Exponential inequalities for nonstationary Markov Chains

Pierre Alquier*, Paul Doukhan† and Xiequan Fan‡

April 3, 2019

Abstract

Exponential inequalities are main tools in machine learning theory. To prove exponential inequalities for non i.i.d random variables allows to extend many learning techniques to these variables. Indeed, much work has been done both on inequalities and learning theory for time series, in the past 15 years. However, for the non independent case, almost all the results concern stationary time series. This excludes many important applications: for example any series with a periodic behaviour is nonstationary. In this paper, we extend the basic tools of Dedecker and Fan (2015) to nonstationary Markov chains. As an application, we provide a Bernstein-type inequality, and we deduce risk bounds for the prediction of periodic autoregressive processes with an unknown period.

1 Introduction

Exponential inequalities are corner stones of machine learning theory. For example, distribution free generalization bounds were proven by Vapnik and Cervonenkis based on Hoeffding's inequality, see Vapnik (1998). Model selection bounds in Massart (2007) also rely on exponential moment inequalities.

To prove such inequalities in the non i.i.d setting is thus crucial to study the generalization ability of machine learning algorithms on time series. As an example, a Bernstein type inequality for α-mixing time series is proven in Modha and Masry (2002). This result is used by Steinwart and Christmann (2009) to prove generalization bounds for general learning problems with α-mixing observations.

*CREST, ENSAE, Université Paris Saclay. Pierre Alquier's work has been supported by GENES and by the French National Research Agency (ANR) under the grant Labex Ecodec (ANR-11-LABEX-0047).

†AGM UMR8088 University Paris-Seine and CIMFAV, Universidad de Valparaiso, Chile. Paul Doukhan's work has been developed within the MME-DII center of excellence (ANR-11-LABEX-0023-01) & PAI-CONICYT MEC N^o80170072.

‡CAM, Tianjin University, Tianjin, China. Fan Xiequan has been partially supported by the National Natural Science Foundation of China (Grant N^o11601375.

Exponential inequalities and machine learning with non-i.i.d observations actually became an important research direction, a more detailed list of references is given below. However, most of these references assume stationarity. That is, only the independence assumption was removed. The observations are still assumed to be identically distributed, or at least erdogic. This excludes many applications: in addition to trends, data related to a human activity such as in industry or economics has some periodicity (hourly, daily, yearly...) and some regime switching; the same remark applies to data with a physical origin, such as in geology, astrophysics...

In this paper, the inequalities proven by Dedecker and Fan (2015) for time homogeneous Markov chains to non-homogeneous chains. This allows to study a large set of nonstationary processes. We obtain Bernstein, McDiarmid inequality as well as moments inequalities. As an application, we study periodic autoregressive processes of the form $X_t = f_t(X_{t-1}) + \varepsilon_t$ where $f_{t+T} = f_t$ for any t, for some period T. Thanks to our version of Bernstein's inequality we show that the Empirical Risk Minimizer (ERM) leads to consistent predictions in this setting. We also show that a penalized version of the ERM enjoys the same property even when T is unknown.

The paper is organized as follows. The rest of this introduction is dedicated to a state-of-the-art on exponential inequalities for time series. Section 2 introduces the notations and assumptions that will be used in the whole paper. In Section 3, we state an extension of Proposition 2.1 of Dedecker and Fan (2015): this is Lemma 3.1. As a proof of concept, we use this lemma to prove a version of Bernstein inequality for nonstationary Markov chains. We also provide Cramer and McDiarmid inequalities based on this lemma. We study periodic autoregressive series in Section 4. Finally, Section 5 contains the proof of Lemma 3.1 and of the results in Section 4.

1.1 State of the art

We refer the reader to Boucheron et al. (2013) for an overview on exponential and concentration inequalities in the i.i.d case. This book also provides references for applications of these results to machine learning theory.

Exponential inequalities were proven for time under a various range of assumptions. We refer the reader to Doukhan (2018) for various approaches on modelling time series.

Inequalities for Markov chains $(X_t)_{t\geq 1}$ are proven in Catoni (2003); Adamczak (2008); Bertail and Clémençon (2010); Joulin and Ollivier (2010); Wintenberger (2017); Bertail and Portier (2018); Paulin (2018); Bertail and Ciolek (2019). Note that most of these inequalities require the chain to be time homogeneous. While this does not imply the chain to be stationary, in some sense the X_t's are asymptotically identically distributed in these papers. For example, consider the powerful renewal technique used in Bertail and Ciolek (2019) to prove a version of Bernstein inequality. The proof is based on the fact that blocks $(X_{\tau_i}, \dots, X_{\tau_{i+1}-1})$ between two *renewal times* τ_i and τ_{i+1} are actually i.i.d. It is thus possible to apply the i.i.d version of Bernstein inequality to these

blocks. The spectral technique used in Paulin (2018) still relies on the ergodicity of the Markov chain (we thank the anonymous Referee for pointing out some of these references). Exponential inequalities for hidden Markov chains are given in Kontorovich and Ramanan (2008).

It is a well-known fact that Hoeffding's inequality is not only valid for independent observations, but also for martingales increments (it is sometimes refered to as Hoeffding-Azuma inequality in this case). To decompose a function of the process as a sum of martingales increments is actually one of the most powerful techniques to prove exponential inequalities, see Chapter 3 in Boucheron et al. (2013). More exponential inequalities for martingales can be found in Seldin *et al* (2012); Rio (2013a); Bercu et al. (2015). This technique is actually used by Dedecker and Fan (2015); Fan et al. (2018) to prove exponential inequalities for Markov chains.

Markov chains are extremely useful in modelisation and simulations, however, many time series have a very different dependence structure. Mixing coefficients allow to quantify the dependence between observations without giving an explicit structure on this dependence. We refer the reader to Rio (2017) for a comprehensive introduction. Exponential inequalities for mixing processes are proven in Samson (2000); Merlevède et al. (2009); Rio (2017); Hang and Steinwart (2017). Mixing series however exclude many stochastic processes, as discussed in the monograph Dedecker et al. (2007). Weak dependance coefficients cover a wider range of processes for which Bernstein type inequalities are proven for example in Collet et al. (2002); Doukhan and Neumann (2007); Wintenberger (2010); Merlevède et al. (2011); Blanchard and Zadorozhnyi (2017). Dynamic systems are examples of processes where only X_1 is random, each X_t is then a deterministic fonction of X_{t-1}. Based on weak dependence arguments, it is possible to prove exponential inequalities for such processes Collet et al. (2002).

Based on such inequalities, it is possible to prove generalization bounds for machine learning algorithms Steinwart et al. (2009); Steinwart and Christmann (2009); Shalizi and Kontorovich (2013); London et al. (2014); Hang and Steinwart (2014); Sanchez-Perez (2015); Kuznetsov and Mohri (2015); McDonald et al. (2017); Alquier and Guedj (2018). Model selection techniques in the spirit of Massart (2007) are studied in Meir (2000); Lerasle (2011); Alquier and Wintenberger (2012), and aggregation of estimators in Alquier et al. (2013).

In this paper we prove provide tools to prove exponential inequalities for non-stationary, non homogeneous Markov chains. Rather than the renewal or spectral techniques discussed above, we extend the martingale approach of Dedecker and Fan (2015) to non-homogeneous chains.

2 Notations

From now, all the random variables are defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Let $(\mathcal{X}, d)$ and $(\mathcal{Y}, \delta)$ be two complete separable metric spaces. Let $(\varepsilon_t)_{t\geq 2}$ be a sequence of i.i.d $\mathcal{Y}$-valued random variables. Let X_1 be a $\mathcal{X}$-valued random

variable independent of $(\varepsilon_t)_{t\geq 2}$. Let $(X_t)_{t\geq 1}$ be the Markov chain given by

$$X_t = F_t(X_{t-1}, \varepsilon_t), \quad \text{for } t \geq 2, \tag{1}$$

where the functions $F_t : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ are such that

$$\sup_t \mathbb{E}\big[d\big(F_t(x, \varepsilon_1), F_t(x', \varepsilon_1)\big)\big] \leq \rho d(x, x') \tag{2}$$

for some constant $\rho \in [0, 1)$, and

$$\sup_t d(F_t(x, y), F_t(x, y')) \leq C\delta(y, y') \tag{3}$$

for some constant $C > 0$. In particular, when $F_t \equiv F$, this is the model studied by Dedecker and Fan (2015).

