JMLR: Workshop and Conference Proceedings vol (2012) 1–15

Fast rates in learning with dependent observations

Pierre Alquier ALQUIER@MATH.UNIV-PARIS-DIDEROT.FR
CREST-LS & LPMA, Université Paris 7, 175, rue du Chevaleret, 75013 Paris, FRANCE

Olivier Wintenberger WINTENBERGER@CEREMADE.DAUPHINE.FR
CREST-LFA & CEREMADE, Université Paris Dauphine, Place du Marchal De Lattre De Tassigny, 75775 PARIS CEDEX 16, FRANCE

Abstract

In this paper we tackle the problem of fast rates in time series forecasting from a statistical learning perspective. In a serie of papers (e.g. Meir (2000); Modha and Masry (1998); Alquier and Wintenberger (2012)) it is shown that the main tools used in learning theory with iid observations can be extended to the prediction of time series. The main message of these papers is that, given a family of predictors, we are able to build a new predictor that predicts the series as well as the best predictor in the family, up to a remainder of order $1/\sqrt{n}$. It is known that this rate cannot be improved in general. In this paper, we show that in the particular case of the least square loss, and under a strong assumption on the time series (ϕ-mixing) the remainder is actually of order $1/n$. Thus, the optimal rate for iid variables, see e.g. Tsybakov (2003), and individual sequences, see Cesa-Bianchi and Lugosi (2006) is, for the first time, achieved for uniformly mixing processes. We also show that our method is optimal for aggregating sparse linear combinations of predictors.

Keywords: Statistical learning theory, time series prediction, PAC-Bayesian bounds, oracle inequalities, fast rates, sparsity, mixing.

1. Intro

The problem of time series forecasting is a standard problem in statistics. The parametric approach contains a wide range of models associated with efficient estimation and prediction methods, see e.g. Hamilton (1994); Brockwell and Davis (2009).

In the last few years, several universal approaches emerged from various fields such that non-parametric statistics, machine learning, computer science and game theory. These approaches share some common features: the aim is to to build a prediction procedure that is able to predict the series as well as the best predictor in a given set of initial predictors, say Θ. The set of predictors are usually inspired by different parametric or non-parametric statistical models. The true distribution of the data is not assumed to belong to one of these models. However, we can distinguish two classes in these approaches, with different quantification of the objective, and different terminologies:

- in the "prediction of individual sequences" approach, predictors are usually called "experts". The objective is online prediction: at each date t, a prediction of the future realization x_{t+1} is based on the previous observations x_1, ..., x_t, the objective being to minimize the cumulative prevision loss. See for example Cesa-Bianchi and Lugosi (2006); Stoltz (2010) for an introduction.

- in the statistical learning approach, the given predictors are sometimes referred as "models" or "concepts". The batch setting is more classical in statistics. A prediction procedure is build on a complete sample X_1, ..., X_n. The performance of the procedure is compared on average with the best predictor, called the "oracle". The environment is not deterministic and some hypotheses like mixing or weak dependence is required: see Meir (2000); Modha and Masry (1998); Alquier and Wintenberger (2012).

In both settings, we are able to predict a bounded time series as well as the best expert, up to a small remainder. This type of results is referred in statistical theory as an oracle inequality. In general, neglecting the size of the set of predictors Θ, the remainder is of the order $1/\sqrt{n}$ in both approaches: see, e.g., Cesa-Bianchi and Lugosi (2006) for the "individual sequences" approach; for the "statistical learning approach" the rate $1/\sqrt{n}$ is reached in Alquier and Wintenberger (2012). This paper is based on the following remark: in the case of prediction of individual sequences, under stronger assumption on the loss function (satisfied e.g. by the quadratic loss), a fast rate $1/n$ can be reached. Note that Meir (2000); Modha and Masry (1998) deal with the quadratic loss, their rate can be better than $1/\sqrt{n}$ but cannot reach $1/n$. Here, we prove that the same result is true in the statistical learning setting. Namely, under a ϕ-mixing assumption introduced in Ibragimov (1962), we are able to reach the fast rate in the batch setting for the quadratic loss.

Following Alquier and Wintenberger (2012), we will use tools from the PAC-Bayesian theory to build our prediction procedure. Historically, the PAC-Bayesian point of view emerged in statistical learning to deal with supervised classification (using the 0/1-loss), see the seminal papers Shawe-Taylor and Williamson (1997); McAllester (1999). These results were extended to general loss functions and more accurate bounds were then given, see for example Catoni (2004, 2007); Alquier (2008); Dalalyan and Tsybakov (2008); Audibert (2010); Alquier and Lounici (2011); Seldin et al. (2011); Gerchinovitz (2011). Interestingly enough, PAC-Bayesian methods often lead to a prediction procedure that is an aggregation of the various predictors in Θ with exponential weights, a standard procedure in individual sequences prediction (introduced by Vovk (1990); Littlestone and Warmuth (1994)). It is striking to note that this procedures receives theoretical justification from approaches that have so different philosophies and objectives. This procedures received various names: EWA, for Exponentially Weighted Aggregate, in Dalalyan and Tsybakov (2008); Gerchinovitz (2011), Gibbs estimator in Catoni (2004, 2007); Alquier (2008); Audibert (2010), weighted majority algorithm in Littlestone and Warmuth (1994)... In Audibert (2004), it is also proved that this estimator is simply the Bayesian estimator under suitable model and prior.

