 1999-Q4 for learning).

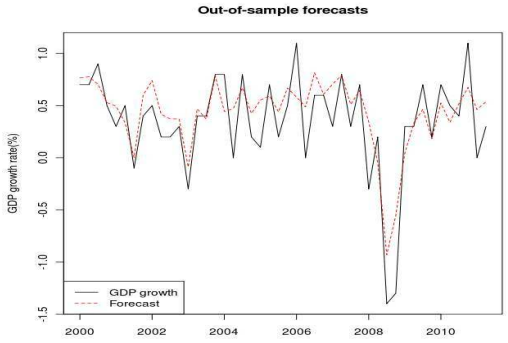

FIG 1. *French GDP forecasting using the quantile loss function with $\tau = 0.5$.*

We denote $\hat{\theta}^{ERM,\tau}[t]$ the estimator computed at time $t-1$, based on the observations X_j, $j < t$. We report the online performance:

$$\begin{aligned} \text{mean abs. pred. error} &= \tfrac{1}{n}\textstyle\sum_{t=1}^n \left|\Delta GDP_t - f_{\hat{\theta}^{ERM,0.5}[t]}(X_{t-1}, X_{t-2})\right| \\ \text{mean quad. pred. error} &= \tfrac{1}{n}\textstyle\sum_{t=1}^n \left[\Delta GDP_t - f_{\hat{\theta}^{ERM,0.5}[t]}(X_{t-1}, X_{t-2})\right]^2 \end{aligned}$$

and compare it to the INSEE performance, see Table 1. We also report the frequency that the GDPs fall above the predicted τ-quantiles for each τ, see Table 2. Note that this quantity should be close to τ.

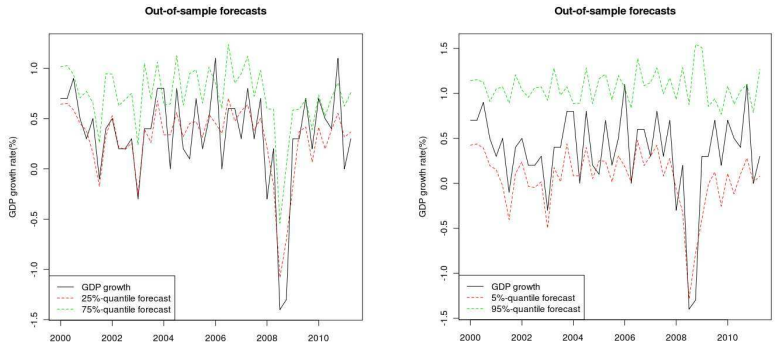

FIG 2. *French GDP online* 50%*-confidence intervals (left) and* 90%*-confidence intervals (right).*

Predictor	Mean absolute prediction error	Mean quadratic prediction error
$\widehat{\theta}^{ERM,0.5}$	0.2249	0.0812
INSEE	0.2579	0.0967

TABLE 1

Performances of the ERM and of the INSEE.

τ	Estimator	Frequency
0.05	$\widehat{\theta}^{ERM,0.05}$	0.1739
0.25	$\widehat{\theta}^{ERM,0.25}$	0.4130
0.5	$\widehat{\theta}^{ERM,0.5}$	0.6304
0.75	$\widehat{\theta}^{ERM,0.75}$	0.9130
0.95	$\widehat{\theta}^{ERM,0.95}$	0.9782

TABLE 2

Empirical frequencies of the event: GDP falls under the predicted τ*-quantile.*

The methodology fails to forecast the importance of the 2008 subprime crisis as it was the case for the INSEE forecaster, see [Cor10]. However, it is interesting to note that the confidence interval is larger at that date: the forecast is less reliable, but thanks to our adaptive confidence interval, it would have been possible to know at that time that the prediction was not reliable. Another interesting point is to remark that the lower bound of the confidence intervals are varying over time while the upper bound is almost constant for $\tau = 0.95$. It supports the idea of asymmetric forecasting errors. A parametric model with gaussian innovations would lead to underestimate the recessions risk.

8. Simulation study

In this section, we finally compare the ERM or Gibbs estimators to the Quasi Maximum Likelihood Estimator (QMLE) based method used by the R function ARMA [R D08]. The idea is not to claim any superiority of one method over another, it is rather to check that the ERM and Gibbs estimators can be safely

n	Model	Innovations	ERM abs.	ERM quad.	QMLE
100	(8.1)	Gaussian	**0.1436** (0.1419)	0.1445 (0.1365)	0.1469 (0.1387)
		Uniform	0.1594 (0.1512)	**0.1591**(0.1436)	0.1628 (0.1486)
	(8.2)	Gaussian	0.1770 (0.1733)	0.1699 (0.1611)	0.1728 (0.1634)
		Uniform	**0.1520** (0.1572)	0.1528 (0.1495)	0.1565 (0.1537)
1000	(8.1)	Gaussian	**0.1336** (0.1291)	0.1343 (0.1294)	0.1345 (0.1296)
		Uniform	**0.1718** (0.1369)	0.1729 (0.1370)	0.1732 (0.1372)
	(8.2)	Gaussian	0.1612(0.1375)	**0.1610** (0.1367)	0.1613 (0.1369)
		Uniform	0.1696 (0.1418)	**0.1687** (0.1404)	0.1691 (0.1407)

TABLE 3

Performances of the ERM estimators and ARMA, on the simulations. The first row "ERM abs." is for the ERM estimator with absolute loss, the second row "ERM quad." for the ERM with quadratic loss. The standard deviations are given in parentheses.

used in various contexts as their performances are close to the standard QMLE even in the context where the series is generated from an ARMA model. It is also the opportunity to check the robustness of our estimators in case of misspecification.

8.1. Parametric family of predictors

Here, we compare the ERM to the QMLE.

We draw simulations from an AR(1) models (8.1) and a non linear model (8.2):

$$X_t = 0.5X_{t-1} + \varepsilon_t \tag{8.1}$$

$$X_t = 0.5\sin(X_{t-1}) + \varepsilon_t \tag{8.2}$$

where ε_t are iid innovations. We consider two cases of distributions for ε_t: the uniform case, $\varepsilon_t \sim \mathcal{U}[-a, a]$, and the Gaussian case, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. Note that, in the first case, both models satisfy the assumptions of Theorem 8: there exists a stationary solutions (X_t) that is ϕ-mixing when the innovations are uniformly distributed and **WeakDep**$(\mathcal{C})$ is satisfied for some $\mathcal{C} > 0$. This paper does not provide any theoretical results for the Gaussian case as it is unbounded. However, we refer the reader to [AW12] for truncations techniques that allows to deal with this case too. We fix $\sigma = 0.4$ and $a = 0.70$ such that $Var(\epsilon_t) \simeq 0.16$ in both cases. For each model, we simulate first a sequence of length n and then we predict X_n using the observations $(X_1, \ldots, X_{n-1})$. Each simulation is repeated 100 times and we report the mean quadratic prediction errors on the Table 3.

It is interesting to note that the ERM estimator with absolute loss performs better on model (8.1) while the ERM with quadratic loss performs slightly better on model (8.2). The difference tends be too small to be significative, however, the numerical results tends to indicate that both methods are robust to model mispecification. Also, both estimators seem to perform better than the R QMLE procedure when $n = 100$, but the differences tends to be less perceptible when n grows.