This class of Markov chains, that we call "one-step contracting", contains a lot of pertinent examples. The classical AR(1)-process is given by $X_t = F(X_{t-1}, \varepsilon_t)$ where $F(x, y) = ax+y$. Condition (3) is satisfied, and Condition (2) will be satisfied as soon as $|a| < 1$. Now, consider a time-varying AR(1) process:

$$X_t = a_t X_{t-1} + \varepsilon_t.$$

This process may be non-stationary. Condition (3) is still satisfied, and $|F_t(x, y)-F_t(x', y)| \leq |a_t||x-x'|$ so Condition (2) will be satisfied as soon as $\sup_t |a_t| < 1$. This process is studied by Bardet and Doukhan (2018) under various assumptions: local stationarity, that means a slow variation of a_t as a function of t, see Dahlhaus (1996), and periodicity, that is, for any t: $a_{t+T} = a_t$ for some (known) period T. If T is unknown, a cross-validation procedure to estimate T is proposed (Remark 2.4) without a consistency result. Below we will propose a penalized procedure with some statistical guarantees.

As a much more general example, consider the following functional auto-regressive model. Let $\mathcal{X}$ be a separable Banach space with norm $|\cdot|$. The functional auto-regressive model is defined by

$$X_t = f_t(X_{t-1}) + \varepsilon_t\,,$$

where $f_t : \mathcal{X} \to \mathcal{X}$ is such that

$$|f_t(x) - f_t(x')| \leq \rho|x - x'|.$$

Clearly (1) and (2) are satisfied once $\rho \in [0, 1)$, see Diaconis and Freedman (1999) for more examples.

We introduce the natural filtration of the chain $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and for $t \in \mathbb{N}$, $\mathcal{F}_t = \sigma(X_1, X_2, \ldots, X_t)$.

Consider a separately Lipschitz function $f : \mathcal{X}^n \to \mathbb{R}$ such that

$$|f(x_1, \ldots, x_n) - f(x'_1, \ldots, x'_n)| \leq \sum_{t=1}^{n} d(x_t, x'_t)\,.$$

We define

$$S_n := f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]\,. \tag{4}$$

The objective of what follows will be to derive inequalities on the tails of $\mathbb{P}(|S_n| > x)$.

3 Main results

Dedecker and Fan (2015) proved several exponential and moments inequalities if $F_t \equiv F$, by using a martingale decomposition. We will first extend this martingale decomposition to the general case. As an example, we will use it to prove a Bernstein type inequality. Other inequalities are given in the appendix. Here $(X_t)_{t\geq 1}$ will be a Markov chain satisfying (1) for some functions $(F_t)_{t\geq 2}$ satisfying (2).

3.1 Main lemma: martingale decomposition

Set $K_t(\rho) = (1-\rho^{t+1})/(1-\rho)$ for $t \geq 0, \rho \in [0,1)$ and

$$g_k(X_1,\dots,X_k) = \mathbb{E}[f(X_1,\dots,X_n)|\mathcal{F}_k]\,,$$

and

$$d_k = g_k(X_1,\dots,X_k) - g_{k-1}(X_1,\dots,X_{k-1})\,.$$

Define $S_t := d_1 + d_2 + \cdots + d_t$, for $t \in [1, n-1]$, and note that the functional S_n introduced in (4) satisfies indeed $S_n = d_1 + d_2 + \cdots + d_n$. Thus (S_t) is a martingale adapted to the filtration $\mathcal{F}_t$, and (d_t) is the martingale difference of (S_t).

Let P_{X_1} denote the distribution of X_1 and P_ε the (common) distribution of the ε_t's. Let G_{X_1}, G_ε and $H_{t,\varepsilon}$ be defined by

$$G_{X_1}(x) = \int d(x,x')P_{X_1}(dx'),$$

$$G_\varepsilon(y) = \int C\delta(y,y')P_\varepsilon(dy') \text{ and}$$

$$H_{t,\varepsilon}(x,y) = \int d(F_t(x,y),F_t(x,y'))P_\varepsilon(dy')\,.$$

We are now in position to state our main lemma.

Lemma 3.1. *Assume* (1) *and* (2)*, then:*

1. *The function g_t is separately Lipschitz and*

$$|g_t(x_1,\dots,x_t) - g_t(x'_1,\dots,x'_t)| \leq \sum_{i=1}^{t-1} d(x_i,x'_i) + K_{n-t}(\rho)d(x_t,x'_t)\,.$$

2. *The martingale difference (d_t) is such that*

$$|d_1| \leq K_{n-1}(\rho)G_{X_1}(X_1)\,,$$
$$\forall t \in [2,n],\ |d_t| \leq K_{n-t}(\rho)H_{t,\varepsilon}(X_{t-1},\varepsilon_t)\,.$$

3. *Assume moreover that the F_t's satisfy (3). Then $H_{t,\varepsilon}(x,y) \leq G_\varepsilon(y)$, and consequently, for $t \in [2,n]$,*

$$|d_t| \leq K_{n-t}(\rho) G_\varepsilon(\varepsilon_t)\,.$$

The proof of this lemma is given in Section 5. First, we want to show that the inequalities in this lemma can be used to prove exponential inequalities on S_n.

3.2 Application: Bernstein inequality

Note that van de Geer (1995) and de la Pena (1999) obtained some tight Bernstein type inequalities for martingales. Here, we can use the martingale decomposition and apply Lemma 3.1 to obtain the following result.

Theorem 3.1. *Assume that there exist some constants $M > 0, V_1 \geq 0$ and $V_2 \geq 0$ such that, for any integer $k \geq 2$,*

$$\mathbb{E}\Big[G_{X_1}(X_1)^k\Big] \leq \frac{k!}{2} V_1 M^{k-2}, \text{ and } \mathbb{E}\Big[G_\varepsilon(\varepsilon)^k\Big] \leq \frac{k!}{2} V_2 M^{k-2}\,. \tag{5}$$

Let $\delta = M K_{n-1}(\rho)$ and

$$V_{(n)} = V_1 \Big(K_{n-1}(\rho)\Big)^2 + V_2 \sum_{k=2}^{n} \Big(K_{n-k}(\rho)\Big)^2.$$

Then, for any $s \in [0, \delta^{-1})$,

$$\mathbb{E}\left[e^{\pm s S_n}\right] \leq \exp\left(\frac{s^2 V_{(n)}}{2(1 - s\,\delta)}\right)\,. \tag{6}$$

Consequently, for any $x > 0$,

$$\begin{aligned}\mathbb{P}\big(\pm S_n \geq x\big) &\leq \exp\left(\frac{-x^2}{V_{(n)}(1 + \sqrt{1 + 2x\delta/V_{(n)}}) + x\delta}\right) \\ &\leq \exp\left(\frac{-x^2}{2\,(V_{(n)} + x\delta)}\right)\,.\end{aligned}$$

The quantity $V_{(n)}$ can be computed explicitely from the definition for each n but note that

$$V_1 + (n-1)V_2 \leq V_{(n)} \leq \frac{V_1 + (n-1)V_2}{(1-\rho)^2}. \tag{7}$$

Proof. For any $s \in [0, \delta^{-1})$,

$$\mathbb{E}\left[e^{s d_1}\right] \leq 1 + \sum_{i=2}^{\infty} \frac{s^i}{i!}\,\mathbb{E}\left[|d_1|^i\right]$$

$$
\begin{aligned}
&\leq 1+\sum_{i=2}^{\infty}\frac{s^i}{i!}\Big(K_{n-1}(\rho)\Big)^i\mathbb{E}\left[\Big(G_{X_1}(X_1)\Big)^i\right]\\
&\leq 1+\sum_{i=2}^{\infty}\frac{s^i}{i!}\Big(K_{n-1}(\rho)\Big)^i\frac{i!}{2}V_1M^{i-2}\\
&= 1+\frac{s^2V_1\Big(K_{n-1}(\rho)\Big)^2}{2(1-s\,\delta)}\\
&\leq \exp\left(\frac{s^2V_1\Big(K_{n-1}(\rho)\Big)^2}{2(1-s\,\delta)}\right).
\end{aligned}
$$

We use Lemma 3.1 for the second inequality, the moment assumption for the third one, and the inequality $1+s\leq e^s$, for the final inequality. Similarly, for any $k\in[2,n]$,

$$
\mathbb{E}\left[e^{sd_k}|\mathcal{F}_{k-1}\right]\leq\exp\left(\frac{s^2V_2\Big(K_{n-k}(\rho)\Big)^2}{2(1-s\delta)}\right).
$$

By the tower property of conditional expectation, it follows that

$$
\begin{aligned}
\mathbb{E}\left[e^{sS_n}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{sS_n}|\mathcal{F}_{n-1}\right]\right]\\
&= \mathbb{E}\left[e^{sS_{n-1}}\mathbb{E}\left[e^{sd_n}|\mathcal{F}_{n-1}\right]\right]\\
&\leq \mathbb{E}\left[e^{sS_{n-1}}\right]\exp\left(\frac{s^2V_2}{2(1-s\,\delta)}\right)\\
&\leq \exp\left(\frac{s^2V_{(n)}}{2(1-s\,\delta)}\right),
\end{aligned}
$$

which gives inequality (6). Using the exponential Markov inequality, we deduce that, for any $x\geq 0$

$$
\begin{aligned}
\mathbb{P}\left(S_n\geq x\right) &\leq \mathbb{E}\left[e^{s\,(S_n-x)}\right]\\
&\leq \exp\left(-s\,x+\frac{s^2V_{(n)}}{2(1-s\,\delta)}\right). \qquad (8)
\end{aligned}
$$

Minimizing the right-hand side with respect to s leads to the result. □

3.3 McDiarmid and Cramer inequalities

Here, we state other consequences of Lemma 3.1. However, as our applications are based on Bernstein inequality, we postpone the proof of these results to Section 5.

