Title	Prediction of time series by statistical learning: general losses and fast rates
Authors(s)	Alquier, Pierre; Li, Xiaoyin; Wintenberger, Olivier
Publication date	2013-12-31
Publication information	Dependence Modelling, 1 : 65-93
Publisher	De Gruyter
Item record/more information	http://hdl.handle.net/10197/8374
Publisher's version (DOI)	10.2478/demo-2013-0004

Dependence Modeling
Research Article • DOI: 10.2478/demo-2013-0004 • DeMo • 2013 • 65-93
Proofreading version

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

Abstract

We establish rates of convergences in statistical learning for time series forecasting. Using the PAC-Bayesian approach, slow rates of convergence $\sqrt{d/n}$ for the Gibbs estimator under the absolute loss were given in a previous work [7], where n is the sample size and d the dimension of the set of predictors. Under the same weak dependence conditions, we extend this result to any convex Lipschitz loss function. We also identify a condition on the parameter space that ensures similar rates for the classical penalized ERM procedure. We apply this method for quantile forecasting of the French GDP. Under additional conditions on the loss functions (satisfied by the quadratic loss function) and for uniformly mixing processes, we prove that the Gibbs estimator actually achieves fast rates of convergence d/n. We discuss the optimality of these different rates pointing out references to lower bounds when they are available. In particular, these results bring a generalization the results of [29] on sparse regression estimation to some autoregression.

Keywords

Statistical learning theory • time series forecasting • PAC-Bayesian bounds • weak dependence • mixing • oracle inequalities • fast rates • GDP forecasting

MSC: 62M20; 60G25; 62M10; 62P20; 65G15; 68Q32; 68T05

Pierre Alquier[1,2*], Xiaoyin Li[3], Olivier Wintenberger[4,5]

1 University College Dublin, School of Mathematical Sciences
2 INSIGHT Centre for Data Analytics
3 Université de Cergy, Laboratoire Analyse Géométrie Modélisation
4 Université Paris-Dauphine, CEREMADE
5 ENSAE, CREST

Received 23 October 2013
Accepted 8 December 2013

1. Introduction

Time series forecasting is a fundamental subject in the statistical processes literature. The parametric approach contains a wide range of models associated with efficient estimation and prediction procedures [36]. Classical parametric models include linear processes such as ARMA models [10]. More recently, non-linear processes such as stochastic volatility and ARCH models received a lot of attention in financial applications - see, among others, the Nobel awarded paper [33], and [34] for a survey of more recent advances. However, parametric assumptions rarely hold on data. Assuming that the observations satisfy a model can bias the prediction and highly underevaluate the risks, see the polemical but highly informative discussion in [61].

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 restricted set of initial predictors Θ, without

* E-mail: pierre.alquier@ucd.ie

any parametric assumption on the distribution of the observed time series. Note however that 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 quantifications of the objective, and different terminologies:

- in the "batch" approach, the family of predictors is sometimes referred to as "model" or "set of concepts". All the observations are given at the same time; the sample $X_1, \dots, X_n$ is modelled as random. Some hypotheses like mixing or weak dependence are required: see [7, 21, 37, 47, 49, 55, 56, 66, 67].

- in the "online" approach, predictors are usually referred to as "experts". At each date t, a prediction of the future realization x_{t+1} is based on the previous observations $x_1, \dots, x_t$, the objective being to minimize the cumulative prediction loss, see [18, 59] for an introduction. The observation are often modeled as deterministic in this context, the problem is then referred as "prediction of individual sequences" - but a probabilistic model is also used sometimes [13].

In both settings, one is usually able to predict the time series as well as the best expert in the set of experts Θ, up to an error term that decreases with the number of observations n. This type of results is referred to as oracle inequalities in statistical theory. In other words, one builds on the basis of the observations a predictor $\hat{\theta}$ such that, with probability at least $1-\varepsilon$,

$$R(\hat{\theta}) \leq \inf_{\theta \in \Theta} R(\theta) + \Delta(n, \varepsilon) \tag{1}$$

where $R(\theta)$ is a measure of the risk of the predictor $\theta \in \Theta$. In general, the remainder term is of order $\Delta(n,\varepsilon) \approx \sqrt{d/n} + \log(\varepsilon^{-1})/\sqrt{n}$ in both approaches, where d is a measure of the complexity or dimension of Θ. We refer the reader to [18] for precise statements in the individual sequences case; for the batch case, the rate $\sqrt{d/n}$ is established in [7] for the absolute loss under a weak dependence assumption (up to a logarithmic term).

The method proposed in [7] is a two-step procedure: first, a set of randomized estimators is drawn, then, one of them is selected by the minimization of a penalized criterion. In this paper, we consider the one step Gibbs estimator introduced in [16] (the Gibbs procedure is related with online approaches like the weighted majority algorithm of [44, 64]). The advantage of this procedure is that it is potentially computationally more efficient when the number of submodels M is very large, this situation is thoroughly discussed in [3, 29] in the context of i.i.d. observations. We discuss the applicability of the procedure for various time series. Also, under additional assumptions on the model, we prove that the classical Empirical Risk Minimization (ERM) procedure can be used instead of the Gibbs estimator. On the contrary to the Gibbs estimator, there is no tuning parameter for the ERM, so this is a very favorable situation. We finally prove that, for a wide family of loss functions including the quadratic loss, the Gibbs estimator reaches the optimal rate $\Delta(n,\varepsilon) \approx d/n + \log(\varepsilon^{-1})/n$ under $\phi-$mixing assumptions. To our knowledge, this is the first time such a result is obtained in this setting. Note however that [1, 22] proves similar results in the online setting, and proves that it is possible to extend the results to the batch setting under $\phi-$mixing assumptions. However, their assumptions on the mixing coefficients is much stronger (our theorem only require summability while their result require exponential decay of the coefficients).

Our main results are based on PAC-Bayesian oracle inequalities. This type of results were first established for supervised classification [46, 60], but were later extended to other problems [3, 4, 16, 17, 28, 40, 57]. In PAC-Bayesian inequalities the complexity term $d = d(\Theta)$ 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 and the ERM estimators in Section 2.2. The main hypotheses necessary to prove theoretical results on these estimators are provided in Section 3. We give examples of inequalities of the form (1) for classical sets of predictors Θ in Section 4. 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 [38] in Section 5. As a special case, we generalize the results of [3, 29, 35] on sparse regression estimation to the case of autoregression. In Section 6 we provide an application to French GDP forecasting. A short simulation study is provided in Section 7. Finally, the proofs of all the theorems are given in Appendices 9 and 10.

2. Preliminaries

2.1. 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 takes values in $\mathbb{R}^p$. We denote $\|\cdot\|$ and $\|\cdot\|_1$ respectively the Euclidean and the ℓ^1 norms of $\mathbb{R}^p$, $p \geq 1$. We denote k an integer $k(n) \in \{1, \dots, n\}$ that might depend on n. We consider a family of predictors $\left\{f_\theta : (\mathbb{R}^p)^k \to \mathbb{R}^p, \theta \in \Theta\right\}$. 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:

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

Notice that no assumptions on k will be used on the paper, the choice of k is determined by the context. For example, if X is a Markov process, it makes sense to fix $k = 1$. In a completely agnostic setting, one might consider larger k. We assume that Θ is a subset of a vector space and that $\theta \mapsto f_\theta$ is linear. We consider a loss function $\ell : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$ that measures a distance between the forecast and the actual realization of the series. Assumptions on ℓ will be given in Section 3.