In Section 2 we introduce the notations used in the whole paper, in particular the time series $(X_t)_{t\in\mathbb{Z}}$ and the set of predictors Θ. Section 3 is devoted to the description of the Gibbs estimator. Our main result is Theorem 1, it is stated in Section 4. In Section 5 we provide examples of time series satisfying the main assumption of Theorem 1 (ϕ-mixing). In Section 6 we discuss the implementation of our procedure using MCMC methods and show the results of some simulations. Finally, proofs are given Section 8, with some technical results postponed to the appendix. As we will see, the main tool needed to apply PAC-Bayesian techniques is a control of the Laplace transform of the prevision risk. In the iid setting, this might be done using classical Hoeffding's or Bernstein's Inequalities. In the context of ϕ-mixing, such a result is provided by a powerful result in Samson (2000).

Note that in this paper, we focus on the case where the set of predictors is the linear span of a finite family of basic predictors. Theorem 1 will be of particular interest in the case where a sparse combination of those basic predictors provide a good prediction. But the results in these paper can be extended in other contexts (e.g. if we only want to predict as well as the best basic predictor). The proof of Theorem 1 involves a general result, Lemma 2, that can be adapted to these various context.

2. The context

2.1. The observation

We assume that we observe $(X_1, \ldots, X_n)$ where $(X_t)_{t\in\mathbb{Z}}$ is a real, stationary process, bounded by a constant B. We remind the ϕ-mixing coefficients of the process (X_t) as introduced by Ibragimov (1962):

Definition 1 (ϕ-mixing coefficients) *We define the ϕ-mixing coefficients of the process $(X_t)_{t\in\mathbb{Z}}$ by*

$$\phi_r = \sup_{(A,B)\in \mathfrak{S}_0 \times \mathfrak{F}_r} |\pi(B/A) - \pi(B)|$$

where $\mathfrak{S}_0 = \sigma(X_t, t \leq 0)$ and $\mathfrak{F}_r = \sigma(X_t, t \geq r)$. We also define:

$$K_\phi^{(n)}(q) := 1 + \sum_{r=1}^{n-q} \sqrt{\phi_{\lfloor r/q \rfloor}}.$$

2.2. Set of predictors

We set a value q and a family of functions: g_1, ..., $g_p : [-B, B]^q \to [-B, B]$. The set of predictors, for a given $b > 0$, is defined by:

$$\{f_\theta, \theta \in \Theta(b)\}$$

where $\Theta(b) = \{\theta \in \mathbb{R}^p : \|\theta\|_1 < b\}$, and

$$f_\theta = \sum_{j=1}^{p} \theta_j g_j.$$

We also put $\Theta = \mathbb{R}^p$ and our objective is to find a θ such that X_{q+1} is well predicted by $f_\theta(X_q, ..., X_1)$ on average under the stationary distribution.

Note that we will allow very large set of predictors (experts, ...). Actually, we will allow $n \ll p$. In this case, a sparsity assumption will be necessary: namely, it is possible to build a good predictor θ such that most of its coordinates are close to 0. This is now a classical assumption in statistical learning theory, see e.g. Tibshirani (1996); Bühlmann and van de Geer (2011).

Example 1 (Auto-regressive predictors) *A very classical example is to design predictors based on auto-regressive models (AR). We put* $p = q$ *and* $g_1(x_q, ..., x_1) = x_q$, *...,* $g_q(x_q, ..., x_1) = x_1$ *so we obtain AR predictors*

$$f_\theta(X_q, ..., X_1) = \sum_{j=1}^{q} \theta_j X_{p-j}.$$

Note that in this case, $p < n$.

Example 2 *We can extend the previous setting to non-linear AR predictors. For example, We take* $p = 2^q$ *and* $g_1(x_q, ..., x_1) = 1(x_q > 0, ..., x_1 > 0)$, *then* $g_2(x_q, ..., x_1) = 1(x_q > 0, ..., x_2 > 0, x_1 \leq 0)$, *..., up to* $g_{2^q}(x_q, ..., x_1) = 1(x_q \leq 0, ..., x_1 \leq 0)$.

Definition 2 (Prevision and empirical risks) *We define the prevision risk*

$$R(\theta) = \mathbb{E}_{\mathbb{P}} \left\{ \left[X_{q+1} - f_\theta(X_q, ..., X_1)\right]^2 \right\}$$

and the empirical risk

$$r(\theta) = \frac{1}{n-q} \sum_{i=q+1}^{n} \left[X_i - f_\theta(X_{i-1}, ..., X_{i-q})\right]^2$$

and

$$\overline{\theta} \in \arg\min_{\Theta} R.$$

The objective is to build an estimator $\hat{\theta}$ based on the observations $(X_1, \ldots, X_n)$ such that $R(\hat{\theta})$ is as small as possible. We see in the next sections that the Gibbs estimator reaches this objective.

3. Description of the method

Ths Gibbs estimator as defined in Catoni (2007) requires a prior distribution on the parameter space.