When the Laplace transform of the dominating random variables $G_{X_1}(X_1)$ and $G_\varepsilon(\varepsilon_k)$ satisfy the Cramér condition, we obtain the following proposition.

Proposition 3.1. *Assume that there exist some constants $a > 0$, $K_1 \geq 1$ and $K_2 \geq 1$ such that*

$$\mathbb{E}\Big[\exp\Big(aG_{X_1}(X_1)\Big)\Big] \leq K_1$$

and

$$\mathbb{E}\Big[\exp\Big(aG_{\varepsilon}(\varepsilon)\Big)\Big] \leq K_2.$$

Let

$$K = \frac{2}{e^2}\left(K_1 + K_2 \sum_{i=2}^{n} \Big(\frac{K_{n-i}(\rho)}{K_{n-1}(\rho)}\Big)^2\right)$$

and $\delta = a/K_{n-1}(\rho)$. Then, for any $s \in [0, \delta)$,

$$\mathbb{E}\left[e^{\pm s S_n}\right] \leq \exp\left(\frac{s^2 K \delta^{-2}}{1 - s\delta^{-1}}\right).$$

Consequently, for any $x > 0$,

$$\begin{aligned}\mathbb{P}\big(\pm S_n \geq x\big) &\leq \exp\left(\frac{-(x\delta)^2}{2K(1+\sqrt{1+x\delta/K})+x\delta}\right)\\ &\leq \exp\left(\frac{-(x\delta)^2}{4K+2x\delta}\right).\end{aligned}$$

Now, consider the case where the increments d_k are bounded. We shall use an improved version of the well known inequality by McDiarmid, stated by Rio (2013b). For this inequality, we do not assume that (3) holds. Thus, Proposition 3.2 applies to any Markov chain $X_i = F_i(X_{i-1}, \varepsilon_i)$ for F_i satisfying (2). Following Rio (2013b), let

$$\ell(t) = (t - \ln t - 1) + t(e^t - 1)^{-1} + \ln(1 - e^{-t}) \quad \text{for all } t > 0,$$

and let

$$\ell^*(x) = \sup_{t>0}\big(xt - \ell(t)\big) \quad \text{for all } x > 0,$$

be the Young transform of $\ell(t)$. As quoted by Rio (2013b), the following inequality holds

$$\ell^*(x) \geq (x^2 - 2x)\ln(1-x) \quad \text{for any } x \in [0, 1).$$

Let also $(X_1', (\varepsilon_i')_{i\geq 2})$ be an independent copy of $(X_1, (\varepsilon_i)_{i\geq 2})$.

Proposition 3.2. *Assume that there exist some positive constants M_k such that*

$$\big\|d(X_1, X_1')\big\|_\infty \leq M_1$$

and for $k \in [2, n]$,

$$\big\|d\big(F_k(X_{k-1}, \varepsilon_k), F_k(X_{k-1}, \varepsilon_k')\big)\big\|_\infty \leq M_k.$$

Let

$$M^2(n,\rho) = \sum_{k=1}^{n} \big(K_{n-k}(\rho) M_k\big)^2$$

and

$$D(n,\rho) = \sum_{k=1}^{n} K_{n-k}(\rho) M_k \,.$$

Then, for any $s \geq 0$,

$$\mathbb{E}[e^{\pm s S_n}] \;\leq\; \exp\left(\frac{D^2(n,\rho)}{M^2(n,\rho)}\; \ell\Big(\frac{M^2(n,\rho)\, s}{D(n,\rho)}\Big)\right) \tag{9}$$

and, for any $x \in [0, D(n,\rho)]$,

$$\mathbb{P}\big(\pm S_n > x\big) \leq \exp\left(-\frac{D^2(n,\rho)}{M^2(n,\rho)}\; \ell^*\Big(\frac{x}{D(n,\rho)}\Big)\right). \tag{10}$$

Consequently, for any $x \in [0, D(n,\rho)]$,

$$\mathbb{P}\big(\pm S_n > x\big) \;\;\leq\;\; \left(\frac{D(n,\rho) - x}{D(n,\rho)}\right)^{\frac{2D(n,\rho)x - x^2}{M^2(n,\rho)}}. \tag{11}$$

Remark 3.2. Since $(x^2 - 2x)\ln(1-x) \geq 2\,x^2$, $\forall x \in [0,1)$, (11) implies the following McDiarmid inequality

$$\mathbb{P}\big(\pm S_n > x\big) \;\leq\; \exp\left(-\frac{2x^2}{M^2(n,\rho)}\right).$$

Remark 3.3. Taking $\Delta(n,\rho) = K_{n-1}(\rho) \max\limits_{1\leq k\leq n} M_k$, we obtain, for any $x \in [0, n\Delta(n,\rho)]$,

$$\mathbb{P}\big(\pm S_n > x\big) \leq \exp\left(-n\ell^*\Big(\frac{x}{n\Delta(n,\rho)}\Big)\right) \leq \exp\left(-\frac{2x^2}{n\Delta^2(n,\rho)}\right).$$

4 Application to periodic autoregressive models

In this section, we apply Theorem 3.1 to predict a nonstationary Markov chain. We will use periodic autoregressive predictors. Of course, these predictors will work well when the Markov chain is indeed periodic autoregressive. However, we will state the results in a more general context – when the model is wrong, we simply estimate its best prediction by a periodic autoregression.

4.1 Context

Let $(X_t)_{t\geq 1}$ be an $\mathbb{R}^d$-valued process defined by the distribution of X_1 and, for $t > 0$,

$$X_t = f_t^*(X_{t-1}) + \varepsilon_t,$$

where the ε_t are i.i.d and centered, and each f_t^* belong to a fixed family of functions $\mathcal{F}$ with $\forall f \in \mathcal{F}$, $\forall (x,y) \in \mathbb{R}^d$, $\|f(x) - f(y)\| \leq \rho \|x - y\|$, $\rho \in [0,1)$.

We are interested by periodic predictors: $f_{t+T} = f_t$, defined by a sequence $(f_1, \dots, f_T) \in \mathcal{F}^T$. Of course, if the series (X_t) actually satisfies $f^*_{t+T} = f^*_t$, then this family of predictors can give optimal predictions. But they might also perform well when this equality is not exact (for example, when there is a very small drift).

Prediction is assessed with respect to a non-negative loss function: $\ell(\cdot)$. We assume that ℓ is L-Lipschitz. Note that this includes the absolute loss, the Huber loss and all the quantile losses. This also includes the quadratic loss if we assume that X_t, and hence ε_t, is bounded. Given a sample $X_1, \dots, X_n$ we define the empirical risk, for any $f_{1:T} = (f_1, \dots, f_T) \in \mathcal{F}^T$:

$$r_n(f_{1:T}) = \frac{1}{n-1} \sum_{i=2}^{n} \ell\left(X_i - f_{i[T]}(X_{i-1})\right),$$

where $i[T] \in \{1, \dots, T\}$ is such that $i - i[T] \in T \cdot \mathbb{Z}$. We then define

$$R_n(f_{1:T}) = \mathbb{E}[r_n(f_{1:T})].$$

Note that when the process has actually T-periodic distribution, in the sense that the distribution of the vectors $(X_{kT+1}, \dots, X_{(k+1)T})$ are the same for any k, then alsmot surely $f_t^* = f_{t+T}^*$ for any t and

$$R_n(f_{1:T}) \xrightarrow[n\to\infty]{} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\ell\left(X_t - f_t(X_{t-1})\right)]$$

the prediction averaged over one period, which appears to be equal to $R_{T+1}(f_{1:T})$. We can actually give a more accurate statement.

Proposition 4.1. *When the distribution of $(X_{kT+1}, \dots, X_{(k+1)T})$ does not depend on k,*

$$|R_{T+1}(f_{1:T}) - R_n(f_{1:T})| \leq C_0 \frac{2T+1}{n-1},$$

where $C_0 = L(1+\rho)\left[\frac{G_\varepsilon(0)}{1-\rho} + G_{X_1}(0)\right]$.