Definition 1.

For any $\theta \in \Theta$ we define the prediction risk as

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

The main assumption on the time series X is that the risk function $R(\theta)$ does effectively not depend on t. This is true for any strongly stationary time series. 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 AR predictors with j parameters as

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

for $\theta = (\theta_0, \theta_1, \dots, \theta_{j-1}) \in \Theta \subset \mathbb{R}^j$.

In order to deal with non-parametric settings, we will also use a model-selection type notation. By this, we mean that we will consider many possible models $\Theta_1, \dots, \Theta_M$, coming for example from different levels of approximation, and finally consider $\Theta = \cup_{j=1}^M \Theta_j$ (see e.g. Chapter 1 in [45]). Note that M can be any integer and might depend on n.

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 $(\mathbb{R}^p)^k \to \mathbb{R}^p$ (following e.g. [29], by dictionnary of function we actually mean any family of functions, for example the Fourier basis, a wavelet basis, splines, polynomials...). It is well-known that a small j will lead to poor approximations properties of Θ_j (large bias), while a large j leads to a huge variability in the estimation. In this situation, our main results allow the case $M = n$ and provide an estimator on $\Theta = \cup_{j=1}^n \Theta_j$ that will achieve the optimal balance between bias and variance.

2.2. The ERM and Gibbs estimators

Consider that $\Theta = \cup_{j=1}^{M} \Theta_j$. 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).
For any $1 \leq j \leq M$, the ERM in Θ_j is defined by

$$\hat{\theta}_j^{ERM} \in \arg\min_{\theta_j \in \Theta_j} r_n(\theta_j).$$

We will denote $\hat{\theta}^{ERM}$ the ERM defined on the entire parameter space Θ. It is well known that if Θ is a high dimensional space then the ERM suffers overfitting. In such cases, we reduce the dimension of the statistical problem by selecting a model through a penalization procedure. The resulting procedure is known under the name SRM (Structural Risk Minimization) [63], or penalized risk minimization [12].

Definition 4 (SRM).
Define the two steps estimator as $\hat{\theta}_{\hat{j}}^{ERM}$ where $\hat{j}$ minimizes the function of j

$$r_n(\hat{\theta}_j^{ERM}) + pen_j$$

for some penalties $pen_j > 0$, $1 \leq j \leq M$.

In some models, risk bounds on the ERM are not available. In order to deal with these models, we introduce another estimator: the Gibbs estimator. 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*. However, π should not necessarily be seen as a Bayesian prior: as in [17], the prior will be used to define a measure of the complexity of Θ (in the same way than the VC dimension of a set [63] measures its complexity).

Definition 5 (Gibbs estimator).
Define the Gibbs estimator with inverse temperature $\lambda > 0$ as

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

The choice of π and λ is discussed in Section 4. Consider the model-selection set-up $\Theta = \cup_{j=1}^{M} \Theta_j$ for disjoints Θ_j. In [7], the following penalization procedure was studied: first, calculate a Gibbs estimator $\hat{\theta}_{\lambda,j}$ in each Θ_j, then, choose one of them based on a penalized minimization criterion similar to the one in Definition 4. In this paper, even in the model selection setup, we will define a probability distribution on the whole space $\Theta = \cup_{j=1}^{M} \Theta_j$ and use Definition 5 to define the Gibbs estimator on Θ.

2.3. Oracle inequalities

Consider some parameter space Θ that is the union of M disjoint sets $\Theta = \cup_{j=1}^{M} \Theta_j$. Our results assert that the risk of the estimators are close to the best possible risk up to a remainder term with high probability $1-\varepsilon$. The rate at which the remainder term tends to zero with n is called the rate of convergence. We introduce the notation $\overline{\theta}_j$ and $\overline{\theta}$ with

$$R(\overline{\theta}_j) = \inf_{\theta \in \Theta_j} R(\theta) \text{ and } R(\overline{\theta}) = \inf_{\theta \in \Theta} R(\theta)$$

(we assume that these minimizers exists, they don't need to be unique; when they don't exist, we can replace these by approximate minimizers). We want to prove that the ERM or Gibbs estimators satisfy, for any $\varepsilon \in (0,1)$ and any $n \geq 0$, the so-called oracle inequality:

$$\mathbb{P}\left(R(\hat{\theta}) \leq \min_{1 \leq j \leq M} \left[R(\overline{\theta}_j) + \Delta_j(n,\varepsilon) \right] \right) \geq 1-\varepsilon \tag{2}$$

where the error terms $\Delta_j(n,\varepsilon) \to 0$ as $n \to \infty$ (we will also consider oracle inequalities when $M=1$, in this case, we will use the notation $\Delta(n,\varepsilon)$ instead of $\Delta_1(n,\varepsilon)$). Slow (resp. fast) rates of convergence correspond to $\Delta_j(n,\varepsilon) = O(n^{-1/2})$ (resp. $O(n^{-1})$) when $\varepsilon > 0$ is fixed for all $1 \leq j \leq M$. It is also important to estimate the increase of the error terms $\Delta_j(n,\varepsilon)$ when $\varepsilon \to 0$. Here it is proportional to $\log(\varepsilon^{-1})$; that corresponds to an exponential tail behavior of the risk.
To establish oracle inequalities, we require some assumptions discussed in the next section.

3. Main assumptions

We prove in Section 4 oracle inequalities under assumptions of three different types. First, assumptions **Bound**($\mathcal{B}$), **WeakDep**($\mathcal{C}$) and **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, these assumptions are satisfied for many classical time series as shown in [23, 26].
Second, assumptions **LipLoss**(K), **Lip**(L), **Dim**(d, D) and **L1**(Ψ) hold respectively on the loss function ℓ, the predictors $\hat{X}_t^{\theta}$ and the parameter spaces Θ_j. These assumptions can be checked in practice as the statistician know the loss function and the predictors.
Finally, the assumption **Margin**($\mathcal{K}$) involve both the observed time series and the loss function ℓ. As in the iid case, it is only required to prove oracle inequalities with fast rates.

3.1. Assumptions on the time series

Assumption Bound($\mathcal{B}$)**,** $\mathcal{B} > 0$: for any $t > 0$ we have $||X_t|| \leq \mathcal{B}$ almost surely.
It is possible to extend some of the results in this paper to unbounded time series using the truncation technique developed in [7]. The price to pay is an increased complexity in the bounds, so, for the sake of simplicity, we only deal with bounded series in this paper.
Assumption **WeakDep**($\mathcal{C}$) is about the $\theta_{\infty,n}(1)$-weak dependence coefficients of [23, 53].