Definition 3 (The prior) *For* $I \subset \{1, ..., p\}$, $b > 0$,

$$\Theta_I(b) = \left\{ \theta \in \Theta(b) : \quad \forall i \notin I, \theta_i = 0 \right\}$$

and

$$\Theta_I = \left\{ \theta \in \Theta : \quad \forall i \notin I, \theta_i = 0 \right\}.$$

Finally, let us put π_b^I *the uniform probability measure on* $\Theta_I(b+1)$. *We put, for some* $b > 0$,

$$\pi_b \propto \sum_{k=0}^{n} 2^{-k-1} \sum_{\substack{I \subset \{1, ..., p\} \\ |I| = k}} \binom{p}{k}^{-1} \pi_b^I.$$

Remark that in order to predict as well as the best predictor in $\Theta(b)$, the prior distribution has to be defined on $\Theta(b+1)$, for technical reasons that will become clear in the proofs (see the appendix). We are now ready to give the definition of the Gibbs estimator.

Definition 4 (Gibbs estimator) *We define, for any $b>0$ and $\lambda>0$, $\hat{\rho}_{\lambda,b}$ such that*

$$\frac{d\hat{\rho}_{\lambda,b}}{d\pi_b}(\theta)=\frac{\exp\left[-\lambda r(\theta)\right]}{\int_{\Theta(b)}\exp[-\lambda r]d\pi_b},$$

and we put

$$\hat{\theta}_{\lambda,b}=\int_{\Theta(b)}\theta\hat{\rho}_{\lambda}(d\theta). \tag{1}$$

The parameter λ is called the inverse temperature parameter. Its choice is a problem in practice, see the discussions in Catoni (2003, 2004, 2007); Alquier (2008). In theory, we will see that λ of the order n will lead to fast rates for prediction. In practice, $\lambda=n/\hat{\mathrm{var}}(X)$ leads to satisfying results in our simulations, where $\hat{\mathrm{var}}(X)$ is the empirical variance of the observed time series. The practical computation of $\hat{\theta}_{\lambda,b}$ can also be a problem. In Dalalyan and Tsybakov (2008) a Langevin Monte-Carlo algorithm is used. Here, as in Alquier and Lounici (2011), the Reversible Jump MCMC of Green (1995) is used, see Section 6.

4. Theoretical results

Theorem 1 (Oracle inequality for the Gibbs estimator) *Assume that $\|\overline{\theta}\|_1<b$ and that there exists a constant $\Phi(q)$ such that for any $n\in\mathbb{N}$, $\Phi(q)\geq K_{\phi}^{(n)}(q)$. Choose*

$$\eta\in\left(0,\frac{16}{\Phi(q)}\right]\quad\text{and}\quad\lambda=\frac{\eta(n-q)}{64\Phi(q)(2+b)^2B^2}.$$

We have, with probability at least $1-\varepsilon$ on the drawing of the sample $(X_1,\cdots,X_n)$,

$$R(\hat{\theta}_{\lambda,b})-R(\overline{\theta})\leq\inf_{\substack{I\subset\{1,\dots,p\}\\ |I|<\frac{\eta(n-q)}{32\Phi(q)(2+b)^2}\\ \theta\in\Theta_I(b)}}\left\{\left(\frac{2+\eta}{2-\eta}\right)\left(R(\theta)-R(\overline{\theta})\right)\right.$$
$$\left.+\frac{64\Phi(q)(2+b)^2B^2}{(n-q)\eta}\left[|I|\left(B+2\log\left(\frac{Bbpe}{|I|}\sqrt{\frac{2\eta(n-q)}{|I|}}\right)\right)+2\log\left(\frac{2}{\varepsilon}\right)\right].\right\}$$

The full proof is given in the appendix. In order to understand this result, it is particularly useful to think of a particular case where there is a sparse optimal predictor: we assume that there is a $\overline{\theta}\in\arg\min_{\Theta(b)}R$ that has only a few number p_0 of non-zero coordinates. This is the classical "sparsity" assumption. Then in this case, taking $\theta=\overline{\theta}$ in the previous result leads to

$$R(\hat{\theta}_{\lambda,b})-R(\overline{\theta})\leq\frac{64\Phi(q)(2+b)^2B^2}{(n-q)\eta}\left\{p_0\left[B+2\log\left(\frac{Bbpe}{p_0}\sqrt{\frac{2\eta(n-q)}{p_0}}\right)\right]+2\log\left(\frac{2}{\varepsilon}\right)\right\} \tag{2}$$

for n large enough - actually, $n > q+p_0[32\Phi(q)(2+b)^2]/\eta$. We obtain that this is not the true dimension p of Θ that determines a rate p/n, but the intrinsic dimension p_0 of $\overline{\theta}$ as the rate is $p_0 \log(pn)/n$. With iid observations, Dalalyan and Tsybakov (2008); Alquier and Lounici (2011) obtained the same result, with rate $p_0 \log(p)/n$. In Gerchinovitz (2011), the same rate is reached in the context of prediction of individual sequences.

Note that of course the strength of Theorem 1 when compared to Inequality 2 is that it ensures that $\hat{\theta}_{\lambda,b}$ will give good prediction not only when $\overline{\theta}$ is sparse, but also when it can only be approximated by a sparse parameter θ.