TABLE 4
Performances of the Gibbs, AIC and "full model" predictors on simulations.

n	Model	Innovations	Gibbs	AIC	Full Model
100	(8.3)	Uniform	0.165 (0.022)	0.165 (0.023)	0.182 (0.029)
		Gaussian	0.167 (0.023)	0.161 (0.023)	0.173 (0.027)
	(8.4)	Uniform	0.163 (0.020)	0.169 (0.022)	0.178 (0.022)
		Gaussian	0.172 (0.033)	0.179 (0.040)	0.201 (0.049)
	(8.5)	Uniform	0.174 (0.022)	0.179 (0.028)	0.201 (0.040)
		Gaussian	0.179 (0.025)	0.182 (0.025)	0.202 (0.031)
1000	(8.3)	Uniform	0.163 (0.005)	0.163 (0.005)	0.166 (0.005)
		Gaussian	0.160 (0.005)	0.160 (0.005)	0.162 (0.005)
	(8.4)	Uniform	0.164 (0.004)	0.166 (0.004)	0.167 (0.004)
		Gaussian	0.160 (0.008)	0.161 (0.008)	0.163 (0.008)
	(8.5)	Uniform	0.171 (0.005)	0.172 (0.006)	0.175 (0.006)
		Gaussian	0.173 (0.009)	0.173 (0.009)	0.176 (0.010)

8.2. Sparse autoregression

To illustrate Corollary 1, we compare the Gibbs predictor to the model selection approach of the ARMA procedure in the R software. This procedure computes the QMLE estimator in each AR(p) model, $1 \leq p \leq q$, and then selects the order p by Akaike's AIC criterion [Aka73]. The Gibbs estimator is computed using a Reversible Jump MCMC algorithm as in [AL11]. The parameter λ is taken as $\lambda = n/\hat{\text{var}}(X)$, the empirical variance of the observed time series.

We draw the data according to the following models:

$$X_t = 0.5X_{t-1} + 0.1X_{t-2} + \varepsilon_t \tag{8.3}$$

$$X_t = 0.6X_{t-4} + 0.1X_{t-8} + \varepsilon_t \tag{8.4}$$

$$X_t = \cos(X_{t-1})\sin(X_{t-2}) + \varepsilon_t \tag{8.5}$$

where ε_t are iid innovations. We still consider the uniform ($\varepsilon_t \sim \mathcal{U}[-a, a]$) and the Gaussian ($\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$) cases with $\sigma = 0.4$ and $a = 0.70$. We compare the Gibbs predictor performances to those of the estimator based on the AIC criterion and to the QMLE in the $AR(q)$ model, so called "full model". For each model, we first simulate a time series of length $2n$, use the observations 1 to n as a learning set and $n+1$ to $2n$ as a test set, for $n = 100$ and $n = 1000$. Each simulation is repeated 20 times and we report in Table 4 the mean and the standard deviation of the empirical quadratic errors for each method and each model.

Note that the Gibbs predictor performs better on Models (8.4) and (8.5) while the AIC predictor performs slightly better on Model (8.3). The difference tends to be negligible when n grows - this is coherent with the fact that we develop here a non-asymptotic theory. Note that the Gibbs predictor performs also well in the case of a Gaussian noise where the boundedness assumption is not satisfied.

References

[AD11] A. Agarwal and J. C. Duchi, *The generalization ability of online algorithms for dependent data*, IEEE Transactions on Information Theory (to appear), 2011.

[Aka73] H. Akaike, *Information theory and an extension of the maximum likelihood principle*, 2nd International Symposium on Information Theory (B. N. Petrov and F. Csaki, eds.), Budapest: Akademia Kiado, 1973, pp. 267–281.

[AL11] P. Alquier and P. Lounici, *PAC-Bayesian bounds for sparse regression estimation with exponential weights*, Electronic Journal of Statistics **5** (2011), 127–145.

[AL12] P. Alquier and X. Li, *Prediction of quantiles bu statistical learning and application to gdp forecasting*, Accepted for DS'12, 2012.

[Alq08] P. Alquier, *PAC-Bayesian bounds for randomized empirical risk minimizers*, Mathematical Methods of Statistics **17** (2008), no. 4, 279–304.

[AP86] K. B. Athreya and S. G. Pantula, *Mixing properties of Harris chains and autoregressive processes*, J. Appl. Probab. **23** (1986), no. 4, 880–892. MR867185 (88c:60127)

[Aud10] J.-Y. Audibert, *PAC-Bayesian aggregation and multi-armed bandits*, HDR Université Paris Est, 2010.

[AW12] P. Alquier and O. Wintenberger, *Model selection for weakly dependent time series forecasting*, Bernoulli **18** (2012), no. 3, 883–193.

[BBR08] G. Biau, O. Biau, and L. Rouvière, *Nonparametric forecasting of the manufacturing output growth with firm-level survey data*, Journal of Business Cycle Measurement and Analysis **3** (2008), 317–332.

[BC11] A. Belloni and V. Chernozhukov, *L1-penalized quantile regression in high-dimensional sparse models*, The Annals of Statistics **39** (2011), no. 1, 82–130.

[BCV01] Yannick Baraud, F. Comte, and G. Viennet, *Model selection for (auto-)regression with dependent data*, ESAIM Probab. Statist. **5** (2001), 33–49. MR1845321 (2002h:62117)

[BD09] P. Brockwell and R. Davis, *Time series: Theory and methods (2nd edition)*, Springer, 2009.

[BFW98] E. Britton, P. Fisher, and J. Whitley, *The inflation report projections: Understanding the fan chart*, Bank of England Quarterly Bulletin **38** (1998), no. 1, 30–37.

[BM01] L. Birgé and P. Massart, *Gaussian model selection*, Journal of the European Mathematical Society **3** (2001), no. 3, 203–268.

[BP11] G. Biau and B. Patra, *Sequential quantile prediction of time series*, IEEE Transactions on Information Theory **57** (2011), 1664–1674.

[Cat] O. Catoni, *A PAC-Bayesian approach to adaptative classification*, 2003.

[Cat04] O. Catoni, *Statistical learning theory and stochastic optimization*, Springer Lecture Notes in Mathematics, 2004.

[Cat07] O. Catoni, *PAC-Bayesian supervised classification (the thermodynamics of statistical learning)*, Lecture Notes-Monograph Series, vol. 56, IMS, 2007.

[CBL06] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*, Cambridge University Press, New York, 2006.

[CM09] L. Clavel and C. Minodier, *A monthly indicator of the french business climate*, Documents de Travail de la DESE, 2009.

[Cor10] M. Cornec, *Constructing a conditional gdp fan chart with an application to french business survey data*, 30th CIRET Conference, New York, 2010.

[DAJJ12] J. C. Duchi, A. Agarwal, M. Johansson, and M. I. Jordan, *Ergodic mirror descent*, Preprint arXiv:1105.4681, 2012.

[DDL+07] J. Dedecker, P. Doukhan, G. Lang, J. R. León, S. Louhichi, and C. Prieur, *Weak dependence, examples and applications*, Lecture Notes in Statistics, vol. 190, Springer-Verlag, Berlin, 2007.

[Dev84] M. Devilliers, *Les enquêtes de conjoncture*, Archives et Documents, no. 101, INSEE, 1984.

[DM06] E. Dubois and E. Michaux, *étalonnages à l'aide d'enquêtes de conjoncture: de nouvaux résultats*, Économie et Prévision, no. 172, INSEE, 2006.

[Dou94] P. Doukhan, *Mixing*, Lecture Notes in Statistics, Springer, New York, 1994.

[Dow04] K. Dowd, *The inflation fan charts: An evaluation*, Greek Economic Review **23** (2004), 99–111.

[DS12] A. Dalalyan and J. Salmon, *Sharp oracle inequalities for aggregation of affine estimators*, The Annals of Statistics (to appear), 2012.

[DT08] A. Dalalyan and A. Tsybakov, *Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity*, Machine Learning **72** (2008), 39–61.