(All the proofs are postponed to Section 5 for the clarity of exposition). The simplest use of Bernstein's inequality is to control the deviation between $r_n(f_{1:T})$ and $R_n(f_{1:T})$ for a fixed predictor $f_{1:T}$.

Corollary 4.1. *Assume that the moment assumption in Theorem 3.1, given by 5, is satisfied. Define $V_{(n)}$ and δ (depending on M, ρ, V_1 and V_2) as in Theorem 3.1. Then for any $0 \leq s < (n-1)/(L(1+\rho)\delta)$,*

$$\mathbb{E} \exp\left(\pm s(r_n(f_{1:T}) - R_n(f_{1:T}))\right) \leq \exp\left(\frac{s^2(1+\rho)^2 L^2 \frac{V_{(n)}}{n-1}}{2(n-1) - 2s(1+\rho)\delta L}\right).$$

From (7) above, we know that

$$\mathcal{V} := \frac{V_{(n)}}{n-1} \leq \frac{\frac{V_1}{n-1} + V_2}{(1-\rho)^2} \leq \frac{V_1 + V_2}{(1-\rho)^2}$$

that does not depend on n.

4.2 Estimation with a fixed period

In this subsection we assume that T is known (we will later show how to deal with the case were it is unknown). Thus, we define the estimator

$$\hat{f}_{1:T} = (\hat{f}_1, \dots, \hat{f}_T) = \underset{f_{1:T}=(f_1,\dots,f_T)}{\operatorname{argmin}} r_n(f_{1:T}).$$

In order to study the statistical performances of $\hat{f}_{1:T}$, a few definitions are in order. For any function $f : \mathbb{R}^d \to \mathbb{R}^d$ we will use the notation

$$\|f\|_{\sup} := \sup_{x \neq 0} \frac{\|f(x)\|}{\|x\|}.$$

When considering linear functions, this actually coincides with the operator norm.

Definition 4.1. *Define the covering number $\mathcal{N}(\mathcal{F}, \epsilon)$ as the cardinality of the smallest set $\mathcal{F}_\epsilon$ such that $\forall f \in \mathcal{F}$, $\exists f_\epsilon \in \mathcal{F}_\epsilon$ such that $\|f - f^\epsilon\|_{\sup} \leq \epsilon$. Define the entropy of $\mathcal{F}$ by $\mathcal{H}(\mathcal{F}, \epsilon) = 1 \vee \log \mathcal{N}(\mathcal{F}, \epsilon)$.*

Covering numbers are standard tools to measure the complexity of set of predictors in machine learning.

Example 4.1. *Consider the class of AR(1) predictors $f(x) = ax$, $|a| \leq \rho$. Define $\mathcal{F}_\epsilon$ as the set of all functions $f(x) = i\varepsilon x$ for $i \in \{0, \pm 1, \dots, \pm\lfloor \rho/\epsilon \rfloor\}$. It is clear that $\mathcal{F}_\epsilon$ satisfies the above definition and that $\operatorname{card}(\mathcal{F}_\epsilon) \leq 1 + 2\rho/\epsilon \leq 1 + 2/\epsilon$. Thus, $\mathcal{N}(\mathcal{F}, \epsilon) \leq 1 + 2/\epsilon$ and so $\mathcal{H}(\mathcal{F}, \epsilon) \leq 1 \vee \log(1 + 2/\epsilon)$. In the VAR(1) case, $f(x) = Ax$ where $\|A\|_{\sup} \leq \rho$. Using the set $\mathcal{F}_\epsilon$ of all matrices with entries in $(\epsilon/\sqrt{d})\{0, \pm 1, \dots, \pm\lfloor \rho\sqrt{d}/\epsilon \rfloor\}$, we prove that $\mathcal{N}(\mathcal{F}, \epsilon) \leq (1 + 2\sqrt{d}/\epsilon)^d$ and thus $\mathcal{H}(\mathcal{F}, \epsilon) \leq 1 \vee d \log(1 + 2\sqrt{d}/\epsilon)$.*

We are now in position to state the following result on the convergence of $R_n(\hat{f}_{1:T})$.

Theorem 4.1. *As soon as $n \geq 1 + 4\delta^2 T \mathcal{H}(\mathcal{F}, \frac{1}{Ln})/\mathcal{V}$ we have, for any $\eta > 0$,*

$$\mathbb{P}\Bigg\{ R_n(\hat{f}_{1:T}) \leq \inf_{f_{1:T} \in \mathcal{F}^T} R_n(f_{1:T}) + C_1 \sqrt{\frac{T\mathcal{H}(\mathcal{F}, \frac{1}{Ln})}{n-1}} + C_2 \frac{\log\left(\frac{4}{\eta}\right)}{\sqrt{n-1}} + \frac{C_3}{n} \Bigg\} \geq 1 - \eta,$$

where $C_1 = 4(1+\rho)L\sqrt{\mathcal{V}}$, $C_2 = 2(1+\rho)L\sqrt{\mathcal{V}} + 2\delta$ *and* $C_3 = 3[G_\varepsilon(0) + G_{X_1}(0)]/(1-\rho) + \mathcal{V}/(2\delta)$.

The theorem states that the predictor $\hat{f}_{1:T}$ predict as well as the best possible one up to an estimation error that vanishes at rate $\sqrt{n}$. For example, using (periodic) VAR(1) predictors in dimension d, we get a bound in

$$R_n(\hat{f}_{1:T}) \leq \inf_{f_{1:T}\in\mathcal{F}^T} R_n(f_{1:T}) + \mathcal{O}\left(d\sqrt{\frac{T\log(nd)}{n}}\right).$$

Remark 4.2. When the series is indeed stationary for a known T, it is to be noted that $(X_{iT+1},\dots,X_{i(T+1)})_{i\geq 0}$ is a time homogeneous Markov chain. In this case, our technique is not really necessary: it would be possible to apply the inequality from Dedecker and Fan (2015). However, when T is not known, this becomes impossible. In this case, one has to compare the empirical risks of $\hat{f}_{1:T}$ for the various possible T's, and for most of them, $(X_{iT+1},\dots,X_{i(T+1)})_{i\geq 0}$ is not homogeneous. In this case, vectorization cannot help. On the other hand, our inequality can be used for period selection, as detailed in the next subsection.

4.3 Period and model selection

We define a penalized estimator in the spirit of Massart (2007). Fix a maximal period $T_{\max}$, for example $T_{\max} = \lfloor n/2 \rfloor$. We propose the following penalized estimator for T:

$$\hat{T} = \arg\min_{1\leq T\leq T_{\max}} \left[r_n(\hat{f}_{1:T}) + \frac{C_1}{2}\sqrt{\frac{T\mathcal{H}(\mathcal{F},\frac{1}{Ln})}{n-1}} \right].$$

Using this estimator, we have the following result.

Theorem 4.3. *For any* $\eta > 0$ *we have,*

$$\mathbb{P}\Bigg\{ R_n(\hat{f}_{1:\hat{T}}) \leq \inf_{1\leq T\leq T_{\max}} \inf_{f_{1:T}\in\mathcal{F}^T} \Bigg[R_n(f_{1:T}) + C_1\sqrt{\frac{T\mathcal{H}(\mathcal{F},\frac{1}{Ln})}{n-1}} + C_2\frac{\log\left(\frac{4T_{\max}}{\eta}\right)}{\sqrt{n-1}} + \frac{C_3}{n} \Bigg] \Bigg\} \geq 1-\eta,$$

as soon as $n \geq 1 + 4\delta^2 T_{\max}\mathcal{H}(\mathcal{F},\frac{1}{Ln})/\mathcal{V}$.

Note that $\hat{T}$ depends on $C_1 = 4(1+\rho)L\sqrt{\mathcal{V}} \leq 4(1+\rho)L\sqrt{V_1+V_2}/(1-\rho)$. While L depends only on the loss that is chosen by the statistician, in many applications ρ, V_1 and V_2 are unknown. We recommend to use an empirical criterion like the slope heuristic, introduced by Birgé and Massart (2006), to calibrate C_1. This procedure is as follows:

1. define, for any $c > 0$, $\hat{T}(c) = \arg\min_{1 \leq T \leq T_{\max}} [r_n(\hat{f}_{1:T}) + c\sqrt{T}]$.
2. fix a small step $\epsilon > 0$ and define $\hat{c}$ as the maximiser of the jump $J_\epsilon(c) = \sqrt{\hat{T}(c+\epsilon)} - \sqrt{\hat{T}(c)}$.
3. select $\hat{T}(2\hat{c})$.