Remark 1 *The value of λ proposed in the Theorem depends on the ϕ-mixing coefficients of the time series. Of course, these coefficients are unknown. One can check in the proof of Theorem 1 that any λ of the order of n would lead to the same rate of convergence, but with less precise constants. However, in practice, this does not tell us how to calibrate λ. It is of course possible to use a procedure such as cross-validation. However, in Dalalyan and Tsybakov (2008) or Alquier and Lounici (2011), it is observed that the value $\lambda = n/(4\sigma^2)$ or $\lambda = n/(2\sigma^2)$, where σ^2 is the variance of then noise, performs well in practice, and receives a theoretical justification in the iid setting. So we propose here the heuristic value $\lambda = n/\hat{\mathrm{var}}(X)$ leads to satisfying results in our simulations, where $\hat{\mathrm{var}}(X)$ is the empirical variance of the observed time series. We will see in Section 6 that it performs well on a set of simulations.*

5. Some examples of ϕ-mixing processes

In this section we study the behavior of the prediction procedure on some classical ϕ-mixing processes. In all the section (ϵ_t) denotes an iid sequence called the innovations.

5.1. The AR(p) model

We consider the case where the observations (X_t) satisfy an AR(p) model:

$$X_t = \sum_{j=1}^{p} a_j X_{t-j} + \epsilon_t, \qquad \forall t \in \mathbb{Z}. \tag{3}$$

Here both $p \in \{1, 2, \dots\}$ and (a_j) are unknown, (ϵ_t) is bounded with a distribution possessing an absolutely continuous component. Assume that $\mathcal{A}(z) = \sum_{j=1}^{p} a_j z^j$ has no root inside the unit disk in $\mathbb{C}$. Then it exists a stationary solution (X_t) that is an exponentially ϕ-mixing processes, in the sense that the coefficients ϕ_r decay exponentially fast, see Athreya and Pantula (1986).

5.2. The MA(q) model

We consider now observations (X_t) such that $X_t = \sum_{j=1}^{q} b_j \epsilon_{t-j}$ for all $t \in \mathbb{Z}$. Assume that $\mathcal{B}(z) = \sum_{j=1}^{q} b_j z^j$ has no root inside the unit disk in $\mathbb{C}$ so that (X_t) is invertible (admits an AR(∞) representation). By definition the process (X_t) is stationary and ϕ-dependent - it is even q-dependent, in the sense that $\phi_r = 0$ for $r > q$. Moreover it is bounded iff the innovations are bounded. So this process satisfies the assumptions of Theorem 1.

5.3. Non linear models

Consider an extension of the AR(p) model of the form

$$X_t = F(X_{t-1}, \dots, X_{t-p}; \epsilon_t), \qquad \forall t \in \mathbb{Z}. \tag{4}$$

To prepare the general case we recall some material from Meyn and Tweedie (1993). Remember that the observations are assumed to belong to the compact set $[-B, B]$. The Lagrange stability, irreducibility and aperiodicity conditions hold when the innovations admits a lower semi-continuous density on $[-B, B]$ and for any $|x| \le B$ we have

$$[-B, B] = A^+(x) := \{F_k(x, w_1, \dots, w_k); \quad k \ge 1,\ (w_1, \dots, w_k) \in \mathrm{Support}^k(\epsilon)\}$$

with $F_k : \mathbb{R}^{k+1} \mapsto \mathbb{R}$ defined recursively by the relation $F_{k+1}(\cdot, w) = F(F_k(\cdot), w)$, $F_1 = F$. A direct application of Proposition 7.5 of Meyn and Tweedie (1993) yields that (X_t) is a T-chain (we refer to Meyn and Tweedie (1993) for the definition) if $F_1(x, w)$ is continuously differentiable on w and for each $x_0 \in \mathbb{R}^p$ there exists $(w_k)_{1 \le k \le p}$ such that $\partial F_k / \partial w_k(x_0, w_1, \dots, w_k) \ne 0$ for all $1 \le k \le p$. For example the generalized AR-GARCH models of the form $F(x, w) = R(x) + \sigma(x) w$ with R and $\sigma > 0$ continuously differentiable is a T-chain.

Assume that (X_t) is an irreducible, aperiodic, Lagrange stable T-chain. Then it satisfies the Doeblin condition and is thus exponentially ϕ-mixing, see Theorem 16.2.7 of Meyn and Tweedie (1993).

6. Implementation and simulations

6.1. RJMCMC method

The Gibbs estimator, given by (1), takes the form of an integral over a large dimensional space. It can thus be computed by Monte Carlo methods. This is actually a classical approach for Bayesian estimators, see e.g. Marin and Robert (2007); Robert (1996). Here, we use the RJMCMC algorithm - Reversible Jumb Markov Chain Monte Carlo, Green (1995). This method is implemented for example in Alquier and Lounici (2011) to compute a Gibbs estimator that takes exactly the same form than ours.

6.2. Simulations study in the AR case

We compare here the Gibbs estimator given by (1) to the “classical approach” in the AR case. This approach, for example as implemented in the R software (R Development Core Team (2008)), computes the least square estimator in each submodel AR(p) and then selects the order p by Akaike's AIC criterion Akaike (1973).