[DTW97] F. X. Diebold, A. S. Tay, and K. F. Wallis, *Evaluating density forecasts of inflation: the survey of professional forecasters*, Discussion Paper No.48, ESRC Macroeconomic Modelling Bureau, University of Warwick and Working Paper No.6228, National Bureau of Economic Research, Cambridge, Mass., 1997.

[DV76] M. D. Donsker and S. S. Varadhan, *Asymptotic evaluation of certain markov process expectations for large time. iii.*, Communications on Pure and Applied Mathematics **28** (1976), 389–461.

[Eng82] R. F. Engle, *Autoregressive conditional heteroscedasticity with estimates of variance of united kingdom inflation*, Econometrica **50** (1982), 987–1008.

[FZ10] C. Francq and J.-M. Zakoian, *Garch models: Structure, statistical inference and financial applications*, Wiley-Blackwell, 2010.

[Ger11] S. Gerchinovitz, *Sparsity regret bounds for individual sequences in online linear regression*, Proceedings of COLT'11, 2011.

[Ham94] J. Hamilton, *Time series analysis*, Princeton University Press, 1994.

[Ibr62] I. A. Ibragimov, *Some limit theorems for stationary processes*, Theory of Probability and its Application **7** (1962), no. 4, 349–382.

[KB78] R. Koenker and G. Jr. Bassett, *Regression quantiles*, Econometrica **46** (1978), 33–50.

[Koe05] R. Koenker, *Quantile regression*, Cambridge University Press, Cambridge, 2005.

[Kul59] S. Kullback, *Information theory and statistics*, Wiley, New York, 1959.

[Lec11] G. Lecué, *Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis*, HDR Thesis, Université Paris-Est Marne-la-Vallée, 2011.

[Li10] X. Li, *Agrégation de prédicteurs appliquée à la conjoncture*, Rapport de stage de M2 - Université Paris 6 - INSEE sous la direction de Matthieu

Cornec, 2010.
[LW94] N. Littlestone and M.K. Warmuth, *The weighted majority algorithm*, Information and Computation **108** (1994), 212–261.
[McA99] D. A. McAllester, *PAC-Bayesian model averaging*, Procs. of of the 12th Annual Conf. On Computational Learning Theory, Santa Cruz, California (Electronic), ACM, New-York, 1999, pp. 164–170.
[Mei00] R. Meir, *Nonparametric time series prediction through adaptive model selection*, Machine Learning **39** (2000), 5–34.
[Min10] C. Minodier, *Avantages comparés des séries premières valeurs publiées et des séries des valeurs révisées*, Documents de Travail de la DESE, 2010.
[MM98] D. S. Modha and E. Masry, *Memory-universal prediction of stationary random processes*, IEEE transactions on information theory **44** (1998), no. 1, 117–133.
[MT93] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Communications and Control Engineering Series, Springer-Verlag London Ltd., London, 1993. MR1287609 (95j:60103)
[MT99] E. Mammen and A. B. Tsybakov, *Smooth discrimination analysis*, The Annals of Statistics **34** (1999), no. 5, 1808–1829.
[R D08] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, 2008.
[Rio00] E. Rio, *Ingalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes*, Comptes Rendus de l'Académie des Sciences de Paris, Série I **330** (2000), 905–908.
[Sam00] P.-M. Samson, *Concentration of measure inequalities for markov chains and ϕ-mixing processes*, The Annals of Probability **28** (2000), no. 1, 416–461.
[SLCB+12] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, J. Peters, and P. Auer, *Pac-bayesian inequalities for martingales*, IEEE Transactions on Information Theory (to appear), 2012.
[Sto10] G. Stoltz, *Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique*, Journal de la SFDS **151** (2010), no. 2, 66–106.
[STW97] J. Shawe-Taylor and R. Williamson, *A PAC analysis of a bayes estimator*, Proceedings of the Tenth Annual Conference on Computational Learning Theory, COLT'97, ACM, 1997, pp. 2–9.
[Tal07] N. N. Taleb, *Black swans and the domains of statistics*, The American Statistician **61** (2007), no. 3, 198–200.
[Tsy03] A. Tsybakov, *Optimal rates of aggregation*, Learning Theory and Kernel Machines (B. Schölkopf and M. K. Warmuth, eds.), Springer LNCS, 2003, pp. 303–313.
[TW00] A. S. Tay and K. F. Wallis, *Density forecasting: a survey*, Journal of Forecasting **19** (2000), 235–254.
[Vap99] V. Vapnik, *The nature of statistical learning theory*, Springer, 1999.
[Vov90] V.G. Vovk, *Aggregating strategies*, Proceedings of the 3rd Annual Workshop on Computational Learning Theory (COLT), 1990, pp. 372–283.
[Win10] O. Wintenberger, *Deviation inequalities for sums of weakly dependent time series*, Electronic Communications in Probability **15** (2010), 489–503.

Appendix A: A general PAC-Bayesian inequality

Theorems 1 and 5 are actually both corollaries of a more general result that we would like to state for the sake of completeness. This result is the analogous of the PAC-Bayesian bounds proved by Catoni in the case of iid data [Cat07].

Theorem 9 (PAC-Bayesian Oracle Inequality for the Gibbs estimator)**.** *Let us assume that* **LowRates**(κ) *is satisfied for some* $\kappa > 0$*. Then, for any* λ, $\varepsilon > 0$ *we have*

$$\mathbb{P}\left\{ R\left(\hat{\theta}_\lambda\right) \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left[\int R \mathrm{d}\rho + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2\mathcal{K}(\rho,\pi) + 2\log\left(2/\varepsilon\right)}{\lambda} \right] \right\} \geq 1-\varepsilon.$$

This result is proved in Appendix B, but we can now provide the proofs of Theorems 1 and 5.

Proof of Theorem 1. We apply Theorem 9 for $\pi = \frac{1}{M}\sum_{\theta\in\Theta}\delta_\theta$ and restrict the inf in the upper bound to Dirac masses $\rho \in \{\delta_\theta, \theta \in \Theta\}$. We obtain $\mathcal{K}(\rho,\pi) = \log M$, and the upper bound for $R(\hat{\theta}_\lambda)$ becomes:

$$\begin{aligned} R\left(\hat{\theta}_\lambda\right) &\leq \inf_{\rho\in\{\delta_\theta,\theta\in\Theta\}} \left[\int R\mathrm{d}\rho + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2\log\left(2M/\varepsilon\right)}{\lambda} \right] \\ &= \inf_{\theta\in\Theta} \left[R(\theta) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2\log\left(2M/\varepsilon\right)}{\lambda} \right]. \end{aligned}$$

■

Proof of Theorem 5. An application of Theorem 9 yields that with probability at least $1-\varepsilon$

$$R(\hat{\theta}_\lambda) \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta)} \left[\int R\mathrm{d}\rho + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2\mathcal{K}(\rho,\pi) + 2\log\left(2/\varepsilon\right)}{\lambda} \right].$$

Let us estimate the upper bound at the probability distribution ρ_δ defined as

$$\frac{\mathrm{d}\rho_\delta}{\mathrm{d}\pi}(\theta) = \frac{\mathbf{1}\{R(\theta) - R(\overline{\theta}) < \delta\}}{\int_{t\in\Theta} \mathbf{1}\{R(t) - R(\overline{\theta}) < \delta\}\pi(\mathrm{d}t)}.$$

Then we have:

$$\begin{aligned} R\left(\hat{\theta}_\lambda\right) \leq \inf_{\delta>0} \Bigg[& R(\overline{\theta}) + \delta + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} \\ & + 2\frac{-\log\int_{t\in\Theta}\mathbf{1}\{R(t) - \inf_\Theta R < \delta\}\pi(\mathrm{d}t) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda} \Bigg]. \end{aligned}$$