Many variants, details on fast implementations and references for theoretical results (in the i.i.d case) can be found in see Baudry et al. (2012). A theoretical study of the slope heuristic in the context could be the object of future works.

4.4 Simulation study

As an illustration we simulate $X_{t+1} = a_t X_t + \varepsilon_t$ for $t = 1, \ldots, 400$, where $a_{t+4} = a_t$, $(a_1, a_2, a_3, a_4) = (0.8, 0.5, 0.9, -0.7)$ and $\varepsilon_t \sim \mathcal{N}(0, 1)$. The data is shown in Figure 1 and the autocorrelation function in Figure 2. It is clear that a statistician trying to estimate an AR(1) model with a fixed coefficient would be puzzled by this situation.

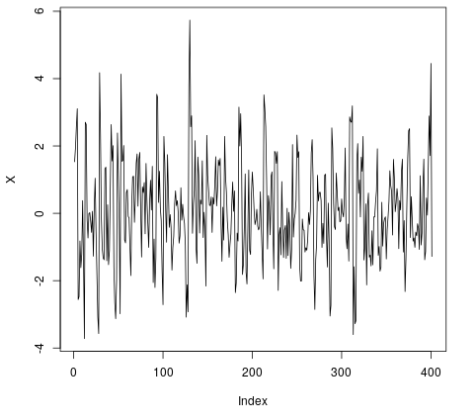

Figure 1: Simulated data.

The dependence of $r_n(\hat{a}_{1:T})$ with respect to $T \in \{1, ..., T_{\max}\}$ with $T_{\max} = 20$ is shown in Figure 3. The choice $T = 4$ leads to an improvement with respect to $T < 4$. On the other hand, we observe a slow linear decrease of $r_n(\hat{a}_{1:4})$, $r_n(\hat{a}_{1:8})$, $r_n(\hat{a}_{1:12})$ ... this is a sign of overfitting. And indeed,

1. for $c < 0.008$, $\hat{T}(c) = 20$,
2. for $0.009 < c < 0.239$, $\hat{T}(c) = 4$,
3. for $0.240 < c$, $\hat{T}(c) = 1$.

Thus, $\hat{c} \simeq 0.0085$ and we choose $\hat{T}(2\hat{c}) = \hat{T}(0.017) = 4$.

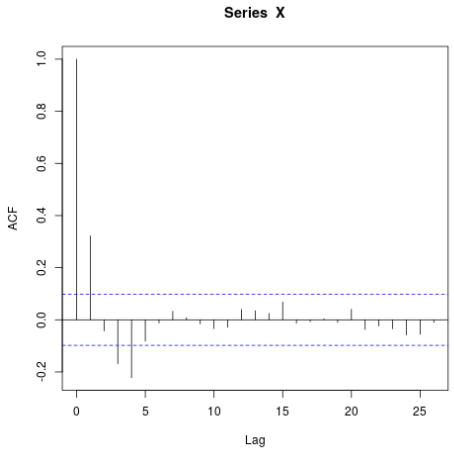

Figure 2: Autocorrelation function of the data.

5 Proofs

Proof of Lemma 3.1. The first point will be proved by backward induction. The result is obvious for $t = n$, since $g_n = f$. Assume that it is true at step t, and let us prove it at step $t-1$. By definition

$$g_{t-1}(X_1, \dots, X_{t-1}) = \mathbb{E}[g_t(X_1, \dots, X_t)|\mathcal{F}_{t-1}] \\ = \int g_t(X_t, \dots, X_{t-1}, F_t(X_{t-1}, y))P_\varepsilon(dy)\,.$$

It follows that

$$|g_{t-1}(x_1, \dots, x_{t-1}) - g_{t-1}(x'_1, \dots, x'_{t-1})| \\ \leq \int |g_t(x_1, \dots, x_{t-1}, F_t(x_{t-1}, y)) \\ - g_t(x'_1, \dots, x'_{t-1}, F_t(x'_{t-1}, y))|P_\varepsilon(dy)\,. \quad (12)$$

Now, by assumption and condition (2),

$$\int |g_t(x_1, \dots, F_t(x_{t-1}, y)) - g_t(x'_1, \dots, F_t(x'_{t-1}, y))|P_\varepsilon(dy) \\ \leq d(x_1, x'_1) + \cdots + d(x_{t-1}, x'_{t-1}) \\ + K_{n-t}(\rho) \int d(F_t(x_{t-1}, y), F_t(x'_{t-1}, y))P_\varepsilon(dy) \\ \leq d(x_1, x'_1) + \cdots + (1 + \rho K_{n-t}(\rho))d(x_{t-1}, x'_{t-1}) \\ \leq d(x_1, x'_1) + \cdots + K_{n-t+1}(\rho)d(x_{t-1}, x'_{t-1})\,.$$

Point 1 follows from this last inequality and (12).

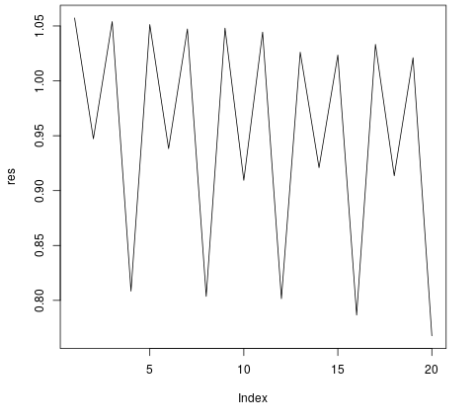

Figure 3: The empirical risk as a function of T.

Let us now prove Point 2. First note that

$$\begin{aligned} |d_1| &= \left| g_1(X_1) - \int g_1(x) P_{X_1}(dx) \right| \\ &\leq K_{n-1}(\rho) \int d(X_1, x) P_{X_1}(dx) \\ &= K_{n-1}(\rho) G_{X_1}(X_1) \end{aligned} \tag{13}$$

where the inequality comes from the first point of Lemma 3.1. In the same way, for $t \geq 2$,

$$\begin{aligned} |d_t| &= \left| g_t(X_1, \cdots, X_t) - \mathbb{E}[g_t(X_1, \cdots, X_t) | \mathcal{F}_{t-1}] \right| \\ &\leq \int \big| g_t(X_1, \cdots, F_t(X_{t-1}, \varepsilon_t)) \\ &\qquad - g_t(X_1, \cdots, F_t(X_{t-1}, y)) \big| P_\varepsilon(dy) \\ &\leq K_{n-t}(\rho) \int d(F_t(X_{t-1}, \varepsilon_t), F_t(X_{t-1}, y)) P_\varepsilon(dy) \\ &= K_{n-t}(\rho) H_{t,\varepsilon}(X_{t-1}, \varepsilon_t) \,. \end{aligned}$$

Finally, the proof of Point 3 is direct: if (3) is true, then

$$H_{t,\varepsilon}(x, y) = \int d(F_t(x, y), F_t(x, y')) P_\varepsilon(dy') \leq \int C\delta(y, y') P_\varepsilon(dy') = G_\varepsilon(y) \,.$$

The proof of the proposition is now complete. □

We state a lemma that will be used in the following proofs.

Lemma 5.1. *Under the assumptions of Section 4 we have*

$$\forall n \in \mathbb{N} \setminus \{0\}, \, \mathbb{E}\|X_n\| \leq \frac{G_\varepsilon(0)}{1 - \rho} + \rho^{n-1} G_{X_1}(0).$$

Proof of Lemma 5.1 By definition of G_{X_1}, $\mathbb{E}\|X_1\| = \int \|x-0\| \mathrm{d}P_{X_1}(x) = G_{X_1}(0)$. Then,

$$\mathbb{E}\|X_n\| = \mathbb{E}\|f_n(X_{n-1}) + \varepsilon\| \leq \mathbb{E}\|f_n(X_{n-1})\| + \mathbb{E}\|\varepsilon\| \leq \rho\mathbb{E}\|X_{n-1}\| + G_\varepsilon(0).$$

So, by induction, for $n > 1$,

$$\begin{aligned}\mathbb{E}\|X_n\| &\leq (1+\rho+\cdots+\rho^{n-2})G_\varepsilon(0) + \rho^{n-1}G_{X_1}(0) \\ &\leq \frac{G_\varepsilon(0)}{1-\rho} + \rho^{n-1}G_{X_1}(0). \qquad \square\end{aligned}$$

Proof of Proposition 3.1. Let $\delta = a/K_{n-1}(\rho)$. Since $\mathbb{E}[d_1] = 0$, it follows that, for any $s \in [0,\delta)$,

$$\begin{aligned}\mathbb{E}[e^{sd_1}] &= 1 + \sum_{i=2}^{\infty} \frac{s^i}{i!}\mathbb{E}[(d_1)^i] \\ &\leq 1 + \sum_{i=2}^{\infty}\Big(\frac{s}{\delta}\Big)^i \mathbb{E}\Big[\frac{1}{i!}|\delta d_1|^i\Big]. \end{aligned} \tag{14}$$