We generate the data according to the following models:

$$X_t = 0.5X_{t-1} + 0.1X_{t-2} + \varepsilon_t \tag{5}$$

$$X_t = 0.6X_{t-4} + 0.1X_{t-8} + \varepsilon_t \tag{6}$$

$$X_t = \cos(X_{t-1}) \sin(X_{t-2}) + \varepsilon_t \tag{7}$$

Table 1: Performances of the Gibbs estimator, AIC and least square estimator in the full model, on the simulations. Each simulation is repeated 20 times, we report on Table 1 the mean performance and standard deviation of each method. We highlight the best result for each line.

n	Model	Innovations	Gibbs	AIC	Full Model
100	(5)	unif.	**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)
	(6)	unif.	**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)
	(7)	unif.	**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	(5)	unif.	**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)
	(6)	unif.	**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)
	(7)	unif.	**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)

where ε_t is the innovation. We will use two models for the innovation: the uniform case, $\varepsilon_t \sim \mathcal{U}[-a, a]$, and the Gaussian case, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. In the first case, the processes defined in (5), (6) and (7) satisfy the assumptions of Theorem 1 (see Section 5) while the Gaussian case is more classical in statistics, so it is worth testing if our method performs well in this context too - even if our method does not receive any theoretical justification in this case, as it is show in Doukhan (1994) that autoregressive processes with gaussian noise are not ϕ-mixing. We take $\sigma = 0.4$ and $a = 0.70$ (In both cases this leads to $\mathrm{Var}(\varepsilon_t) \simeq 0.16$). The Gibbs estimator is used on all the possible AR models as in Example 1; we fix $q = 20$ and $\lambda = n/\hat{\mathrm{var}}(X)$, where $\hat{\mathrm{var}}(X)$ is the empirical variance of the observed time series. We compare its performances to the ones of AIC criterion as implemented in the R software and to the basic least square estimator in the model $AR(q)$ - that we will call “full model”. The experimental design is the following: for each model, we 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. We report the performances on the test set. We take $n = 100$ and $n = 1000$ in the simulations. Each simulation is repeated 20 times, we report on Table 1 the mean performance and standard deviation of each method.

It is interesting to note that our estimator performs better on Model (6) and Model (7) while AIC performs slightly better on Model (5). The differences tends to be less perceptible when n grows - this is coherent with the fact that we develop here a non-asymptotic theory. It is also interesting to note that our estimator seems to work well even in the case of a Gaussian noise.

7. Conclusion

We proved that the Gibbs estimator can reach fast rates in the case of ϕ-mixing time series. It would now be interesting to extend this result to a more general class of processes, e.g. weakly dependent ones. Note however the versions of Bernstein’s inequality known in the context of weak dependence (see e.g. Dedecker et al. (2007); Wintenberger (2010)) do not allow to reach this rate up to our knowledge. More generally, the question of concentration of measure for time series is on a large part still open.

Another question is to provide a theoretical justification to our heuristic for the tuining of λ in practice.

8. Proof of Theorem 1

We start by a short overview of the proof. First, we state a result, Lemma 1, that provides a control of the difference between the risk and the empirical risk of a predictor. The main tool for the proof of this result is Samson's version of Bernstein's inequality in Lemma 3, that we remind in the appendix. Lemma 1 is then used together with Donsker-Varadhan variational formula (also reminded in the appendix, Lemma 4) to prove a PAC-Bayesian type oracle inequality similar to the ones in Catoni (2004), Lemma 2, that is the main tool used to prove Theorem 1

Lemma 1 *Under the hypothesis of Theorem 1, we have, for any $\theta \in \Theta(b+1)$, for any $0 \leq \lambda \leq (n-q)/[4(2+b)^2 B^2 \Phi^2(q)]$,*

$$\mathbb{E}\exp\left\{\lambda\left[\left(1-\frac{32\Phi(q)\lambda(2+b)^2B^2}{n-q}\right)\left(R(\theta)-R(\overline{\theta})\right)-r(\theta)+r(\overline{\theta})\right]\right\}\leq 1,$$

and

$$\mathbb{E}\exp\left\{\lambda\left[\left(1+\frac{32\Phi(q)\lambda(2+b)^2B^2}{n-q}\right)\left(R(\overline{\theta})-R(\theta)\right)-r(\overline{\theta})+r(\theta)\right]\right\}\leq 1.$$

Proof [Proof of Lemma 1] We apply Samson's version of Bernstein's inequality (see Lemma (3) in the Appendix) to $N = n-q$, $Z_i = (X_{i+1}, \ldots, X_{i+q})$,

$$f(Z_i) = \frac{1}{n-q}\Big[R(\theta)-R(\overline{\theta}) \\ - (X_{i+q} - f_\theta(X_{i+q-1},\ldots,X_{i+1}))^2 + \big(X_{i+q} - f_{\overline{\theta}}(X_{i+q-1},\ldots,X_{i+1})\big)^2\Big].$$

Note that we have:

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

and the Z_i are uniformly mixing with coefficients $\phi^Z_r = \phi_{\lfloor r/q \rfloor}$. Note that $K_{\phi^Z} = 1 + \sum_{r=1}^{n-q}\sqrt{\phi_{\lfloor r/q\rfloor}} = K_\phi^{(n)}(q) \leq \Phi(q)$. For any θ and θ' in Θ let us put

$$V(\theta,\theta') = \mathbb{E}_{\mathbb{P}}\left\{\left[\Big(X_{q+1} - f_\theta(X_q,...,X_1)\Big)^2 - \Big(X_{q+1} - f_{\theta'}(X_q,...,X_1)\Big)^2\right]^2\right\}.$$

Noticing that $\sigma^2(f) \leq V(\theta,\overline{\theta})/(n-q)^2$ and that $\|f\|_\infty \leq 4(2+b)^2B^2/(n-q)$, for any $0 \leq \lambda \leq (n-q)/[4(2+b)^2B^2\Phi^2(q)]$, we have