Under the assumptions of Theorem 5 we have:

$$R\left(\hat{\theta}_\lambda\right) \leq \inf_{\delta>0}\left[R(\overline{\theta}) + \delta + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + 2\frac{d\log\left(D/\delta\right) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}\right].$$

The infimum is reached for $\delta = d/\lambda$ and we have:

$$R\left(\hat{\theta}_\lambda\right) \leq R(\overline{\theta}) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + 2\frac{d\log\left(D\sqrt{e}\lambda/d\right) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}.$$

■

Appendix B: Proofs

B.1. Preliminaries

We will use Rio's inequality [Rio00] that is an extension of Hoeffding's inequality in a dependent context. For the sake of completeness, we provide here this result when the observations $(X_1, \ldots, X_n)$ come from a stationary process (X_t)

Lemma 1 (Rio [Rio00])**.** *Let h be a function $(\mathbb{R}^p)^n \to \mathbb{R}$ such that for all x_1, ..., x_n, y_1, ..., $y_n \in \mathbb{R}^p$,*

$$|h(x_1, \ldots, x_n) - h(y_1, \ldots, y_n)| \leq \sum_{i=1}^{n} \|x_i - y_i\|. \tag{B.1}$$

Then, for any $t > 0$, we have

$$\mathbb{E}\left(\exp(t\left\{\mathbb{E}\left[h(X_1, \ldots, X_n)\right] - h(X_1, \ldots, X_n)\right\})\right) \leq \exp\left(\frac{t^2 n\left(\mathcal{B} + \theta_{\infty,n}(1)\right)^2}{2}\right).$$

Others exponential inequalities can be used to obtain PAC-Bounds in the context of time series: the inequalities in [Dou94, Sam00] for mixing time series, and [DDL+07, Win10] under weakest "weak dependence" assumptions, [SLCB+12] for martingales. Lemma 1 is very general and yields optimal low rates of convergence. For fast rates of convergence, we will use Samson's inequality that is an extension of Bernstein's inequality in a dependent context.

Lemma 2 (Samson [Sam00])**.** *Let $N \geq 1$, $(Z_i)_{i\in\mathbb{Z}}$ be a stationary process on $\mathbb{R}^k$ and ϕ_r^Z denote its ϕ-mixing coefficients. For any measurable function $f : \mathbb{R}^k \to [-M, M]$, any $0 \leq t \leq 1/(MK_{\phi z}^2)$, we have*

$$\mathbb{E}(\exp(t(S_N(f) - \mathbb{E}S_N(f)))) \leq \exp\left(8K_{\phi z} N\sigma^2(f)t^2\right),$$

where $S_N(f) := \sum_{i=1}^N f(Z_i)$, $K_{\phi z} = 1 + \sum_{r=1}^N \sqrt{\phi_r^Z}$ and $\sigma^2(f) = \mathrm{Var}(f(Z_i))$.

Proof of Lemma 2. This result can be deduced easily from the proof of Theorem 3 of [Sam00] which states a more general result on empirical processes. In page 457 of [Sam00], replace the definition of $f_N(x_1,\dots,x_n)$ by $f_N(x_1,\dots,x_n) = \sum_{i=1}^n g(x_i)$ (following the notations of [Sam00]). Then check that all the arguments of the proof remain valid, the claim of Lemma 2 is obtained page 460, line 7. ■

We also remind the variational formula of the Kullback divergence.

Lemma 3 (Donsker-Varadhan [DV76] variational formula)**.** *For any $\pi \in \mathcal{M}_+^1(E)$, for any measurable upper-bounded function $h : E \to \mathbb{R}$ we have:*

$$\int \exp(h) d\pi = \exp\left(\sup_{\rho \in \mathcal{M}_+^1(E)} \left(\int h d\rho - \mathcal{K}(\rho,\pi) \right) \right). \tag{B.2}$$

Moreover, the supremum with respect to ρ in the right-hand side is reached for the Gibbs measure $\pi\{h\}$ defined by $\pi\{h\}(\mathrm{d}x) = e^{h(x)}\pi(\mathrm{d}x)/\pi[\exp(h)]$.

Actually, it seems that in the case of discrete probabilities, this result was already known by Kullback (Problem 8.28 of Chapter 2 in [Kul59]). For a complete proof of this variational formula, even in the non integrable cases, we refer the reader to [DV76, Cat, Cat07].

B.2. Technical lemmas for the proofs of Theorems 2, 6, 7 and 9

Lemma 4. *We assume that* **LowRates(κ)** *is satisfied for some $\kappa > 0$. For any $\lambda > 0$ and $\theta \in \Theta$ we have*

$$\mathbb{E}\left(e^{\lambda(R(\theta)-r_n(\theta))}\right) \vee E\left(e^{\lambda(r_n(\theta)-R(\theta))}\right) \leq \exp\left(\frac{\lambda^2\kappa^2}{n\left(1-k/n\right)^2}\right).$$

Proof of Lemma 4. Let us fix $\lambda > 0$ and $\theta \in \Theta$. Let us define the function h by:

$$h(x_1,\dots,x_n) = \frac{1}{K(1+L)} \sum_{i=k+1}^{n} \ell(f_\theta(x_{i-1},\dots,x_{i-k}),x_i).$$

We now check that h satisfies (B.1), remember that $\ell(x,x') = g(x-x')$ so

$$\begin{aligned}
&\left|h\left(x_1,\dots,x_n\right) - h\left(y_1,\dots y_n\right)\right| \\
&\quad\leq \frac{1}{K(1+L)} \sum_{i=k+1}^{n} \left| g(f_\theta(x_{i-1},\dots,x_{i-k}) - x_i) - g(f_\theta(y_{i-1},\dots,y_{i-k}) - y_i) \right| \\
&\quad\leq \frac{1}{1+L} \sum_{i=k+1}^{n} \left\| \left(f_\theta(x_{i-1},\dots,x_{i-k}) - x_i\right) - \left(f_\theta(y_{i-1},\dots,y_{i-k}) - y_i\right) \right\|
\end{aligned}$$

where we used Assumption **LipLoss**(K) for the last inequality. So we have

$$\begin{aligned}
&\Big|h\left(x_1,\dots,x_n\right)-h\left(y_1,\dots y_n\right)\Big| \\
&\quad\leq \frac{1}{1+L}\sum_{i=k+1}^{n}\left(\Big\|f_\theta(x_{i-1},\dots,x_{i-k})-f_\theta(y_{i-1},\dots,y_{i-k})\Big\|+\Big\|x_i-y_i\Big\|\right) \\
&\quad\leq \frac{1}{1+L}\sum_{i=k+1}^{n}\left(\sum_{j=1}^{k}a_j(\theta)\|x_{i-j}-y_{i-j}\|+\|x_i-y_i\|\right) \\
&\quad\leq \frac{1}{1+L}\sum_{i=1}^{n}\left(1+\sum_{j=1}^{k}a_j(\theta)\right)\|x_i-y_i\|\leq\sum_{i=1}^{n}\|x_i-y_i\|
\end{aligned}$$

where we used Assumption **Lip**(L). So we can apply Lemma 1 with $h(X_1,\dots,X_n)=\frac{n-k}{K(1+L)}r_n(\theta)$, $\mathbb{E}(h(X_1,\dots,X_n))=\frac{n-k}{K(1+L)}R(\theta)$, and $t=K(1+L)\lambda/(n-k)$:

$$\begin{aligned}
\mathbb{E}\left(e^{\lambda[R(\theta)-r_n(\theta)]}\right)&\leq\exp\left(\frac{\lambda^2K^2(1+L)^2\left(\mathcal{B}+\theta_{\infty,n}(1)\right)^2}{2n\left(1-k/n\right)^2}\right)\\
&\leq\exp\left(\frac{\lambda^2K^2(1+L)^2\left(\mathcal{B}+\mathcal{C}\right)^2}{2n\left(1-\frac{k}{n}\right)^2}\right)
\end{aligned}$$

by Assumption **WeakDep**$(\mathcal{C})$. This ends the proof of the first inequality. The reverse inequality is obtained by replacing the function h by $-h$. ■

We are now ready to state the following key Lemma.