Note that, for $s \geq 0$,

$$\frac{s^i}{i!}e^{-s} \leq \frac{i^i}{i!}e^{-i} \leq 2e^{-2}, \quad \text{for any } i \geq 2, \tag{15}$$

where the last inequality follows from the fact that $i^i e^{-i}/i!$ is decreasing in i. Notice that the equality in (15) is reached at $s = i = 2$. By (15), Lemma 3.1 and the hypothesis of the proposition, we deduce that

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{i!}|\delta d_1|^i\Big] &\leq 2e^{-2}\mathbb{E}[e^{\delta|d_1|}] \\ &\leq 2e^{-2}\mathbb{E}\Big[\exp\Big(aG_{X_1}(X_1)\Big)\Big] \\ &\leq 2e^{-2}K_1. \end{aligned} \tag{16}$$

Combining Inequalities (14) and (16), we get, for any $s \in [0,\delta)$,

$$\begin{aligned}\mathbb{E}[e^{sd_1}] \;\leq\; 1 + \sum_{n=2}^{\infty}\frac{2}{e^2}\Big(\frac{s}{\delta}\Big)^n K_1 \;=\; 1 + \frac{2}{e^2}\frac{s^2K_1\delta^{-2}}{1-s\delta^{-1}} \\ \leq\; \exp\left(\frac{2}{e^2}\frac{s^2K_1\delta^{-2}}{1-s\delta^{-1}}\right).\end{aligned}$$

Similarly, since $K_{n-i}(\rho)/K_{n-1}(\rho) \leq 1$ for any $i \in [2,n]$, we obtain, for any $s \in [0,\delta)$,

$$\mathbb{E}[e^{sd_i}|\mathcal{F}_{i-1}] \;\leq\; \exp\left(\frac{2}{e^2}\frac{s^2K_2\delta^{-2}}{1-s\delta^{-1}}\Big(\frac{K_{n-i}(\rho)}{K_{n-1}(\rho)}\Big)^2\right).$$

Using the tower property of conditional expectation, we have, for any $s \in [0, \delta)$,

$$\begin{aligned}\mathbb{E}\left[e^{sS_n}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{sS_n}|\mathcal{F}_{n-1}\right]\right] \\ &= \mathbb{E}\left[e^{sS_{n-1}}\mathbb{E}\left[e^{sd_n}|\mathcal{F}_{n-1}\right]\right] \\ &\leq \mathbb{E}\left[e^{sS_{n-1}}\right]\exp\left(\frac{2}{e^2}\frac{t^2K_2\delta^{-2}}{1-t\delta^{-1}}\left(\frac{K_{n-i}(\rho)}{K_{n-1}(\rho)}\right)^2\right).\end{aligned}$$

By recursion,

$$\begin{aligned}\mathbb{E}\left[e^{sS_n}\right] &\leq \mathbb{E}\left[e^{sS_1}\right]\exp\left(\frac{2}{e^2}\frac{t^2K_2\delta^{-2}}{1-t\delta^{-1}}\sum_{i=2}^{n}\left(\frac{K_{n-i}(\rho)}{K_{n-1}(\rho)}\right)^2\right) \\ &\leq \exp\left(\frac{2}{e^2}\frac{s^2K_1\delta^{-2}}{1-s\delta^{-1}}\right)\exp\left(\frac{2}{e^2}\frac{t^2K_2\delta^{-2}}{1-t\delta^{-1}}\sum_{i=2}^{n}\left(\frac{K_{n-i}(\rho)}{K_{n-1}(\rho)}\right)^2\right) \\ &= \exp\left(\frac{s^2K\delta^{-2}}{1-s\delta^{-1}}\right),\end{aligned}$$

where

$$K = \frac{2}{e^2}\left(K_1 + K_2\sum_{i=2}^{n}\left(\frac{K_{n-i}(\rho)}{K_{n-1}(\rho)}\right)^2\right).$$

Then using the exponential Markov inequality, we deduce that, for any $x \geq 0$ and $s \in [0, \delta)$,

$$\begin{aligned}\mathbb{P}\left(S_n \geq x\right) &\leq \mathbb{E}\left[e^{s\left(S_n - x\right)}\right] \\ &\leq \exp\left(-sx + \frac{s^2K\delta^{-2}}{1-s\delta^{-1}}\right).\end{aligned}$$

The minimum is reached at

$$s = s(x) := \frac{x\delta^2/K}{x\delta/K + 1 + \sqrt{1 + x\delta/K}}.$$

The proposition is proven. □

Proof of Proposition 3.2. Denote

$$u_{k-1}(x_1, \ldots, x_{k-1}) = \operatorname{ess\,inf}_{\varepsilon_k} g_k(x_1, \ldots, F_k(x_{k-1}, \varepsilon_k))$$

and

$$v_{k-1}(x_1, \ldots, x_{k-1}) = \operatorname{ess\,sup}_{\varepsilon_k} g_k(x_1, \ldots, F_k(x_{k-1}, \varepsilon_k)).$$

From the proof of Lemma 3.1, it is easy to see that

$$u_{k-1}(X_1, \ldots, X_{k-1}) \leq d_k \leq v_{k-1}(X_1, \ldots, X_{k-1}).$$

By Lemma 3.1 and the hypothesis of the proposition, it follows that

$$v_{k-1}(X_1,\dots,X_{k-1}) - u_{k-1}(X_1,\dots,X_{k-1}) \leq K_{n-k}(\rho)M_k\,.$$

Following exactly the proof of Theorem 3.1 of Rio (2013b) with $\Delta_k = K_{n-k}(\rho)M_k$, we get (9) and (10). Since $\ell^*(x) \geq (x^2-2x)\ln(1-x)$ for any $x \in [0,1]$, (11) follows from (10). □

Proof of Proposition 4.1. Put $k = \lfloor n/T \rfloor$, then

$$\begin{aligned} R_n(f_{1:T}) &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=2}^{n} \ell\left(X_i - f_{i[T]}(X_{i-1})\right)\right] \\ &= \frac{kTR_{T+1}(f_{1:T})}{n} + \mathbb{E}\left[\frac{1}{n-1}\sum_{i=kT}^{n} \ell\left(X_i - f_{i[T]}(X_{i-1})\right)\right]. \end{aligned}$$

First,

$$\begin{aligned} \left|R_n(f_{1:T}) - \frac{kTR_{T+1}(f_{1:T})}{n}\right| &\leq \frac{1}{n-1}\sum_{i=kT}^{(k+1)T-1} \mathbb{E}\left[\left|\ell\left(X_i - f_{i[T]}(X_{i-1})\right)\right|\right] \\ &\leq \frac{1}{n-1}\sum_{i=kT}^{(k+1)T-1} L\mathbb{E}\|X_i\| + \rho L\mathbb{E}\|X_{i-1}\| \\ &\leq \frac{T}{n-1}L(1+\rho)\left[\frac{G_\varepsilon(0)}{1-\rho} + G_{X_1}(0)\right] \end{aligned}$$

where we used Lemma 5.1 and $\rho^{n-1} < 1$ for the last inequality. In the same way,

$$\begin{aligned} \left|\frac{kTR_T(f_{1:T})}{n} - R_{T+1}(f_{1:T})\right| &= \left|\frac{kT}{n-1} - 1\right| R_{T+1}(f_{1:T}) \\ &= \frac{|kT-n+1|}{n-1}\frac{1}{T}\sum_{i=1}^{T}\mathbb{E}\left[\left|\ell\left(X_i - f_i(X_{i-1})\right)\right|\right] \\ &= \frac{T+1}{n-1}(L(1+\rho)\left[\frac{G_\varepsilon(0)}{1-\rho} + G_{X_1}(0)\right]). \end{aligned}$$

Combining both inequalities leads to the result. □

Proof of Corollary 4.1. By definition, we have $X_t = F_t(X_{t-1},\varepsilon_t) = f_t^*(X_{t-1}) + \varepsilon_t$. So $\|F_t(x,\varepsilon_t) - F_t(x',\varepsilon_t)\| = \|f_t^*(x) - f_t^*(x')\| \leq \rho\|x-x'\|$ and so (2) is satisfied, and $\|F_t(x,y) - F_t(x,y')\| = \|y-y'\|$ so that (3) is satisfied with $C=1$. We consider the random variable $S_n = f(X_1,\dots,X_n) - \mathbb{E}[f(X_1,\dots,X_n)]$, where

$$f(x_1,\dots,x_n) = \frac{1}{L(\rho+1)}\sum_{t=2}^{n}\ell\left(x_t - f_{t[T]}(x_{t-1})\right).$$