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

Notice also that

$$
\begin{aligned}
V(\theta,\overline{\theta}) &= \mathbb{E}_{\mathbb{P}}\left\{\left[2X_{q+1}-(f_{\theta}+f_{\overline{\theta}})(X_q,...,X_1)\right]^2\left[(f_{\theta}-f_{\overline{\theta}})(X_q,...,X_1)\right]^2\right\} \\
&\leq (2+\|\theta\|_1+\|\overline{\theta}\|_1)^2B^2\mathbb{E}_{\mathbb{P}}\left\{\left[(f_{\theta}-f_{\overline{\theta}})(X_q,...,X_1)\right]^2\right\} \\
&= (2+\|\theta\|_1+\|\overline{\theta}\|_1)^2B^2\left[R(\theta)-R(\overline{\theta})\right] \leq 4(2+b)^2B^2\left[R(\theta)-R(\overline{\theta})\right]
\end{aligned}
$$

as $\theta\in\Theta(b+1)$ and $\overline{\theta}\in\Theta(b)\subset\Theta(b+1)$. This proves the first inequality of Lemma 1. The second inequality is proved exacly in the same way, but replacing f by $-f$. ■

We are now ready to state the following key result. Note that the very classical definition of the Kullback divergence $\mathcal{K}(\rho,\pi)$ is reminded in the appendix.

Lemma 2 (PAC-Bayesian oracle inequality for a ϕ-mixing process) *Under the hypothesis of Theorem 1, we have, for any $0\leq\lambda\leq(n-q)/[4(2+b)^2B^2\Phi^2(q)]$, for any $0<\varepsilon<1$,*

$$
\mathbb{P}\left\{\begin{array}{l}
\forall\rho\in\mathcal{M}_+^1(\Theta(b+1)), \\
\left(1-\frac{32\Phi(q)\lambda(2+b)^2B^2}{n-q}\right)\left(\int R\mathrm{d}\rho-R(\overline{\theta})\right)\leq\int r\mathrm{d}\rho-r(\overline{\theta})+\frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{2}{\varepsilon}\right)}{\lambda} \\
\textit{and} \\
\int r\mathrm{d}\rho-r(\overline{\theta})\leq\left(\int R\mathrm{d}\rho-R(\overline{\theta})\right)\left(1+\frac{32\Phi(q)\lambda(2+b)^2B^2}{n-q}\right)+\frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{2}{\varepsilon}\right)}{\lambda}
\end{array}\right\}\geq 1-\varepsilon.
$$

Proof [Proof of Lemma 2] Let us fix ε, λ and $\theta\in\Theta(b+1)$, and apply the first inequality of Lemma 1. We have:

$$
\mathbb{E}\exp\left\{\lambda\left[\left(1-\frac{32\Phi(q)\lambda(2+b)^2B^2}{n-q}\right)\left(R(\theta)-R(\overline{\theta})\right)-r(\theta)+r(\overline{\theta})\right]\right\}\leq 1,
$$

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

$$
\mathbb{E}\int\exp\left\{\lambda\left[\left(1-\frac{32\Phi(q)\lambda(2+b)^2B^2}{n-q}\right)\left(R(\theta)-R(\overline{\theta})\right)-r(\theta)+r(\overline{\theta})+\log\left(\frac{\epsilon}{2}\right)\right]\right\}\pi_b(\mathrm{d}\theta) \leq\frac{\varepsilon}{2}.
$$

We apply Donsker-Varadhan variational formula (see Lemma 4 in the appendix) and we get:

$$
\mathbb{E}\exp\left\{\sup_{\rho}\lambda\left[\left(1-\frac{32\Phi(q)\lambda(2+b)^2B^2}{n-q}\right)\left(\int R\mathrm{d}\rho-R(\overline{\theta})\right)-\int r\mathrm{d}\rho+r(\overline{\theta})+\log\left(\frac{\epsilon}{2}\right) -\mathcal{K}(\rho,\pi)\right]\right\}\leq\frac{\varepsilon}{2}.
$$

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

$$\mathbb{P}\left\{\sup_{\rho} \lambda \left[\left(1 - \frac{32\Phi(q)\lambda(2+b)^2 B^2}{n-q}\right)\left(\int R \mathrm{d}\rho - R(\overline{\theta})\right) - \int r \mathrm{d}\rho + r(\overline{\theta}) \right.\right.$$
$$\left.\left. + \log\left(\frac{\epsilon}{2}\right)\right] - \mathcal{K}(\rho, \pi) \geq 0\right\} \leq \frac{\varepsilon}{2}.$$

Now, we follow the same proof again but starting with the second inequality of Lemma 1. We obtain:

$$\mathbb{P}\left\{\sup_{\rho} \lambda \left[\left(1 + \frac{32\Phi(q)\lambda(2+b)^2 B^2}{n-q}\right)\left(R(\overline{\theta}) - \int R \mathrm{d}\rho\right) - r(\overline{\theta}) + \int r \mathrm{d}\rho \right.\right.$$
$$\left.\left. + \log\left(\frac{\epsilon}{2}\right) - \mathcal{K}(\rho, \pi)\right] \geq 0\right\} \leq \frac{\varepsilon}{2}.$$

A union bound ends the proof. ■

We are now ready to give the proof of Theorem 1.