Lemma 5. *Let us assume that* **LowRates**(κ) *is satisfied satisfied for some* $\kappa>0$*. Then for any* $\lambda>0$ *we have*

$$\mathbb{P}\left\{\begin{array}{l}\forall\rho\in\mathcal{M}_+^1(\Theta),\\ \int R\mathrm{d}\rho\leq\int r_n\mathrm{d}\rho+\frac{\lambda\kappa^2}{n(1-k/n)^2}+\frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}\\ \quad\textit{and}\\ \int r_n\mathrm{d}\rho\leq\int R\mathrm{d}\rho+\frac{\lambda\kappa^2}{n(1-k/n)^2}+\frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}\end{array}\right\}\geq 1-\varepsilon. \tag{B.3}$$

Proof of Lemma 5. Let us fix $\theta>0$ and $\lambda>0$, and apply the first inequality of Lemma 4. We have:

$$\mathbb{E}\Big(\exp\left(\lambda\Big(R(\theta)-r_n(\theta)-\frac{\lambda\kappa^2}{n\left(1-k/n\right)^2}\Big)\right)\Big)\leq 1,$$

and we multiply this result by $\varepsilon/2$ and integrate it with respect to $\pi(\mathrm{d}\theta)$. An application of Fubini's Theorem yields

$$\mathbb{E}\int\exp\left(\lambda(R(\theta)-r_n(\theta))-\frac{\lambda^2\kappa^2}{n\left(1-k/n\right)^2}-\log\left(2/\varepsilon\right)\right)\pi(\mathrm{d}\theta)\leq\frac{\varepsilon}{2}.$$

We apply Lemma 3 and we get:

$$\mathbb{E}\exp\left(\sup_{\rho}\left\{\lambda\int(R(\theta)-r_n(\theta))\rho(\mathrm{d}\theta)-\frac{\lambda^2\kappa^2}{n\left(1-k/n\right)^2}-\log\left(2/\varepsilon\right)-\mathcal{K}(\rho,\pi)\right\}\right)\leq\frac{\varepsilon}{2}.$$

As $e^x\geq\mathbf{1}_{\mathbb{R}_+}(x)$, we have:

$$\mathbb{P}\left\{\sup_{\rho}\left\{\lambda\int\left(R(\theta)-r_n(\theta)\right)\rho(\mathrm{d}\theta)-\frac{\lambda^2\kappa^2}{n\left(1-k/n\right)^2}-\log\left(2/\varepsilon\right)-\mathcal{K}(\rho,\pi)\right\}\geq 0\right\}\leq\frac{\varepsilon}{2}.$$

Using the same arguments than above but starting with the second inequality of Lemma 4:

$$\mathbb{E}\exp\left(\lambda\Big(r_n(\theta)-R(\theta)-\frac{\lambda\kappa^2}{n\left(1-k/n\right)^2}\Big)\right)\leq 1.$$

we obtain:

$$\mathbb{P}\left\{\sup_{\rho}\left\{\lambda\int\left[r_n(\theta)-R(\theta)\right]\rho(\mathrm{d}\theta)-\frac{\lambda^2\kappa^2}{n\left(1-\frac{k}{n}\right)^2}-\log\left(\frac{2}{\varepsilon}\right)-\mathcal{K}(\rho,\pi)\right\}\geq 0\right\}\leq\frac{\varepsilon}{2}.$$

A union bound ends the proof. ■

The following variant of Lemma 5 will also be useful.

Lemma 6. *Let us assume that* **LowRates**(κ) *is satisfied satisfied for some* $\kappa>0$. *Then for any* $\lambda>0$ *we have*

$$\mathbb{P}\left\{\begin{array}{l}\forall\rho\in\mathcal{M}_+^1(\Theta),\\ \int R\mathrm{d}\rho\leq\int r_n\mathrm{d}\rho+\frac{\lambda\kappa^2}{n(1-k/n)^2}+\frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}\\ \quad and\\ r_n(\overline{\theta})\leq R(\overline{\theta})+\frac{\lambda\kappa^2}{n(1-k/n)^2}+\frac{\log(2/\varepsilon)}{\lambda}\end{array}\right\}\geq 1-\varepsilon.$$

Proof of Lemma 6. Following the proof of Lemma 5 we have:

$$\mathbb{P}\left\{\sup_{\rho}\left\{\lambda\int\left(R(\theta)-r_n(\theta)\right)\rho(\mathrm{d}\theta)-\frac{\lambda^2\kappa^2}{n\left(1-k/n\right)^2}-\log\left(2/\varepsilon\right)-\mathcal{K}(\rho,\pi)\right\}\geq 0\right\}\leq\frac{\varepsilon}{2}.$$

Now, we use the second inequality of Lemma 4, with $\theta=\overline{\theta}$:

$$\mathbb{E}\Big(\exp\left(\lambda\Big(r_n(\overline{\theta})-R(\overline{\theta})-\frac{\lambda\kappa^2}{n\left(1-k/n\right)^2}\Big)\right)\Big)\leq 1.$$

But then, we directly apply Markov's inequality to get:

$$\mathbb{P}\left\{r_n(\overline{\theta})\geq R(\overline{\theta})+\frac{\lambda\kappa^2}{n\left(1-k/n\right)^2}+\frac{\log\left(2/\varepsilon\right)}{\lambda}\right\}\leq\frac{\varepsilon}{2}.$$

Here again, a union bound ends the proof. ■

B.3. Proof of Theorems 9 and 7

In this subsection we prove the general result on the Gibbs predictor.

Proof of Theorem 9. We apply Lemma 5. So, with probability at least $1-\varepsilon$ we are on the event given by (B.3). From now, we work on that event. The first inequality of (B.3), when applied to $\hat{\rho}_\lambda(\mathrm{d}\theta)$, gives

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \leq \int r_n(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) + \frac{\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{1}{\lambda}\log\left(2/\varepsilon\right) + \frac{1}{\lambda}\mathcal{K}(\hat{\rho}_\lambda,\pi).$$

According to Lemma 3 we have:

$$\int r_n(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) + \frac{1}{\lambda}\mathcal{K}(\hat{\rho}_\lambda,\pi) = \inf_\rho\left(\int r_n(\theta)\rho(\mathrm{d}\theta) + \frac{1}{\lambda}\mathcal{K}(\rho,\pi)\right)$$

so we obtain

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \leq \inf_\rho\left\{\int r_n(\theta)\rho(\mathrm{d}\theta) + \frac{\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\mathcal{K}(\rho,\pi)+\log\left(2/\varepsilon\right)}{\lambda}\right\}. \tag{B.4}$$

We now estimate from above $r(\theta)$ by $R(\theta)$. Applying the second inequality of (B.3) and plugging it into Inequality B.4 gives

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \leq \inf_\rho\left\{\int R\mathrm{d}\rho + \frac{2}{\lambda}\mathcal{K}(\rho,\pi) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2}{\lambda}\log\left(2/\varepsilon\right)\right\}.$$

We end the proof by the remark that $\theta \mapsto R(\theta)$ is convex and so by Jensen's inequality $\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \geq R\left(\int \theta\hat{\rho}_\lambda(\mathrm{d}\theta)\right) = R(\hat{\theta}_\lambda)$. ■