Remark that

$$
\begin{aligned}
&|f(x_1,\dots,x_n) - f(x_1,\dots,x'_t,\dots,x_n)| \\
&\leq \frac{\left|\ell\left(x_{t+1} - f_{(t+1)[T]}(x_t)\right) - \ell\left(x_{t+1} - f_{(t+1)[T]}(x'_t)\right)\right|}{L(\rho+1)} \\
&\qquad + \frac{\left|\ell\left(x_t - f_{t[T]}(x_{t-1})\right) - \ell\left(x'_t - f_{t[T]}(x_{t-1})\right)\right|}{L(\rho+1)} \\
&\leq \frac{\|f_{(t+1)[T]}(x_t) - f_{(t+1)[T]}(x'_t)\| + \|x_t - x'_t\|}{\rho+1} \\
&\leq \|x_t - x'_t\|.
\end{aligned}
$$

So the assumptions of Theorem 3.1 are satisfied and

$$
\mathbb{E}\exp\left(\pm t S_n\right) \leq \exp\left(\frac{t^2 V_{(n)}}{2 - 2t\delta}\right).
$$

Remind that $S_n = \frac{n-1}{L(1+\rho)}\left[r_n(f_{1:T}) - \mathbb{E}[r_n(f_{1:T})]\right]$, and $R_n(f_{1:T}) = \mathbb{E}[r_n(f_{1:T})]$, so that by setting $s = t(n-1)/L(1+\rho)$ we end the proof. □

Proof of Theorem 4.1. Fix $\epsilon > 0$. We have, for any $f_{1:T} \in \mathcal{F}_\epsilon$, the deviation inequality from Corollary 4.1. A union bound on $f_{1:T} \in \mathcal{F}_\epsilon$ leads to, for any $s \in \left[0, \frac{n-1}{L(1+\rho)\delta}\right)$,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{(f_{1:T})\in\mathcal{F}_\epsilon^T} |r_n(f_{1:T}) - R_n(f_{1:T})| > x\right) \\
&\leq \sum_{(f_{1:T})\in\mathcal{F}_\epsilon} \mathbb{P}\left(|r_n(f_{1:T}) - R_n(f_{1:T})| > x\right) \\
&\leq \sum_{(f_{1:T})\in\mathcal{F}_\epsilon^T} \mathbb{E}\exp\left(s|r_n(f_{1:T}) - R_n(f_{1:T})| - sx\right) \\
&\leq 2\mathcal{N}(\mathcal{F},\epsilon)^T \exp\left(\frac{s^2(1+\rho)^2 L^2 \mathcal{V}}{2(n-1) - 2s(1+\rho)\delta L} - sx\right).
\end{aligned}
$$

Now, for any $f_{1:T} \in \mathcal{F}^T$ we construct $f^\epsilon_{1:T} = (f^\epsilon_1,\dots,f^\epsilon_T)$ by chosing, for any $t \in \{1,\dots,T\}$, a function f^ϵ_t such that $\|f_t - f^\epsilon_t\|_{\sup} \leq \epsilon$, as allowed from the definition of $\mathcal{F}_\epsilon$. Obviously

$$
\begin{aligned}
&\left|\ell(X_t - f^\epsilon_{t[T]}(X_{t-1})) - \ell(X_t - f_{t[T]}(X_{t-1}))\right| \\
&\qquad\qquad \leq L\|f^\epsilon_{t[T]}(X_{t-1}) - f_{t[T]}(X_{t-1})\| \leq L\epsilon\|X_{t-1}\|
\end{aligned}
$$

and as a consequence,

$$
|r_n(f_{1:T}) - r_n(f^\epsilon_{1:T})| \leq L\epsilon \cdot \frac{\sum_{t=1}^{n-1}\|X_t\|}{n-1},
$$

and

$$|R_n(f_{1:T}) - R_n(f^\epsilon_{1:T})| \leq \epsilon L \cdot \frac{\sum_{t=1}^{n-1} \mathbb{E}\|X_t\|}{n-1}.$$

Using Theorem 3.1 with $f(X_1, \ldots, X_n) = \sum_{t=1}^{n-1} \|X_t\|$ we have, for any $y > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{t=1}^{n-1} \|X_t\| > \sum_{t=1}^{n-1} \mathbb{E}\|X_t\| + y\right) \\
&\leq \mathbb{E} \exp\left[\frac{1}{2\delta}\left(\sum_{t=1}^{n-1} \|X_t\| - \sum_{t=1}^{n-1} \mathbb{E}\|X_t\| - y\right)\right] \\
&\leq \exp\left(\frac{\left(\frac{1}{2\delta}\right)^2 V_{(n)}}{2\left(1-\frac{1}{2}\right)} - \frac{y}{2\delta}\right) = \exp\left(\frac{V_{(n)}}{4\delta^2} - \frac{y}{2\delta}\right).
\end{aligned}$$

Lemma 5.1 leads to

$$\begin{aligned}
\sum_{t=1}^{n-1} \mathbb{E}\|X_t\| &\leq \sum_{t=1}^{n-1} \left[\frac{G_\epsilon(0)}{1-\rho} + \rho^{t-1} G_{X_1}(0)\right] \\
&\leq \frac{(n-1)G_\epsilon(0) + G_{X_1}(0)}{1-\rho} =: z_{\rho,n}
\end{aligned}$$

where we introduce the last notation for short. Now let us consider the "favorable" event

$$\mathcal{E} = \left\{\sum_{t=1}^{n-1} \|X_t\| \leq z_{\rho,n} + y\right\} \bigcap \left\{\sup_{f_{1:T} \in \mathcal{F}_\epsilon} |r_n(f_{1:T}) - R_n(f_{1:T})| \leq x\right\}.$$

The previous inequalities show that

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}^c\right) \leq \exp&\left(\frac{V_{(n)}}{4\delta^2} - \frac{y}{2\delta}\right) \\
&+ 2\mathcal{N}(\mathcal{F}, \epsilon)^T \exp\left(\frac{s^2(1+\rho)^2 L^2 \mathcal{V}}{2(n-1) - 2s(1+\rho)\delta L} - sx\right). \quad (17)
\end{aligned}$$

On $\mathcal{E}$, we have:

$$\begin{aligned}
R_n(\hat{f}_{1:T}) &\leq R_n(\hat{f}^\epsilon_{1:T}) + \epsilon L \frac{z_{\rho,n}}{n-1} \\
&\leq r_n(\hat{f}^\epsilon_{1:T}) + x + \epsilon L \frac{z_{\rho,n}}{n-1} \\
&\leq r_n(\hat{f}_{1:T}) + x + \epsilon L \left[2\frac{z_{\rho,n}}{n-1} + \frac{y}{n-1}\right] \\
&= \min_{f_{1:T} \in \mathcal{F}^T_\epsilon} r_n(f_{1:T}) + x + \epsilon L \frac{2z_{\rho,n} + y}{n-1} \\
&\leq \min_{f_{1:T} \in \mathcal{F}^T_\epsilon} R_n(f_{1:T}) + 2x + \epsilon L \frac{2z_{\rho,n} + y}{n-1}
\end{aligned}$$

$$\leq \min_{f_{1:T}\in\mathcal{F}^T} R_n(f_{1:T}) + 2x + \epsilon L\frac{3z_{\rho,n}+y}{n-1}.$$

In particular, the choice $\epsilon = 1/(Ln)$ ensures:

$$R_n(\hat{f}_{1:T}) \leq \min_{f_{1:T}\in\mathcal{F}^T} R_n(f_{1:T}) + 2x + \frac{3z_{\rho,n}+y}{n(n-1)}. \tag{18}$$

Fix $\eta > 0$ and put:

$$x = \frac{s(1+\rho)^2L^2\mathcal{V}}{2(n-1)-2s(1+\rho)\delta L} + \frac{T\mathcal{H}(\mathcal{F},\frac{1}{Ln}) + \log\left(\frac{4}{\eta}\right)}{s}$$

and $y = 2\delta\log\left(\frac{2}{\eta}\right) + \frac{V_{(n)}}{2\delta}$. Note that, plugged into (17), these choices ensure $\mathbb{P}(\mathcal{E}^c) \leq \eta/2 + \eta/2 = \eta$. Put

$$s = \frac{1}{(1+\rho)L}\sqrt{(n-1)T\mathcal{H}(\mathcal{F},\frac{1}{Ln})\Big/\mathcal{V}}.$$