Definition 6.
For any $k > 0$, define the $\theta_{\infty,k}(1)$-weak dependence coefficients of a bounded stationary sequence (X_t) by the relation

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

($||V||_\infty$ refers to the essential supremum of the random variable $||V||$ where V is a random vector in $\mathbb{R}^p$) where Λ_1^k is the set of 1-Lipshitz functions of k variables

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

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 [23] for details.

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

Example 3.

Examples of processes satisfying **WeakDep**$(\mathcal{C})$ *and* **Bound**$(\mathcal{B})$ *are provided in [7, 23, 32]. 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**$(\mathcal{C})$*, as well a large class of Markov models and non-linear* ARCH *models, see [32] p. 2003-2004.*

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 [38].

Definition 7.

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

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

where $\sigma(X_t, t \in I)$ *is the* σ*-algebra generated by the set of random variables* $\{X_t, t \in I\}$.

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 ([53]).

Let (X_t) *be any time series that satisfies* **Bound**$(\mathcal{B})$ *and* **PhiMix**$(\mathcal{C})$. *Then it also satisfies* **WeakDep** $(\mathcal{C}\mathcal{B})$.

(This is a direct consequence of the last inequality in the proof of Corollaire 1 p. 907 in [53]).

3.2. Assumptions on the loss function

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 4.

A classical example in statistics is given by $\ell(x, x') = \|x - x'\|$*, it is the loss used in [7], this loss is the absolute loss in the case of univariate time series. It satisfies* **LipLoss**(K) *with* $K = 1$. *In [47, 49], the loss function used is the quadratic loss* $\ell(x, x') = \|x - x'\|^2$. *When* **Bound**$(\mathcal{B})$ *is satisfied, the quadratic loss satisfies* **LipLoss**$(2\mathcal{B})$.

Example 5.

The class of quantile loss functions introduced in [41] 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 [9, 13, 42]. In this case, **LipLoss**(K) *is satisfied with $K = \max(\tau, 1-\tau) \le 1$.*

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

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

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

To define the Gibbs estimator we set a prior measure π on the parameter space Θ. The complexity of the parameter space is determined by the growth of the volume of sets around the oracle $\overline{\theta}_j$:

Assumption Dim(d_j, D_j**):** there are constants $d_j = d(\Theta_j, \pi_j)$ and $D_j = D(\Theta_j, \pi_j)$ satisfying

$$\forall \delta > 0, \quad \pi_j(\{\theta,\ R(\theta) - R(\overline{\theta}_j) < \delta\}) \le \left(\frac{\delta}{D_j}\right)^{d_j}.$$

This assumption basically states that the prior gives enough weight to the sets $\{\theta : R(\theta) - R(\overline{\theta}) < \delta\}$. As discussed in [7, 17], it holds for reasonable priors when Θ_j is a compact set in a finite dimensional space with d_j depending on the dimension and D_j depending on the diameter of Θ_j. In the case of the ERM, we need a more restrictive assumption that states that we can compare the set $\{\theta : R(\theta) - R(\overline{\theta}) < \delta\}$ to some ℓ^1 ball in Θ.

Assumption L1(Ψ)**,** $\psi > 0$: $\left\|\hat{X}_1^{\theta_1} - \hat{X}_1^{\theta_2}\right\| \le \psi\left\|\theta_1 - \theta_2\right\|_1$ a.s. for all $(\theta_1, \theta_2) \in \Theta^2$.

3.3. Margin assumption

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 or Bernstein hypothesis.

Assumption Margin($\mathcal{K}$)**,** $\mathcal{K} > 0$: for any $\theta \in \Theta$,

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

As assumptions **Margin**($\mathcal{K}$) and **PhiMix**($\mathcal{C}$) won't be used before Section 5, we postpone examples to this section.

4. Slow rates oracle inequalities

In this section, we give oracle inequalities in the sense of Equation 2 with slow rates of convergence $\Delta_j(n, \varepsilon)$. The proofs of these results are given in Section 10. Note that the results concerning the Gibbs estimator are actually corollaries of a general result, Theorem 7, stated in Section 9. We introduce the following notation for the sake of shortness.

Definition 8.

When **Bound**($\mathcal{B}$), **LipLoss**(K), **Lip**(L_j) *and* **WeakDep**($\mathcal{C}$) *are satisfied, we say that the model Θ satisfies Assumption* **SlowRates**(κ_j) *for $\kappa_j := K(1+L_j)(\mathcal{B}+\mathcal{C})/\sqrt{2}$.*

4.1. The experts selection problem with slow rates

Consider the so-called V-aggregation problem [51] with a finite set of predictors.

Theorem 1.
Assume that $|\Theta| = N \in \mathbb{N}$ and that **SlowRates(κ)** *is satisfied for $\kappa > 0$. Let π be the uniform probability distribution on Θ. Then the oracle inequality (2) is satisfied by the Gibbs estimator $\hat{\theta}_\lambda$ for $\lambda > 0$, $\varepsilon > 0$ with*

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

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

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

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

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

But $\kappa = \kappa(K, L, \mathcal{B}, \mathcal{C})$ and the constants $\mathcal{B}$ and $\mathcal{C}$ are, usually, unknown. However, under our assumptions, the ERM predictor reaches the same bound without any calibration parameter.

Theorem 2.
Assume that $|\Theta| = M$ and that **SlowRates(κ)** *is satisfied for $\kappa > 0$. Then the ERM estimator $\hat{\theta}^{ERM}$ satisfies the oracle inequality (2) for any $\varepsilon > 0$ with*

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

We now discuss the optimality of these results. First, note that to our knowledge, no lower bounds for estimation under dependence assumption are known in this context. But, as iid observations satisfy our weak dependence assumptions, we can compare our upper bounds to the lower bounds known in the iid case. In the context of a finite parameter set and bounded outputs, the optimal rate in the iid case still depends on the loss function. For the absolute loss, it is proved in [6], Theorem 8.3 p. 1618 that the rate $\sqrt{\log(N)/n}$ cannot be improved. This means that the rates in Theorems 1 and 2 cannot be improved without any additional assumption.

4.2. The Gibbs and ERM estimators when $M = 1$

In the previous subsection we focused on the case where Θ is a finite set. Here we deal with the general case, in the sense that Θ can be either finite or infinite. Note that we won't consider model selection issues in this subsection, say $M = 1$. The case where $\Theta = \cup_{i=1}^{M}\Theta_i$ with $M > 1$ is postponed to the next subsection.

Theorem 3.
Assume that **SlowRates(κ)** *and* **Dim(d, D)** *are satisfied. Then the oracle inequality (2) is satisfied for the Gibbs estimator $\hat{\theta}_\lambda$ for $\lambda > 0$ with*

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

Here again $\lambda = O(\sqrt{nd})$ yields slow rates of convergence $O(\sqrt{d/n}\log n)$. But an exact minimization of the bound with respect to λ is not possible as the constant κ is not known and cannot be estimated efficiently (estimations of the weak dependence coefficients are too conservative in practice). A similar oracle inequality holds for the ERM estimator, that does not require any calibration, but this time, this result requires a more restrictive assumption on the structure of Θ (see Remark 1 below).

Theorem 4.