Proof of Theorem 7. Let us apply Lemma 5 in each model Θ_j, with a fixed $\lambda_j > 0$ and confidence level $\varepsilon_j > 0$. We obtain, for all j,

$$\mathbb{P}\left\{\begin{array}{l}\forall\rho\in\mathcal{M}_+^1(\Theta_j),\\ \int R\mathrm{d}\rho \leq \int r_n\mathrm{d}\rho + \frac{\lambda_j\kappa_j^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi_j)+\log(2/\varepsilon_j)}{\lambda_j}\\ \text{and}\\ \int r_n\mathrm{d}\rho \leq \int R\mathrm{d}\rho + \frac{\lambda_j\kappa_j^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi_j)+\log(2/\varepsilon_j)}{\lambda_j}\end{array}\right\} \geq 1-\varepsilon_j.$$

We put $\varepsilon_j = p_j\varepsilon$, a union bound gives leads to:

$$\mathbb{P}\left\{\begin{array}{l}\forall j\in\{1,...,M\},\quad \forall\rho\in\mathcal{M}_+^1(\Theta_j),\\ \int R\mathrm{d}\rho \leq \int r_n\mathrm{d}\rho + \frac{\lambda_j\kappa_j^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi_j)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j}\\ \text{and}\\ \int r_n\mathrm{d}\rho \leq \int R\mathrm{d}\rho + \frac{\lambda_j\kappa_j^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi_j)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j}\end{array}\right\} \geq 1-\varepsilon. \tag{B.5}$$

From now, we only work on that event of probability at least $1-\varepsilon$. Remark that

$$R(\hat{\theta}) = R(\hat{\theta}_{\lambda_{\hat{j}},\hat{j}})$$

$$\leq \int R(\theta)\hat{\rho}_{\lambda_{\hat{j}},\hat{j}}(\mathrm{d}\theta) \text{ by Jensen's inequality}$$

$$\leq \int r_n\hat{\rho}_{\lambda_{\hat{j}},\hat{j}}(\mathrm{d}\theta) + \frac{\lambda_j\kappa_j^2}{n\left(1-k/n\right)^2} + \frac{\mathcal{K}(\hat{\rho}_{\lambda_{\hat{j}},\hat{j}},\pi_j) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j}$$

by (B.5)

$$= \inf_{1\leq j\leq M}\left\{\int r_n\hat{\rho}_{\lambda_j,j}(\mathrm{d}\theta) + \frac{\lambda_j\kappa_j^2}{n\left(1-k/n\right)^2} + \frac{\mathcal{K}(\hat{\rho}_{\lambda_j,j},\pi_j) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j}\right\}$$

by definition of $\hat{j}$

$$= \inf_{1\leq j\leq M}\inf_{\rho\in\mathcal{M}_+^1(\Theta_j)}\left\{\int r_n\rho(\mathrm{d}\theta) + \frac{\lambda_j\kappa_j^2}{n\left(1-k/n\right)^2} + \frac{\mathcal{K}(\rho,\pi_j) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j}\right\}$$

by Lemma 3

$$\leq \inf_{1\leq j\leq M}\inf_{\rho\in\mathcal{M}_+^1(\Theta_j)}\left\{\int R\rho(\mathrm{d}\theta) + \frac{2\lambda_j\kappa_j^2}{n\left(1-k/n\right)^2} + 2\frac{\mathcal{K}(\rho,\pi_j) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j}\right\}$$

by (B.5) again

$$\leq \inf_{1\leq j\leq M}\inf_{\delta>0}\left\{R(\overline{\theta}_j) + \delta + \frac{2\lambda_j\kappa_j^2}{n\left(1-k/n\right)^2} + 2\frac{d_j\log\left(D_j/\delta\right) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j}\right\}$$

by restricting ρ as in the proof of Cor. 5 page 10

$$\leq \inf_{1\leq j\leq M}\left\{R(\overline{\theta}_j) + \frac{2\lambda_j\kappa_j^2}{n\left(1-k/n\right)^2} + 2\frac{d_j\log\left(\frac{D_j e\lambda_j}{d_j}\right) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda_j}\right\}$$

by taking $\delta = \dfrac{d_j}{\lambda_j}$

$$= \inf_{1\leq j\leq M}\left\{R(\overline{\theta}_j) + \inf_{\lambda>0}\left\{\frac{2\lambda\kappa_j^2}{n\left(1-k/n\right)^2} + 2\frac{d_j\log\left(\frac{D_j e\lambda}{d_j}\right) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda}\right\}\right\}$$

by definition of λ_j

$$\leq \inf_{1\leq j\leq M}\left\{R(\overline{\theta}_j) + 2\frac{\kappa_j}{1-k/n}\left\{\sqrt{\frac{d_j}{n}}\log\left(\frac{D_j e^2}{\kappa_j}\sqrt{\frac{n}{d_j}}\right) + \frac{\log\left(\frac{2}{\varepsilon p_j}\right)}{\sqrt{nd_j}}\right\}\right\}.$$

■

B.4. Proof of Theorems 2 and 6

Let us now prove the results about the ERM.

Proof of Theorem 2. We choose π as the uniform probability distribution on Θ and $\lambda > 0$. We apply Lemma 6. So we have, with probability at least $1-\varepsilon$,

$$\left\{ \begin{array}{ll} \forall \rho \in \mathcal{M}_+^1(\Theta'), & \int R\mathrm{d}\rho \leq \int r_n \mathrm{d}\rho + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda} \\ \text{and} & r_n(\overline{\theta}) \leq R(\overline{\theta}) + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\log(2/\varepsilon)}{\lambda}. \end{array} \right.$$

We restrict the inf in the first inequality to Dirac masses $\rho \in \{\delta_\theta, \theta \in \Theta\}$ and we obtain:

$$\left\{ \begin{array}{ll} \forall \theta \in \Theta, & R(\theta) \leq r_n(\theta) + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\log\left(\frac{2M}{\varepsilon}\right)}{\lambda} \\ \text{and} & r_n(\overline{\theta}) \leq R(\overline{\theta}) + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\log(2/\varepsilon)}{\lambda}. \end{array} \right.$$

In particular, we apply the first inequality to $\hat{\theta}^{ERM}$. We remind that $\overline{\theta}$ minimizes R on Θ and that $\hat{\theta}^{ERM}$ minimizes r_n on Θ, and so we have

$$\begin{aligned} R(\hat{\theta}^{ERM}) &\leq r_n(\hat{\theta}^{ERM}) + \frac{\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\log(M)+\log\left(2/\varepsilon\right)}{\lambda} \\ &\leq r_n(\overline{\theta}) + \frac{\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\log(M)+\log\left(2/\varepsilon\right)}{\lambda} \\ &\leq R(\overline{\theta}) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\log(M)+2\log\left(2/\varepsilon\right)}{\lambda} \\ &\leq R(\overline{\theta}) + \frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2\log\left(2M/\varepsilon\right)}{\lambda}. \end{aligned}$$

The result still holds if we choose λ as a minimizer of

$$\frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{2\log\left(2M/\varepsilon\right)}{\lambda}.$$

■

Proof of Theorem 6. We put $\Theta' = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq D+1\}$. We choose π as the uniform probability distribution on Θ'. We apply Lemma 6. So we have, with probability at least $1-\varepsilon$,

$$\left\{ \begin{array}{ll} \forall \rho \in \mathcal{M}_+^1(\Theta'), & \int R\mathrm{d}\rho \leq \int r_n \mathrm{d}\rho + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda} \\ \text{and} & r_n(\overline{\theta}) \leq R(\overline{\theta}) + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\log(2/\varepsilon)}{\lambda}. \end{array} \right.$$