Proof First, we apply Lemma 2. From now, a work on the event of probability at least $1-\varepsilon$ given by this lemma. In particular we have $\forall \rho \in \mathcal{M}_+^1(\Theta)$,

$$\int R \mathrm{d}\rho - R(\overline{\theta}) \leq \frac{\int r \mathrm{d}\rho - r(\overline{\theta}) + \frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}}{1 - \frac{32\Phi(q)\lambda(2+b)^2 B^2}{n-q}}.$$

For the sake of simplicity, during this proof, we will use the following notation:

$$C = 32\Phi(q)(2+b)^2 B^2.$$

Taking $\rho = \hat{\rho}_{\lambda,b}$ leads to:

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

We apply Lemma 4 to see that:

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

Now, we use the second inequality of Lemma 2 to see that

$$\begin{aligned}
\int R \mathrm{d}\hat{\rho}_{\lambda,b} - R(\overline{\theta}) &\leq \inf_{\rho} \frac{\left(1 + \frac{\lambda C}{n-q}\right)\left(\int R \mathrm{d}\rho - R(\overline{\theta})\right) + 2\frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}}{1 - \frac{\lambda C}{n-q}} \\
&\leq \inf_{I \subset \{1,\dots,q\}} \inf_{\rho \ll \pi_b^I} \frac{\left(1 + \frac{\lambda C}{n-q}\right)\left(\int R \mathrm{d}\rho - R(\overline{\theta})\right) + 2\frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}}{1 - \frac{\lambda C}{n-q}}. \quad (8)
\end{aligned}$$

By Jensen's inequality,

$$\int R\mathrm{d}\hat{\rho}_{\lambda,b} \geq R\left(\hat{\theta}_{\lambda,b}\right).$$

Also remark that, as soon as $\rho \ll \pi_b^I$,

$$\begin{aligned}\mathcal{K}(\rho,\pi) &= (|I|+1)\log(2) + \log\binom{p}{|I|} + \mathcal{K}(\rho,\pi_b^I) \\ &\leq (|I|+1)\log(2) + |I|\log\left(\frac{pe}{|I|}\right) + \mathcal{K}(\rho,\pi_b^I)\end{aligned}$$

(see, e.g., Catoni (2003) page 190). Now, for any $0<\delta<1$, for any $I \subset \{1,...,p\}$, and $\theta \in \Theta_I(B)$, we take $\rho_{\delta,I,\theta}$ as the uniform measure on $\{t \in \Theta_I(b) : \|t-\theta\|_1 \leq \delta\}$. Note that as $\theta \in \Theta_I(B)$ and $\delta < 1$, the support of $\rho_{\delta,I,\theta}$ is included in $\Theta(b+1)$ the support of π_b. This is the reason why π_b is defined in this way. Inequality (8) leads to

$$\begin{aligned}&R\left(\hat{\theta}_{\lambda,b}\right) - R(\overline{\theta}) \\ &\leq \frac{1}{1-\frac{\lambda C}{n-q}} \inf_{\delta>0} \inf_{I\subset\{1,\dots,q\}} \inf_{\theta\in\Theta_I(b)} \Bigg\{ \left(1+\frac{\lambda C}{n-q}\right)\left(R(\theta)+B^2\delta^2 - R(\overline{\theta})\right) \\ &\qquad + 2\frac{(|I|+1)\log(2) + |I|\log\left(\frac{pe}{|I|}\right) + |I|\log\left(\frac{b}{\delta}\right) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}\Bigg\}\end{aligned}$$

and so, by choosing $\delta = \sqrt{|I|/(2B^2\lambda)}$, we get:

$$\begin{aligned}&R\left(\hat{\theta}_{\lambda,b}\right) - R(\overline{\theta}) \\ &\leq \frac{1}{1-\frac{\lambda C}{n-q}} \inf_{\delta>0} \inf_{I\subset\{1,\dots,q\}} \inf_{\theta\in\Theta_I(b)} \Bigg\{ \left(1+\frac{\lambda C}{n-q}\right)\left(R(\theta) - R(\overline{\theta})\right) \\ &\qquad + \frac{|I|\left(B + 2\log\left(\frac{2Bbpe}{|I|}\sqrt{\frac{\lambda}{|I|}}\right)\right) + 2\log\left(\frac{4}{\varepsilon}\right)}{\lambda}\Bigg\}\end{aligned}$$

Remember that $\lambda \leq (n-q)c$ where we put for short $c = 1/[4(2+b)^2B^2\Phi^2(q)]$. Let us take $\lambda = \eta(n-q)/(2C)$ for some constant η. Remark that $\eta \leq 2cC$ ensures that $\lambda \leq (n-q)c$ while we need to impose $|I| < \eta B^2(n-q)/C$ in order to ensure that $\delta < 1$. We obtain:

$$\begin{aligned}\mathbb{P}\Bigg\{ R(\hat{\theta}_{\lambda,b}) - R(\overline{\theta}) \leq \inf_{\substack{I\subset\{1,\dots,p\}\\ |I|<\eta B^2(n-q)/C \\ \theta\in\Theta_I(b)}} \Bigg[\left(\frac{2+\eta}{2-\eta}\right)\left(R(\theta)-R(\overline{\theta})\right) \\ + \frac{2C}{(n-q)\eta}\left(|I|\left(B + 2\log\left(\frac{Bbpe}{|I|}\sqrt{\frac{2\eta(n-q)}{|I|}}\right)\right) + 2\log\left(\frac{2}{\varepsilon}\right)\right)\Bigg]\Bigg\} \geq 1-\varepsilon.\end{aligned}$$

We end the computation by the remark that $\lambda = \eta(n-q)/(2C) = \eta(n-q)/[64\Phi(q)(2+b)^2B^2]$ and that $\eta \leq 2cC = 16/\Phi(q)$. ∎

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281. Budapest: Akademia Kiado, 1973.