Assume that $\Theta = \{\theta \in \mathbb{R}^d : \left\|\theta\right\|_1 \leq R\}$ for some $R > 0$, and that **SlowRates(κ)** *holds on the extended model $\Theta' = \{\theta \in \mathbb{R}^d : \left\|\theta\right\|_1 \leq R+1\}$. If* **L1($\Psi$)** *is satisfied then the oracle inequality (2) is satisfied for any $\varepsilon > 0$ with*

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

For n sufficiently large and $\lambda = ((1-k/n)/\kappa)\sqrt{dn} \geq 2K\psi/d$ we obtain the oracle inequality

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

Thus, the ERM procedure achieves predictions that are close to the oracle, with a slow rate of convergence. On the one hand, this rate of convergence can be improved under more restrictive assumption on the loss, the parameter spaces and the observations. On the other hand, the general result holds for any quantile losses and any parameter spaces in bijection with a ℓ^1 ball in $\mathbb{R}^d$. In particular it applies very easily to linear predictors of any orders.

Example 6.

When $\Theta = \{\theta \in \mathbb{R}^d : ||\theta||_1 \leq R\}$, the linear AR predictors with j parameters satisfy **Lip**(L) *with $L = R+1$. The assumptions of Theorem 4 are satisfied with $d = j$ and $\psi = \mathcal{B}$. Moreover, thanks to Remark 1 below, the assumptions of Theorem 3 are satisfied with $D = (K\mathcal{B} \vee K^2\mathcal{B}^2)(R+1)$.*

Note that the context of Theorem 4 are less general than the one of Theorem 3:

Remark 1.

Under the assumptions of Theorem 4 we have for any $\theta \in \Theta$

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

Consider the Gibbs estimator with prior distribution π uniform on $\Theta' = \{\theta \in \mathbb{R}^d : \theta_1 \leq R+1\}$. We have

$$\log\frac{1}{\pi\{\theta : R(\theta) - R(\overline{\theta}) < \delta\}} \leq \log\frac{1}{\pi\{\theta : ||\theta - \overline{\theta}||_1 < \frac{\delta}{K\psi}\}} \begin{cases} = d\log\left(\frac{K\psi(R+1)}{\delta}\right) & \textit{when } \delta/K\psi \leq 1 \\ \leq d\log\left(K\psi(R+1)\right) & \textit{otherwise.} \end{cases}$$

Thus, in any case,

$$\log\frac{1}{\pi\{\theta : R(\theta) - R(\overline{\theta}) < \delta\}} \leq d\log\left(\frac{(K\psi \vee K^2\psi^2)(R+1)}{\delta}\right)$$

and **Dim**(d, D) *is satisfied for $d = d$ and $D = (K\psi \vee K^2\psi^2)(R+1)$.*

Remark 2.

We obtain oracle inequalities for the ERM on parameter spaces in bijection with the ℓ_1-ball in $\mathbb{R}^d$. The rates of convergence are $O(\sqrt{d/n}\log n)$. In the iid case, procedures can achieve the rates $O(\sqrt{\log(d)/n})$ which is optimal as shown by [39] - but, to our knowledge, lower bounds in the dependent case are still an open issue. It nevertheless indicates that the ERM procedure might not be optimal in the setting considered here. Note however that, as the results on the Gibbs procedure are more general, they do not require the parameter space to be an ℓ_1-ball. For example, when Θ is finite, one can deduce a result similar to Theorem 1 from Theorem 3. This proves that Theorem 3 cannot be improved in general.

4.3. Model aggregation through the Gibbs estimator

We know tackle the case $\Theta = \cup_{i=1}^{M}\Theta_j$ with $M \geq 1$. In this case, it is convenient to define the prior π as $\pi = \sum_{1\leq i\leq M} p_i \pi_i$ where $\pi_i(\Theta_i) = 1$, $p_i \geq 0$ and $\sum_{1\leq i\leq M} p_i = 1$.

Theorem 5.

Assume that **SlowRates**(κ_j) **and Dim**(d_j, D_j) *are satisfied for all $1 \leq j \leq M$. For each $\theta \in \Theta$, as there is only one j such that $\theta \in \Theta_j$, we define $\kappa(\theta) := \kappa_j$. Define a modified Gibbs estimator as:*

$$\tilde{\theta}_\lambda = \int \theta \tilde{\rho}_\lambda(\mathrm{d}\theta) \text{ where } \tilde{\rho}_\lambda(\mathrm{d}\theta) \propto \exp\left[-\lambda r_n(\theta) - \frac{\lambda^2 \kappa(\theta)^2}{n(1-\frac{k}{n})^2}\right]\pi(\mathrm{d}\theta)$$

(note that when all the L_j are equal, this coincides with the Gibbs estimator $\hat{\theta}_\lambda$). Let us define the grid $\Lambda = \{2^0, 2^1, 2^2, \dots\} \cap [1, n]$ and

$$\tilde{\lambda} = \arg\min_{\lambda\in\Lambda}\left\{\int\left[r_n(\theta) + \frac{\lambda\kappa(\theta)^2}{n\,(1-k/n)^2}\right]\tilde{\rho}_\lambda(\mathrm{d}\theta) + \frac{\mathcal{K}(\tilde{\rho}_\lambda, \pi)}{\lambda}\right\}.$$

Then, with probability at least $1-\varepsilon$,

$$R(\tilde{\theta}_{\tilde{\lambda}}) \leq \min_{1\leq j\leq M}\left\{R(\overline{\theta}_j) + \inf_{\lambda\in[1,n]}\left[\frac{4\lambda\kappa_j^2}{n\,(1-k/n)^2} + 2\frac{d_j\log\left(\frac{2D_j\sqrt{e}\lambda}{d_j}\right) + \log\left(\frac{2\log_2(2n)}{\varepsilon p_j}\right)}{\lambda}\right]\right\}.$$

Note that when $M \leq n$, the choice $p_j = 1/M$ leads to a rate $O(\sqrt{d_j/n}\log(n))$. However, when the number of model is large, this is not a good choice. Calibration of p_j is discussed in details in [7], the choice $p_j \geq \exp(-d_j)$, when possible, has the advantage that it does not deteriorate the rate of convergence.

Note that it is possible to prove a similar result for a penalized ERM (or SRM) under additional assumptions: **L1**(Ψ_j) for each model Θ_j. However, as for the Gibbs estimator, the SRM requires the knowledge of κ_j, so there is no advantage at all in using the SRM instead of the Gibbs estimator in the model selection setup.

5. Fast rates oracle inequalities

5.1. Discussion on the assumptions

In this section, we provide oracle inequalities like (2) with fast rates of convergence $\Delta_j(n,\varepsilon) = O(d_j/n)$. One need additional restrictive assumptions.