So for any ρ,

$$\begin{aligned} R(\hat{\theta}^{ERM}) &= \int [R(\hat{\theta}^{ERM}) - R(\theta)]\rho(\mathrm{d}\theta) + \int R\mathrm{d}\rho \\ &\leq \int [R(\hat{\theta}^{ERM}) - R(\theta)]\rho(\mathrm{d}\theta) + \int r_n \mathrm{d}\rho + \frac{\lambda\kappa^2}{n\left(1-k/n\right)^2} + \frac{\mathcal{K}(\rho,\pi)+\log\left(2/\varepsilon\right)}{\lambda} \\ &\leq \int [R(\hat{\theta}^{ERM}) - R(\theta)]\rho(\mathrm{d}\theta) + \int [r_n(\theta) - r_n(\hat{\theta}^{ERM})]\rho(\mathrm{d}\theta) + r_n(\hat{\theta}^{ERM}) \end{aligned}$$

$$+\frac{\lambda\kappa^2}{n\left(1-k/n\right)^2}+\frac{\mathcal{K}(\rho,\pi)+\log\left(2/\varepsilon\right)}{\lambda}$$

$$\leq 2K\psi\int\|\theta-\hat{\theta}^{ERM}\|_1\rho(\mathrm{d}\theta)+r_n(\overline{\theta})+\frac{\lambda\kappa^2}{n\left(1-k/n\right)^2}+\frac{\mathcal{K}(\rho,\pi)+\log\left(2/\varepsilon\right)}{\lambda}$$

$$\leq 2K\psi\int\|\theta-\hat{\theta}^{ERM}\|_1\rho(\mathrm{d}\theta)+R(\overline{\theta})+\frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2}+\frac{\mathcal{K}(\rho,\pi)+2\log\left(2/\varepsilon\right)}{\lambda}.$$

Now we define, for any $\delta>0$, ρ_δ by

$$\frac{\mathrm{d}\rho_\delta}{\mathrm{d}\pi}(\theta)=\frac{\mathbf{1}\{\|\theta-\hat{\theta}^{ERM}\|<\delta\}}{\int_{t\in\Theta'}\mathbf{1}\{\|t-\hat{\theta}^{ERM}\|<\delta\}\pi(\mathrm{d}t)}.$$

So in particular, we have, for any $\delta>0$,

$$R(\hat{\theta}^{ERM})\leq 2K\psi\delta+R(\overline{\theta})$$
$$+\frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2}+\frac{\log\frac{1}{\int_{t\in\Theta'}\mathbf{1}\{\|t-\hat{\theta}^{ERM}\|<\delta\}\pi(\mathrm{d}t)}+2\log\left(2/\varepsilon\right)}{\lambda}.$$

But for any $\delta\leq 1$,

$$-\log\int_{t\in\Theta'}\mathbf{1}\{\|t-\hat{\theta}^{ERM}\|<\delta\}\pi(\mathrm{d}t)=d\log\left(\frac{D+1}{\delta}\right).$$

So we have

$$R(\hat{\theta}^{ERM})\leq\inf_{\delta\leq 1}\left\{2K\psi\delta+R(\overline{\theta})+\frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2}+\frac{d\log\left(\frac{D+1}{\delta}\right)+2\log\left(2/\varepsilon\right)}{\lambda}\right\}.$$

We optimize this result by taking $\delta=d/(2\lambda K\psi)$, which is smaller than 1 as soon as $t\geq 2K\psi/d$, we get:

$$R(\hat{\theta}^{ERM})\leq R(\overline{\theta})+\frac{2\lambda\kappa^2}{n\left(1-k/n\right)^2}+\frac{d\log\left(\frac{2eK\psi(D+1)t}{d}\right)+2\log\left(2/\varepsilon\right)}{\lambda}.$$

We just choose λ as the minimizer of the r.h.s., subject to $t\geq 2K\psi/d$, to end the proof. ■

B.5. Some preliminary lemmas for the proof of Theorem 8

Lemma 7. *Under the hypothesis of Theorem 8, we have, for any $\theta\in\Theta$, for any $0\leq\lambda\leq(n-k)/(2kKL\mathcal{BC})$,*

$$\mathbb{E}\exp\left\{\lambda\left[\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)\left(R(\theta)-R(\overline{\theta})\right)-r(\theta)+r(\overline{\theta})\right]\right\}\leq 1,$$

and

$$\mathbb{E}\exp\left\{\lambda\left[\left(1+\frac{8k\mathcal{C}\lambda}{n-k}\right)\left(R(\overline{\theta})-R(\theta)\right)-r(\overline{\theta})+r(\theta)\right]\right\}\leq 1.$$

Lemma 7. We apply Lemma 2 to $N = n-k$, $Z_i = (X_{i+1}, \dots, X_{i+k})$,

$$f(Z_i) = \frac{1}{n-k}\Big[R(\theta) - R(\overline{\theta}) \\ - \ell\left(X_{i+k}, f_\theta(X_{i+k-1}, \dots, X_{i+1})\right) + \ell\left(X_{i+k}, f_{\overline{\theta}}(X_{i+k-1}, \dots, X_{i+1})\right)^2\Big],$$

and so

$$S_N(f) = [R(\theta) - R(\overline{\theta}) - r(\theta) + r(\overline{\theta})],$$

and the Z_i are uniformly mixing with coefficients $\phi_r^Z = \phi_{\lfloor r/q \rfloor}$. Note that $1 + \sum_{r=1}^{n-q}\sqrt{\phi_r^Z} = 1 + \sum_{r=1}^{n-q}\sqrt{\phi_{\lfloor r/k \rfloor}} \leq k\mathcal{C}$ by **PhiMix**$(\mathcal{C})$. For any θ and θ' in Θ let us put

$$V(\theta, \theta') = \mathbb{E}\left\{\left[\ell\Big(X_{k+1}, f_\theta(X_k, ..., X_1)\Big) - \ell\Big(X_{k+1}, f_{\theta'}(X_k, ..., X_1)\Big)\right]^2\right\}.$$

We are going to apply Lemma 2. Remark that $\sigma^2(f) \leq V(\theta, \overline{\theta})/(n-k)^2$. Also,

$$\Big|\ell\left(X_{i+k}, f_\theta(X_{i+k-1}, \dots, X_{i+1})\right) - \ell\left(X_{i+k}, f_{\overline{\theta}}(X_{i+k-1}, \dots, X_{i+1})\right)\Big| \\ \leq K\left|f_\theta(X_{i+k-1}, \dots, X_{i+1}) - f_{\overline{\theta}}(X_{i+k-1}, \dots, X_{i+1})\right| \leq KL\mathcal{B}$$

where we used LipLoss(K) for the first inequality and Lip(L) and PhiMix$(\mathcal{B}, \mathcal{C})$ for the second inequality. This implies that $\|f\|_\infty \leq 2KL\mathcal{B}/(n-k)$, so we can apply Lemma 2 for any $0 \leq \lambda \leq (n-k)/(2kKL\mathcal{B}\mathcal{C})$], we have

$$\ln \mathbb{E}\exp\left[\lambda\Big(R(\theta) - R(\overline{\theta}) - r(\theta) + r(\overline{\theta})\Big)\right] \leq \frac{8k\mathcal{C}V(\theta, \overline{\theta})\lambda^2}{n-k}.$$

Notice finally that Margin$(\mathcal{K})$ leads to

$$V(\theta, \overline{\theta}) = \mathcal{K}\left[R(\theta) - R(\overline{\theta})\right]$$

This proves the first inequality of Lemma 7. The second inequality is proved exacly in the same way, but replacing f by $-f$. □

We are now ready to state the following key Lemma.