P. Alquier. Pac-bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008.

P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

P. Alquier and O. Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli* (to appear), available on arXiv:0902.2924, 2012.

K. B. Athreya and S. G. Pantula. Mixing properties of Harris chains and autoregressive processes. *J. Appl. Probab.*, 23(4):880–892, 1986. ISSN 0021-9002.

J.-Y. Audibert. Théorie statistique de l'apprentissage: une approche pac-bayésienne. HDR Université Paris VI, 2004.

J.-Y. Audibert. Pac-bayesian aggregation and multi-armed bandits. HDR Université Paris Est, 2010.

P. Brockwell and R. Davis. *Time Series: Theory and Methods (2nd Edition)*. Springer, 2009.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

O. Catoni. A pac-bayesian approach to adaptative classification. *Preprint Laboratoire de Probabilités et Modèles Aléatoires*, 2003.

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

O. Catoni. *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*, volume 56 of *Lecture Notes-Monograph Series*. IMS, 2007.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, 2006.

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

J. Dedecker, P. Doukhan, G. Lang, J. R. León, S. Louhichi, and C. Prieur. *Weak Dependence, Examples and Applications*, volume 190 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 2007.

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

P. Doukhan. *Mixing*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1994.

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

P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

J. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

I. A. Ibragimov. Some limit theorems for stationary processes. *Theory of Probability and its Application*, 7(4):349–382, 1962.

N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

J.-M. Marin and C. P. Robert. *Bayesian Core: A practical approach to computational Bayesian analysis*. Springer, 2007.

D. A. McAllester. Pac-bayesian model averaging. In *Procs. of of the 12th Annual Conf. On Computational Learning Theory, Santa Cruz, California (Electronic)*, pages 164–170. ACM, New-York, 1999.

R. Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39:5–34, 2000.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993. ISBN 3-540-19832-6.

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

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, 2008.

C. P. Robert. *Méthods de Monte Carlo par chaines de Markov*. Economica (Paris), 1996.

P.-M. Samson. Concentration of measure inequalities for markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000.

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, P. Auer, and J. Shawe-Taylor. *PAC-Bayesian inequalities for martingales*. arXiv:1110.6886, 2011.

J. Shawe-Taylor and R. Williamson. A pac analysis of a bayes estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory, COLT'97*, pages 2–9. ACM, 1997.

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(2):66–106, 2010.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

A. Tsybakov. Optimal rates of aggregation. In B. Schölkopf and M. K. Warmuth, editors, *Learning Theory and Kernel Machines*, pages 303–313. Springer LNCS, 2003.

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

O. Wintenberger. Deviation inequalities for sums of weakly dependent time series. *Electronic Communications in Probability*, 15:489–503, 2010.

Appendix A. Samson's version of Bernstein's inequality and Donsker-Varadhan variational formula

Lemma 3 (Samson (2000) (page 460, line7)) *Let $N \in \mathbb{N}$. Let $(Z_i)_{i\in\mathbb{Z}}$ be a stationary process, let (ϕ_r^Z) denote its ϕ-mixing coefficients, let f be a measurable function $\mathbb{R} \to [-M, M]$ and let*

$$S_N(f) := \sum_{i=1}^{N} f(Z_i).$$

Then:

$$\ln \mathbb{E}(\exp(\lambda(S(f) - \mathbb{E}S(f)))) \leq 8K_{\phi^Z} N\sigma^2(f)\lambda^2, \text{ for all } 0 \leq \lambda \leq 1/(MK_{\phi^Z}^2),$$

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

Definition 5 *Given a measurable space $(E, \mathcal{E})$ we let $\mathcal{M}_+^1(E)$ denote the set of all probability measures on $(E, \mathcal{E})$. The Kullback divergence is a pseudo-distance on $\mathcal{M}_+^1(E)$ defined, for any $(\pi, \pi') \in [\mathcal{M}_+^1(E)]^2$ by the equation*

$$\mathcal{K}(\pi, \pi') = \begin{cases} \pi[\log(d\pi/d\pi')] & \text{if } \pi \ll \pi', \\ \\ +\infty & \text{otherwise.} \end{cases}$$

with the convention that $\pi[h] = \int h(x)\pi(\mathrm{d}x)$ for any measurable function h.

Lemma 4 (Donsker and Varadhan (1976) variational formula) *For any π in the set $\mathcal{M}_+^1(E)$, for any measurable function $h : E \to \mathbb{R}$ such that $\pi[\exp(h)] < +\infty$ we have:*

$$\pi[\exp(h)] = \exp\left(\sup_{\rho \in \mathcal{M}_+^1(E)} \left(\rho[h] - \mathcal{K}(\rho, \pi) \right) \right), \tag{9}$$

with convention $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of π, the supremum with respect to ρ in the right-hand side is reached for the Gibbs measure $\pi\{h\}$ defined by

$$\pi\{h\}(\mathrm{d}x) = \frac{e^{h(x)}\pi(\mathrm{d}x)}{\pi[\exp(h)]}.$$