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

As stated above, the margin (or Bernstein) assumption is required even in the iid setting to achieve fast rates. We provide now 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 process, see [5], 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 $X_t = F(X_{t-1}, \dots, X_{t-p}; \epsilon_t)$, the stationary solution is ϕ-mixing, the coefficients are estimated and can satisfy **PhiMix**($\mathcal{C}$). *See e.g. [50].*

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

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

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 **Lip**(L) *is satisfied. Moreover* **LipLoss**(K) *is satisfied with $K = 4\mathcal{B}$. Assume that $\overline{\theta} = \arg\min_{\theta \in \mathbb{R}^N} R(\theta) \in \Theta$ in order to have:*

$$\begin{aligned}
&\mathbb{E}\left\{\left[\left(X_{q+1} - f_\theta(X_q, ..., X_1)\right)^2 - \left(X_{q+1} - f_{\overline{\theta}}(X_q, ..., X_1)\right)^2\right]^2\right\} \\
&= \mathbb{E}\Big\{[f_\theta(X_q, ..., X_1) - f_{\overline{\theta}}(X_q, ..., X_1)]^2 \, [2X_{q+1} - f_\theta(X_q, ..., X_1) - f_{\overline{\theta}}(X_q, ..., X_1)]^2\Big\} \\
&\leq \mathbb{E}\left\{[f_\theta(X_q, ..., X_1) - f_{\overline{\theta}}(X_q, ..., X_1)]^2 \, 16\mathcal{B}^2(1 + L)^2\right\} \\
&\leq 16\mathcal{B}^2(1 + L)^2 \left[R(\theta) - R(\overline{\theta})\right] \text{ by Pythagorean theorem.}
\end{aligned}$$

Assumption **Margin**($\mathcal{K}$) *is satisfied with $\mathcal{K} = 16\mathcal{B}^2(1 + L)^2$. According to Theorem 6 below, the oracle inequality with fast rates holds as soon as Assumption* **PhiMix**($\mathcal{C}$) *is satisfied.*

We introduce the following notation for the sake of shortness.

Definition 9.

When **Margin**($\mathcal{K}$), **LipLoss**(K), **Bound**($\mathcal{B}$), **PhiMix**($\mathcal{C}$), **Lip**(L) *are satisfied, we will say for short that Θ satisfies Assumption* **FastRates**(κ) *for $\kappa := 4\mathcal{K}\mathcal{C}(4 \vee K L \mathcal{B})$.*

5.2. Gibbs estimator for model selection

We only give oracle inequalities for the Gibbs estimator in the model-selection setting. Obviously, when there is only one models, all the results can be obtained by taking $M = 1$ (in this case, this result can be extended to the ERM predictor at the cost of additional assumptions, so we won't present any results on the ERM here).

Theorem 6.
Assume that **FastRates**(κ) *and* **Dim**(d_j, D_j) *hold for any* $j \in \{1, ..., M\}$. *Then for* $\lambda = (n-k)/\kappa$, *we obtain the oracle inequality* (1) *is satisfied with*

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

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 6 is too conservative. In the iid case, it is shown in [29] 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.
The rate d/n is known to be optimal in the i.i.d. case for the quadratic loss, see e.g. (1.3) page 1676 in [14]. Let us compare the rates in Theorem 6 to the ones in [1, 22, 47, 49]. In [47, 49], the rate $1/n$ is never obtained. The paper [1] proves fast rates for online algorithms that are also computationally efficient, see also [22]. The fast rate $1/n$ is achieved when the coefficients (ϕ_r) are geometrically decreasing. In other cases, the rate is slower. Note that we do not suffer such a restriction: we only need the partial sums of the coefficients to converge. The Gibbs estimator of Theorem 6 can also be computed efficiently thanks to MCMC procedures, see [3, 29].

5.3. Corollary: sparse autoregression

Consider the linear predictors

$$\hat{X}_t^{\theta} = \theta_0 + \sum_{i=1}^{j} X_{t-i}\theta_i.$$

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

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

Let us remark that we have the disjoint union $\Theta = \cup_{J \subset \{1,\dots,p\}} \Theta_J = \{\theta \in \mathbb{R}^p : ||\theta||_1 \leq L\}$. 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 **PhiMix**($\mathcal{C}$) *is satisfied for some* $\mathcal{C} > 0$ *as well as* **Bound**($\mathcal{B}$). *Then the oracle inequality* (1) *is satisfied with*

$$\Delta(n, \varepsilon) = \text{cst.} \frac{|\overline{J}| \log\left((n-k)p/|\overline{J}|\right) + \log\left(\frac{2}{\varepsilon}\right)}{n-k}$$

for some constant $\text{cst} = \text{cst}(\mathcal{B}, \mathcal{C}, L)$ *and where* $\overline{J}$ *is the unique subset of* $\{1, \dots, p\}$ *such that* $\overline{\theta} \in \Theta_{\overline{J}}$.

This extends the results of [3, 29, 35] to the case of autoregression. The upper bound is optimal up to the n in the log term, see e.g. (1.3) page 1676 in [14].

Proof. The proof follows the computations of Example 10 that we do not reproduce here: we check the conditions **LipLoss**(K) with $K = 4\mathcal{B}$, **Lip**(L) and **Margin**($\mathcal{K}$) with $\mathcal{K} = 16\mathcal{B}^2(1+L)^2$. We can apply Theorem 6 with $d_J = |J|$ and $D_j = L$. □

6. Application to French GDP forecasting

6.1. Uncertainty in GDP forecasting

Every quarter $t \geq 1$, the French national bureau of statistics, INSEE (*Institut National de la Statistique et des Etudes Economiques* http://www.insee.fr/), 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 ΔGDP_j, $0 < j < t$ (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 [48]);

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 [24] 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 [19, 25]. Following [20], 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 available from the INSEE website. Note that a similar approach is used in other countries, see e.g. [8] 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 [30, 62] for surveys on density forecasting and [11] 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 [20, 27]. 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.

6.2. Application of Theorem 4 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 5 page 71) 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 [20] given by the relation

$$f_\theta(X_{t-1}, X_{t-2}) = \theta_0 + \theta_1 \Delta\text{GDP}_{t-1} + \theta_2 I_{t-1} + \theta_3 (I_{t-1} - I_{t-2})|I_{t-1} - I_{t-2}| \quad (3)$$

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 $R^{\tau}(\theta) := \mathbb{E}[\ell_{\tau}(\Delta\mathrm{GDP}_t, f_{\theta}(X_{t-1}, X_{t-2}))]$ 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 4 as **Lip**(L) is satisfied Θ' with $L = D + 1$ and **LipLoss**(K) with $K = 1$. If the observations are bounded, stationary such that **WeakDep**($\mathcal{C}$) holds for some $\mathcal{C} > 0$, the assumptions of Theorem 4 are satisfied with $\psi = \mathcal{B}$ and $d = 4$:

Corollary 2.

Let $\tau \in (0, 1)$. 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.

6.3. 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).

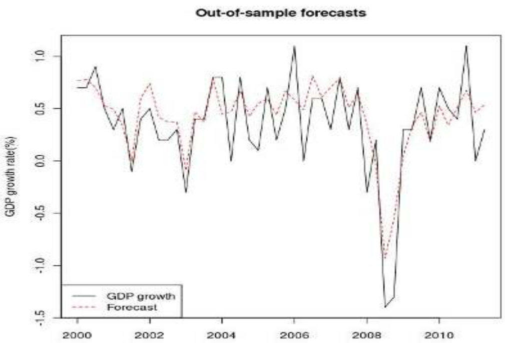

Fig 1. French GDP forecasting using the quantile loss function with $\tau = 0.5$.