Lemma 8. *Under the hypothesis of Theorem 8, we have, for any* $0 \leq \lambda \leq (n-k)/(2kKL\mathcal{B}\mathcal{C})$*, for any* $0 < \varepsilon < 1$*,*

$$\mathbb{P}\left\{\begin{array}{l} \forall \rho \in \mathcal{M}_+^1(\Theta), \\ \left(1 - \frac{8k\mathcal{C}\lambda}{n-k}\right)\left(\int R\mathrm{d}\rho - R(\overline{\theta})\right) \leq \int r\mathrm{d}\rho - r(\overline{\theta}) + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda} \\ \textit{and} \\ \int r\mathrm{d}\rho - r(\overline{\theta}) \leq \left(\int R\mathrm{d}\rho - R(\overline{\theta})\right)\left(1 + \frac{8k\mathcal{C}\lambda}{n-k}\right) + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda} \end{array}\right\} \geq 1 - \varepsilon.$$

Proof of Lemma 8. Let us fix ε, λ and $\theta \in \Theta$, and apply the first inequality of Lemma 7. We have:

$$\mathbb{E}\exp\left\{\lambda\left[\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)\left(R(\theta)-R(\overline{\theta})\right)-r(\theta)+r(\overline{\theta})\right]\right\}\leq 1,$$

and we multiply this result by $\varepsilon/2$ and integrate it with respect to $\pi(\mathrm{d}\theta)$. Fubini's Theorem gives:

$$\mathbb{E}\int\exp\left\{\lambda\left[\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)\left(R(\theta)-R(\overline{\theta})\right)-r(\theta)+r(\overline{\theta})+\log(\epsilon/2)\right]\right\}\pi(\mathrm{d}\theta)\leq\frac{\varepsilon}{2}.$$

We apply Lemma 3 and we get:

$$\mathbb{E}\exp\left\{\sup_{\rho}\lambda\left[\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)\left(\int R\mathrm{d}\rho-R(\overline{\theta})\right)-\int r\mathrm{d}\rho+r(\overline{\theta})+\log(\epsilon/2)-\mathcal{K}(\rho,\pi)\right]\right\}\leq\frac{\varepsilon}{2}.$$

As $e^x \geq \mathbf{1}_{\mathbb{R}_+}(x)$, we have:

$$\mathbb{P}\left\{\sup_{\rho}\lambda\left[\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)\left(\int R\mathrm{d}\rho-R(\overline{\theta})\right)-\int r\mathrm{d}\rho+r(\overline{\theta})+\log(\epsilon/2)\right]-\mathcal{K}(\rho,\pi)\geq 0\right\}\leq\frac{\varepsilon}{2}.$$

Let us apply the same arguments starting with the second inequality of Lemma 7. We obtain:

$$\mathbb{P}\left\{\sup_{\rho}\lambda\left[\left(1+\frac{8k\mathcal{C}\lambda}{n-k}\right)\left(R(\overline{\theta})-\int R\mathrm{d}\rho\right)-r(\overline{\theta})+\int r\mathrm{d}\rho+\log(\epsilon/2)-\mathcal{K}(\rho,\pi)\right]\geq 0\right\}\leq\frac{\varepsilon}{2}.$$

A union bound ends the proof. ■

B.6. Proof of Theorem 8

Proof of Theorem 8. Fix $0 \leq \lambda = (n-k)/(4kKL\mathcal{BC}) \wedge (n-k)/(16k\mathcal{C}) \leq (n-k)/(2kKL\mathcal{BC})$. Applying Lemma 8, we assume from now that the event of

probability at least $1-\varepsilon$ given by this lemma is satisfied. In particular we have $\forall \rho \in \mathcal{M}_+^1(\Theta)$,

$$\int R\mathrm{d}\rho - R(\overline{\theta}) \leq \frac{\int r\mathrm{d}\rho - r(\overline{\theta}) + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}}{\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)}.$$

In particular, thanks to Lemma 3, we have:

$$\int R\mathrm{d}\hat{\rho}_\lambda - R(\overline{\theta}) \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta)} \frac{\int r\mathrm{d}\rho - r(\overline{\theta}) + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}}{\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)}.$$

Now, we apply the second inequality of Lemma 8:

$$\begin{aligned}
&\int R\mathrm{d}\hat{\rho}_\lambda - R(\overline{\theta}) \\
&\qquad \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta)} \frac{\left(1+\frac{8k\mathcal{C}\lambda}{n-k}\right)\left[\int R\mathrm{d}\rho - R(\overline{\theta})\right] + 2\frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}}{\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)} \\
&\qquad \leq \inf_j \inf_{\rho\in\mathcal{M}_+^1(\Theta_j)} \frac{\left(1+\frac{8k\mathcal{C}\lambda}{n-k}\right)\left[\int R\mathrm{d}\rho - R(\overline{\theta})\right] + 2\frac{\mathcal{K}(\rho_j,\pi)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda}}{\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)} \\
&\qquad \leq \inf_j \inf_{\delta>0} \frac{\left(1+\frac{8k\mathcal{C}\lambda}{n-k}\right)\left[R(\overline{\theta}_j) + \delta - R(\overline{\theta})\right] + 2\frac{d_j\log\left(\frac{D_j}{\delta}\right)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda}}{\left(1-\frac{8k\mathcal{C}\lambda}{n-k}\right)}
\end{aligned}$$

by restricting ρ as in the proof of Theorem 5. First, notice that our choice $\lambda \leq (n-k)/(16k\mathcal{C})$ leads to

$$\begin{aligned}
\int R\mathrm{d}\hat{\rho}_\lambda - R(\overline{\theta}) &\leq 2\inf_j \inf_{\delta>0} \left\{ \frac{3}{2}\left[R(\overline{\theta}_j) + \delta - R(\overline{\theta})\right] + 2\frac{d_j\log\left(\frac{D_j}{\delta}\right)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda} \right\} \\
&\leq 4\inf_j \inf_{\delta>0} \left\{ R(\overline{\theta}_j) + \delta - R(\overline{\theta}) + \frac{d_j\log\left(\frac{D_j}{\delta}\right)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda} \right\}.
\end{aligned}$$

Taking $\delta = d_j/\lambda$ leads to

$$\int R\mathrm{d}\hat{\rho}_\lambda - R(\overline{\theta}) \leq 4\inf_j \left\{ R(\overline{\theta}_j) - R(\overline{\theta}) + \frac{d_j\log\left(\frac{D_j e\lambda}{d_j}\right)+\log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda} \right\}.$$

Finally, we replace the last occurences of λ by its value:

$$\int R\mathrm{d}\hat{\rho}_\lambda - R(\overline{\theta})$$
$$\leq 4 \inf_j \left\{ R(\overline{\theta}_j) - R(\overline{\theta}) + (16k\mathcal{C} \vee 4kKL\mathcal{BC}) \frac{d_j \log\left(\frac{D_j e(n-k)}{16k\mathcal{C}d_j}\right) + \log\left(\frac{2}{\varepsilon p_j}\right)}{n-k} \right\}.$$

Jensen's inequality leads to:

$$R\left(\hat{\theta}_\lambda\right) - R(\overline{\theta})$$
$$\leq 4 \inf_j \left\{ R(\overline{\theta}_j) - R(\overline{\theta}) + 4k\mathcal{C}\,(4 \vee KL\mathcal{B}) \frac{d_j \log\left(\frac{D_j e(n-k)}{16k\mathcal{C}d_j}\right) + \log\left(\frac{2}{\varepsilon p_j}\right)}{n-k} \right\}.$$

■