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.

The methodology fails to forecast the importance of the 2008 subprime crisis as it was the case for the INSEE forecaster, see [20]. 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 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.

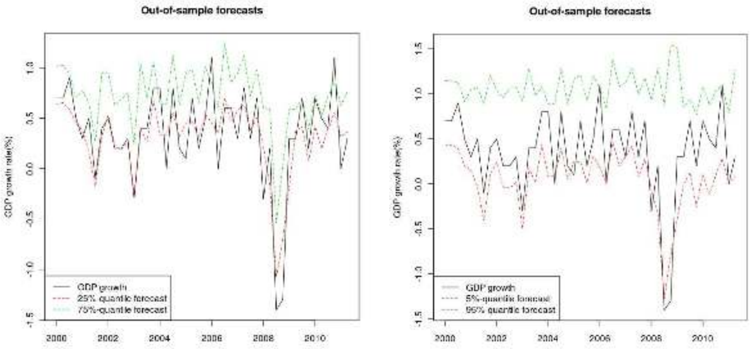

Fig 2. French GDP online 50%-confidence intervals (left) and 90%-confidence intervals (right).

Table 1. *Performances of the ERM and of the INSEE.*

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

Table 2. *Empirical frequencies of the event: GDP falls under the predicted τ-quantile.*

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

7. 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 [52]. We want to check that the ERM and Gibbs estimators can be safely 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.

7.1. Parametric family of predictors

Here, we compare the ERM to the QMLE.
We draw simulations from an AR(1) models (4) and a non linear model (5):

$$X_t = 0.5X_{t-1} + \varepsilon_t \tag{4}$$

$$X_t = 0.5\sin(X_{t-1}) + \varepsilon_t \tag{5}$$

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 6: there

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.*

n	Model	Innovations	ERM abs.	ERM quad.	QMLE
100	(4)	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)
	(5)	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	(4)	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)
	(5)	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)

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 [7] 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 $(X_1, \dots, X_{n-1})$. 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 (4) while the ERM with quadratic loss performs slightly better on model (5). 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.

7.2. 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 [2]. The Gibbs estimator is computed using a Reversible Jump MCMC algorithm described in Section 4 in the preprint version of [3] (available on arXiv as *http://arxiv.org/pdf/1009.2707v1.pdf*). 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:

$$X_t = 0.5X_{t-1} + 0.1X_{t-2} + \varepsilon_t \tag{6}$$

$$X_t = 0.6X_{t-4} + 0.1X_{t-8} + \varepsilon_t \tag{7}$$

$$X_t = \cos(X_{t-1})\sin(X_{t-2}) + \varepsilon_t \tag{8}$$

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.

The three procedures seem not significatively different. Although notice that the Gibbs predictor performs better on Models (7) and (8) while the AIC predictor performs slightly better on Model (6). Note that the Gibbs predictor performs also well in the case of a Gaussian noise where the boundedness assumption is not satisfied.

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

n	Model	Innovations	Gibbs	AIC	Full Model
100	(6)	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)
	(7)	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)	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	(6)	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)
	(7)	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)	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. Conclusion

This paper provides oracle inequalities for the empirical risk minimizer and the Gibbs estimator that generalizes earlier results by Catoni [17] to the context of time series forecasting. While essentially theoretical, these results are used in a real-life example with promising results. Future work might include a more intensive simulation study. Probably, more efficient Monte-Carlo algorithms should be investigated. Just before we submitted the final version of this paper, a preprint appeared on arXiv [58] where the computation time needed to compute an accurate approximation of the estimator by Monte-Carlo is upper-bounded. This is a very promising research direction. Equally important, on the theoretical side, while the assumptions needed to obtain the slow rates of convergence are rather general, the assumptions we used to get the fast rates are restrictive. Further work will include an investigation on the optimality of these assumptions.

Acknowledgements

We would like to thank the anonymous referees for their constructive remarks. We also would like to thank Pr. Olivier Catoni, Pascal Massart and Alexandre Tsybakov for insightfull comments on preliminary versions of this work.

References

[1] A. Agarwal and J. C. Duchi, *The generalization ability of online algorithms for dependent data*, IEEE Trans. Inform. Theory **59** (2011), no. 1, 573–587.

[2] 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.

[3] P. Alquier and P. Lounici, *PAC-Bayesian bounds for sparse regression estimation with exponential weights*, Electron. J. Stat. **5** (2011), 127–145.

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

[5] K. B. Athreya and S. G. Pantula, *Mixing properties of Harris chains and autoregressive processes*, J. Appl. Probab. **23** (1986), no. 4, 880–892. MR 867185 (88c:60127)

[6] J.-Y. Audibert, *Fast rates in statistical inference through aggregation*, Ann. Statist. **35** (2007), no. 2, 1591–1646.

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

[8] 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.

[9] A. Belloni and V. Chernozhukov, *L1-penalized quantile regression in high-dimensional sparse models*, Ann. Statist. **39** (2011), no. 1, 82–130.

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

[11] 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.

[12] L. Birgé and P. Massart, *Gaussian model selection*, J. Eur. Math. Soc. **3** (2001), no. 3, 203–268.

[13] G. Biau and B. Patra, *Sequential quantile prediction of time series*, IEEE Trans. Inform. Theory **57** (2011), 1664–1674.

[14] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, *Aggregation for gaussian regression*, Ann. Statist. **35** (2007), no. 4, 1674–1697.

[15] O. Catoni, *A PAC-Bayesian approach to adaptative classification*, preprint (2003).

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

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

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

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

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

[21] N. V. Cuong, L. S. Tung Ho, and V. Dinh, *Generalization and robustness of batched weighted average algorithm with v-geometrically ergodic markov data*, Proceedings of ALT'13 (Jain S., R. Munos, F. Stephan, and T. Zeugmann, eds.), Springer, 2013, pp. 264–278.

[22] J. C. Duchi, A. Agarwal, M. Johansson, and M. I. Jordan, *Ergodic mirror descent*, SIAM J. Optim. **22** (2012), no. 4, 1549–1578.

[23] 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.

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

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

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

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

[28] A. Dalalyan and J. Salmon, *Sharp oracle inequalities for aggregation of affine estimators*, Ann. Statist. **40** (2012), no. 4, 2327–2355.

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

[30] 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.

[31] M. D. Donsker and S. S. Varadhan, *Asymptotic evaluation of certain markov process expectations for large time. iii.*, Comm. Pure Appl. Math. **28** (1976), 389–461.

[32] P. Doukhan and O. Wintenberger, *Weakly dependent chain with infinite memory*, Stochastic Process. Appl. **118** (2008), no. 11, 1997–2013.

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

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

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

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

[37] H. Hang and I. Steinwart, *Fast learning from α-mixing observations*, Technical report, Fakultät für Mathematik und Physik, Universität Stuttgart, 2012.

[38] I. A. Ibragimov, *Some limit theorems for stationary processes*, Theory Probab. Appl. **7** (1962), no. 4, 349–382.

[39] A. B. Juditsky, A. V. Nazin, A. B. Tsybakov, and N. Vayatis, *Recursive aggregation of estimators bythe mirror descent algorithm with averaging*, Probl. Inf. Transm. **41** (2005), no. 4, 368–384.

[40] A. B. Juditsky, P. Rigollet, and A. B. Tsybakov, *Learning my mirror averaging*, Ann. Statist. **36** (2008), no. 5, 2183–2206.

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

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

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

[44] N. Littlestone and M.K. Warmuth, *The weighted majority algorithm*, Information and Computation **108** (1994), 212–261.

[45] P. Massart, *Concentration inequalities and model selection - ecole d'été de probabilités de saint-flour xxxiii - 2003*, Lecture Notes in Mathematics - J. Picard Editor, vol. 1896, Springer, 2007.

[46] 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.

[47] R. Meir, *Nonparametric time series prediction through adaptive model selection*, Mach. Learn. **39** (2000), 5–34.

[48] C. Minodier, *Avantages comparés des séries premières valeurs publiées et des séries des valeurs révisées*, Documents de Travail de la DESE, 2010.

[49] D. S. Modha and E. Masry, *Memory-universal prediction of stationary random processes*, IEEE Trans. Inform. Theory **44** (1998), no. 1, 117–133.

[50] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Communications and Control Engineering Series, Springer-Verlag London Ltd., London, 1993. MR 1287609 (95j:60103)

[51] A. Nemirovski, *Topics in nonparametric statistics*, Lectures on Probability Theory and Statistics - Ecole d'ét'e de probagilités de Saint-Flour XXVIII (P. Bernard, ed.), Springer, 2000, pp. 85–277.

[52] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, 2008.

[53] E. Rio, *Ingalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes*, C. R. Math. Acad. Sci. Paris **330** (2000), 905–908.

[54] P.-M. Samson, *Concentration of measure inequalities for markov chains and ϕ-mixing processes*, Ann. Probab. **28** (2000), no. 1, 416–461.

[55] I. Steinwart and A. Christmann, *Fast learning from non-i.i.d. observations*, Advances in Neural Information Processing Systems 22 (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, eds.), 2009, pp. 1768–1776.

[56] I. Steinwart, D. Hush, and C. Scovel, *Learning from dependent observations*, J. Multivariate Anal. **100** (2009), 175–194.

[57] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, J. Peters, and P. Auer, *Pac-bayesian inequalities for martingales*, IEEE Trans. Inform. Theory **58** (2012), no. 12, 7086–7093.

[58] A. Sanchez-Perez, *Time series prediction via aggregation : an oracle bound including numerical cost*, Preprint arXiv:1311.4500, 2013.

[59] 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.

[60] 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.

[61] N. N. Taleb, *Black swans and the domains of statistics*, Amer. Statist. **61** (2007), no. 3, 198–200.

[62] A. S. Tay and K. F. Wallis, *Density forecasting: a survey*, J. Forecast **19** (2000), 235–254.

[63] V. Vapnik, *The nature of statistical learning theory*, Springer, 1999.

[64] V.G. Vovk, *Aggregating strategies*, Proceedings of the 3rd Annual Workshop on Computational Learning Theory (COLT), 1990, pp. 372–283.

[65] O. Wintenberger, *Deviation inequalities for sums of weakly dependent time series*, Electron. Commun. Probab. **15** (2010), 489–503.

[66] Y.-L. Xu and D.-R. Chen, *Learning rate of regularized regression for exponentially strongly mixing sequence*, J. Statist. Plann. Inference **138** (2008), 2180–2189.

[67] B. Zou, L. Li, and Z. Xu, *The generalization performance of erm algorithm with strongly mixing observations*, Mach. Learn. **75** (2009), 275–295.

9. A general PAC-Bayesian inequality

Theorems 1 and 3 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 [17].

Theorem 7 (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 10, but we can now provide the proofs of Theorems 1 and 3.

Proof of Theorem 1. We apply Theorem 7 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:

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

■

Proof of Theorem 3. An application of Theorem 7 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:

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

Under the assumptions of Theorem 3 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:

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

■

10. Proofs

10.1. Preliminaries

We will use Rio's inequality [53] 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 $(X_1, \dots, X_n)$ come from a stationary process (X_t)

Lemma 1 (Rio [53]).

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

$$|h(x_1, \dots, x_n) - h(y_1, \dots, y_n)| \leq \sum_{i=1}^{n} ||x_i - y_i||. \tag{9}$$

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

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

Others exponential inequalities can be used to obtain PAC-Bounds in the context of time series: the inequalities in [26, 54] for mixing time series, and [23, 65] under weakest "weak dependence" assumptions, [57] 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 [54]).

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 [54] which states a more general result on empirical processes. In page 457 of [54], 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 [54]). 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 [31] variational formula).

For any $\pi \in \mathcal{M}_+^1(E)$, for any measurable upper-bounded function $h : E \to \mathbb{R}$ we have:

$$\int \exp(h) d\pi = \exp\left(\sup_{\rho\in\mathcal{M}_+^1(E)} \left(\int h d\rho - \mathcal{K}(\rho, \pi)\right)\right). \tag{10}$$

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 [43]). For a complete proof of this variational formula, even in the non integrable cases, we refer the reader to [15, 17, 31].

10.2. Technical lemmas for the proofs of Theorems 2, 4, 5 and 7

Lemma 4.

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

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

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 (9), remember that $\ell(x,x') = g(x-x')$ so

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

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

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

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 $t = K(1+L)\lambda/(n-k)$:

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

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*

$$\mathbb{P}\left\{\begin{array}{l}\forall\rho\in\mathcal{M}^1_+(\Theta), \\ \int R\mathrm{d}\rho \leq \int r_n\mathrm{d}\rho + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda} \\ \textit{and} \\ \int r_n\mathrm{d}\rho \leq \int R\mathrm{d}\rho + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}\end{array}\right\} \geq 1-\varepsilon. \tag{11}$$

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

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

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

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

We apply Lemma 3 and we get:

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

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

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

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

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

we obtain:

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

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*

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

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

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

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

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

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

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

Here again, a union bound ends the proof. ■

10.3. Proof of Theorems 7 and 5

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

Proof of Theorem 7. We apply Lemma 5. So, with probability at least $1-\varepsilon$ we are on the event given by (11). From now, we work on that event. The first inequality of (11), when applied to $\hat{\rho}_\lambda(d\theta)$, gives

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

According to Lemma 3 we have:

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

so we obtain

$$\int R(\theta)\hat{\rho}_\lambda(d\theta) \le \inf_\rho\left\{\int r_n(\theta)\rho(d\theta) + \frac{\lambda\kappa^2}{n(1-k/n)^2} + \frac{\mathcal{K}(\rho,\pi)+\log(2/\varepsilon)}{\lambda}\right\}. \tag{12}$$

We now estimate from above $r(\theta)$ by $R(\theta)$. Applying the second inequality of (11) and plugging it into Inequality 12 gives

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

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

Proof of Theorem 5. Follow, for each $\lambda \in \Lambda$, the proof of Lemma 5 with $\kappa(\theta)$ instead of κ, and fixed confidence level $\varepsilon/|\Lambda| > 0$. We obtain, for all $\lambda \in \Lambda$,

$$\mathbb{P}\left\{\begin{array}{l}\forall\rho\in\mathcal{M}_+^1(\Theta),\\ \int R d\rho \le \int\left[r_n(\theta)+\frac{\lambda\kappa(\theta)^2}{n(1-k/n)^2}\right]\rho(d\theta) + \frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{2|\Lambda|}{\varepsilon}\right)}{\lambda}\\ \text{and}\\ \int r_n d\rho \le \int\left[R(\theta)+\frac{\lambda\kappa(\theta)^2}{n(1-k/n)^2}\right]\rho(d\theta) + \frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{2|\Lambda|}{\varepsilon}\right)}{\lambda}\end{array}\right\} \ge 1-\varepsilon/|\Lambda|.$$

A union bound provides:

$$\mathbb{P}\left\{\begin{array}{l}\forall\lambda\in\Lambda,\forall\rho\in\mathcal{M}_+^1(\Theta),\\ \int R d\rho \le \int\left[r_n(\theta)+\frac{\lambda\kappa(\theta)^2}{n(1-k/n)^2}\right]\rho(d\theta) + \frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{2|\Lambda|}{\varepsilon}\right)}{\lambda}\\ \text{and}\\ \int r_n d\rho \le \int\left[R(\theta)+\frac{\lambda\kappa(\theta)^2}{n(1-k/n)^2}\right]\rho(d\theta) + \frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{2|\Lambda|}{\varepsilon}\right)}{\lambda}\end{array}\right\} \ge 1-\varepsilon. \tag{13}$$

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

$$\begin{aligned} R(\tilde{\theta}) = R(\tilde{\theta}_{\tilde{\lambda}}) &\le \int R(\theta)\tilde{\rho}_{\tilde{\lambda}}(d\theta) \text{ by Jensen's inequality}\\ &\le \int\left[r_n(\theta)+\frac{\tilde{\lambda}\kappa(\theta)^2}{n(1-k/n)^2}\right]\tilde{\rho}_{\tilde{\lambda}}(d\theta) + \frac{\mathcal{K}(\tilde{\rho}_{\tilde{\lambda}},\pi)+\log\left(\frac{2|\Lambda|}{\varepsilon}\right)}{\tilde{\lambda}}\\ &\qquad \text{by (13)}\\ &= \min_{\lambda\in\Lambda}\left\{\int\left[r_n(\theta)+\frac{\lambda\kappa(\theta)^2}{n(1-k/n)^2}\right]\tilde{\rho}_\lambda(d\theta) + \frac{\mathcal{K}(\tilde{\rho}_\lambda,\pi)+\log\left(\frac{2|\Lambda|}{\varepsilon}\right)}{\lambda}\right\}\end{aligned}$$

$$
\begin{aligned}
&\text{by definition of } \hat{\lambda} \\
&= \min_{\lambda\in\Lambda} \inf_{\rho\in\mathcal{M}^1_+(\Theta)} \left\{ \int \left[r_n(\theta) + \frac{\lambda\kappa(\theta)^2}{n\left(1-k/n\right)^2} \right] \rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{2|\Lambda|}{\varepsilon}\right)}{\lambda} \right\} \\
&\text{by Lemma 3} \\
&\leq \min_{\lambda\in\Lambda} \inf_{\rho\in\mathcal{M}^1_+(\Theta)} \left\{ \int \left[R(\theta) + \frac{2\lambda\kappa(\theta)^2}{n\left(1-k/n\right)^2} \right] \rho(\mathrm{d}\theta) + 2\frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{2|\Lambda|}{\varepsilon}\right)}{\lambda} \right\} \\
&\text{by (13) again} \\
&\leq \min_{\lambda\in\Lambda} \min_{1\leq j\leq M} \inf_{\delta>0} \left\{ R(\overline{\theta}_j) + \delta + \frac{2\lambda\kappa_j^2}{n\left(1-k/n\right)^2} + 2\frac{d_j \log\left(D_j/\delta\right) + \log\left(\frac{2|\Lambda|}{\varepsilon p_j}\right)}{\lambda} \right\} \\
&\text{by restricting } \rho \text{ as in the proof of Cor. 3 page 72} \\
&\leq \min_{1\leq j\leq M} \min_{\lambda\in\Lambda} \left\{ R(\overline{\theta}_j) + \frac{2\lambda\kappa_j^2}{n\left(1-k/n\right)^2} + 2\frac{d_j \log\left(\frac{D_j\sqrt{e}\lambda_j}{d_j}\right) + \log\left(\frac{2|\Lambda|}{\varepsilon p_j}\right)}{\lambda} \right\} \\
&\text{by taking } \delta = \frac{d_j}{\lambda_j} \\
&= \min_{1\leq j\leq M} \left\{ R(\overline{\theta}_j) + \inf_{\lambda\in[1,n]} \left\{ \frac{4\lambda\kappa_j^2}{n\left(1-k/n\right)^2} + 2\frac{d_j \log\left(\frac{2D_j\sqrt{e}\lambda}{d_j}\right) + \log\left(\frac{2|\Lambda|}{\varepsilon p_j}\right)}{\lambda} \right\} \right\}
\end{aligned}
$$

as, for any $\lambda \in [1, n]$, there is $\lambda' \in \Lambda$ such that $\lambda' \leq \lambda \leq 2\lambda'$. Finally, note that $|\Lambda| \leq \log_2(n) + 1 = \log_2(2n)$. ■

10.4. Proof of Theorems 2 and 4

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 $1-\varepsilon$,

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

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

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

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

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

So for any ρ,

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

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

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

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

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

But for any $\delta \leq 1$,

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

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 $\lambda \geq 2K\psi/d$, we get:

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

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

10.5. Some preliminary lemmas for the proof of Theorem 6

Lemma 7.
Under the hypothesis of Theorem 6, we have, for any $\theta \in \Theta$, for any $0 \leq \lambda \leq (n-k)/(2kKL\mathcal{B}\mathcal{C})$,

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

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

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

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 6, 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. ■

10.6. Proof of Theorem 6

Fix $0\leq\lambda=(n-k)/(4kKL\mathcal{B}\mathcal{C})\wedge(n-k)/(16k\mathcal{C})\leq(n-k)/(2kKL\mathcal{B}\mathcal{C})$. 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:

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

by restricting ρ as in the proof of Theorem 3. First, notice that our choice $\lambda \leq (n-k)/(16k\mathcal{C})$ leads to

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

$$\leq 4 \inf_j \inf_{\delta>0} \left\{ R(\overline{\theta}_j) + \delta - R(\overline{\theta}) + \frac{d_j \log\left(\frac{D_j}{\delta}\right) + \log\left(\frac{2}{\varepsilon p_j}\right)}{\lambda} \right\}.$$

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:

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

Jensen's inequality leads to:

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

■