As soon as $2s(1+\rho)\delta L \leq n-1$, that is actually ensured by the condition $n \geq 1 + 4\delta^2 T\mathcal{H}(\mathcal{F},\frac{1}{Ln})/\mathcal{V}$, we have:

$$\begin{aligned} x &\leq \frac{s(1+\rho)^2L^2\mathcal{V}}{n-1} + \frac{T\mathcal{H}(\mathcal{F},\frac{1}{Ln}) + \log\left(\frac{4}{\eta}\right)}{s} \\ &= 2(1+\rho)L\sqrt{\frac{\mathcal{V}T\mathcal{H}(\mathcal{F},\frac{1}{Ln})}{n-1}} \\ &\quad + (1+\rho)L\log\left(\frac{4}{\eta}\right)\sqrt{\frac{\mathcal{V}}{(n-1)T\mathcal{H}(\mathcal{F},\frac{1}{Ln})}}. \end{aligned}$$

Pluging the expressions of x and y and the definition of $z_{\rho,n}$ into (18) gives:

$$\begin{aligned} R_n(\hat{f}_{1:T}) \leq \min_{f_{1:T}\in\mathcal{F}^T} R_n(f_{1:T}) &+ 4(1+\rho)L\sqrt{\frac{\mathcal{V}T\mathcal{H}(\mathcal{F},\frac{1}{Ln})}{n-1}} \\ &+ 2(1+\rho)L\log\left(\frac{4}{\eta}\right)\sqrt{\frac{\mathcal{V}}{(n-1)T\mathcal{H}(\mathcal{F},\frac{1}{Ln})}} \\ &+ \frac{1}{n}\left[3\frac{G_\epsilon(0)+\frac{G_{X_1}(0)}{n-1}}{1-\rho} + \frac{2\delta\log\left(\frac{2}{\eta}\right)+\frac{V_{(n)}}{2\delta}}{n-1}\right] \end{aligned}$$

which ends the proof. □

Proof of Theorem 4.3. Fix $\epsilon > 0$. For any $1 \leq T \leq T_{\max}$ and $f_{1:T} = (f_1,\ldots,f_T) \in \mathcal{F}^T$, chose $f^\epsilon_{1:T} = (f^\epsilon_1,\ldots,f^\epsilon_T)$ such that for any i, $\|f^\epsilon_i - f_i\|_{\sup} \leq \epsilon$. Define the event

$$\mathcal{A} = \left\{\sum_{t=1}^{n-1}\|X_t\| \leq z_{\rho,n}+y\right\} \bigcap \bigcap_{T=1}^{T_{\max}} \left\{\sup_{f_{1:T}\in\mathcal{F}_\epsilon} |r_n(f_{1:T}) - R_n(f_{1:T})| \leq x_T\right\}$$

where $z_{\rho,n}$ is defined as in the proof of Theorem 4.1 and $x_1, \ldots, x_{T_{\max}} > 0$. We have, for any $s_1, \ldots, s_{T_{\max}} < (n-1)/[L\delta(1+\rho)]$,

$$\mathbb{P}\left(\mathcal{A}^c\right) \leq \exp\left(\frac{V_{(n)}}{4\delta^2} - \frac{y}{2\delta}\right) + 2\sum_{T=1}^{T_{\max}} \mathcal{N}(\mathcal{F}, \epsilon)^T \exp\left(\frac{s_T^2(1+\rho)^2 L^2 \mathcal{V}}{2(n-1) - 2s_T(1+\rho)\delta L} - s_T x_T\right) \leq \eta,$$

the last inequality being ensured by the choice $y = 2\delta\log(2/\eta) + V_{(n)}/(2\delta)$ and, for any T,

$$x_T = \frac{s_T(1+\rho)^2 L^2 \mathcal{V}}{2(n-1) - 2s_T(1+\rho)\delta L} + \frac{T\mathcal{H}(\mathcal{F}, \frac{1}{Ln}) + \log\left(4T_{\max}/\eta\right)}{s_T},$$

$$s_T = \frac{1}{(1+\rho)L}\sqrt{\frac{(n-1)T\mathcal{H}(\mathcal{F}, \frac{1}{Ln})}{\mathcal{V}}}.$$

Note that this choice also leads to

$$x_T \leq \frac{C_1}{2}\sqrt{\frac{T\mathcal{H}(\mathcal{F}, \frac{1}{Ln})}{n-1}} + \frac{C_2}{2}\frac{\log\left(4T_{\max}/\eta\right)}{\sqrt{n-1}}.$$

On $\mathcal{A}$, $R_n(\hat{f}_{1:\hat{T}}) \leq R_n(\hat{f}^{\epsilon}_{1:\hat{T}}) + \epsilon L z_{\rho,n}/(n-1)$, and

$$\begin{aligned}
R_n(\hat{f}_{1:\hat{T}}) &\leq r_n(\hat{f}^{\epsilon}_{1:\hat{T}}) + x_{\hat{T}} + \epsilon L\frac{z_{\rho,n}}{n-1} \\
&\leq r_n(\hat{f}_{1:\hat{T}}) + x_{\hat{T}} + \epsilon L\frac{2z_{\rho,n}+y}{n-1} \\
&\leq r_n(\hat{f}_{1:\hat{T}}) + \frac{C_1}{2}\sqrt{\frac{\hat{T}\mathcal{H}(\mathcal{F}, \frac{1}{Ln})}{n-1}} + \frac{C_2}{2}\frac{\log\left(4T_{\max}/\eta\right)}{\sqrt{n-1}} + \epsilon L\frac{2z_{\rho,n}+y}{n-1} \\
&= \min_{1\leq T\leq T_{\max}}\left\{r_n(\hat{f}_{1:T}) + \frac{C_1}{2}\sqrt{\frac{T\mathcal{H}(\mathcal{F}, \frac{1}{Ln})}{n-1}}\right\} + \frac{C_2}{2}\frac{\log\left(4T_{\max}/\eta\right)}{\sqrt{n-1}} \\
&\quad + \epsilon L\frac{2z_{\rho,n}+y}{n-1} \\
&\leq \min_{1\leq T\leq T_{\max}}\min_{f_{1:T}\in\mathcal{F}_\epsilon^T}\left\{r_n(f_{1:T}) + \frac{C_1}{2}\sqrt{\frac{T\mathcal{H}(\mathcal{F}, \frac{1}{Ln})}{n-1}}\right\} + \frac{C_2}{2}\frac{\log\left(4T_{\max}/\eta\right)}{\sqrt{n-1}} \\
&\quad + \epsilon L\frac{2z_{\rho,n}+y}{n-1} \\
&\leq \min_{1\leq T\leq T_{\max}}\min_{f_{1:T}\in\mathcal{F}_\epsilon^T}\left\{R_n(f_{1:T}) + C_1\sqrt{\frac{T\mathcal{H}(\mathcal{F}, \frac{1}{Ln})}{n-1}}\right\} + \frac{C_2}{2}\frac{\log\left(4T_{\max}/\eta\right)}{\sqrt{n-1}}
\end{aligned}$$

$$
\begin{aligned}
&+ \epsilon L \frac{2z_{\rho,n} + y}{n-1} \\
&\leq \min_{1 \leq T \leq T_{\max}} \min_{f_{1:T} \in \mathcal{F}^T} \left\{ R_n(f_{1:T}) + C_1 \sqrt{\frac{T\mathcal{H}(\mathcal{F}, \frac{1}{Ln})}{n-1}} \right\} + \frac{C_2}{2} \frac{\log\left(4T_{\max}/\eta\right)}{\sqrt{n-1}} \\
&+ \epsilon L \frac{3z_{\rho,n} + y}{n-1} \\
&\leq \inf_{1 \leq T \leq T_{\max}} \inf_{f_{1:T} \in \mathcal{F}^T} \left[R_n(f_{1:T}) + C_1 \sqrt{\frac{T\mathcal{H}(\mathcal{F}, \frac{1}{Ln})}{n-1}} + C_2 \frac{\log\left(4T_{\max}/\eta\right)}{\sqrt{n-1}} \right. \\
&\left. + \frac{C_3}{n} \right]. \qquad \square
\end{aligned}
$$

References

Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13, 1000–1034.

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

Alquier, P. and Li, X. (2012). Prediction of quantiles by statistical learning and application to GDP forecasting. In *International Conference on Discovery Science*, pages 22–36. Springer.

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

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

Bardet, J.-M. and Doukhan, P. (2018). Non-parametric estimation of time varying $AR(1)$ processes with local stationarity and periodicity. *Electronic Journal of Statistics*, 12(2):2323–2354.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470.

Bercu, B., Delyon, B., and Rio, E. (2015). *Concentration inequalities for sums and martingales*. Springer.

Bertail, P. and Ciolek, G. (2019). *New Bernstein and Hoeffding type inequalities for regenerative Markov chains. ALEA, Latin American Journal of Probability and Mathematical Statistics*, to appear.

Bertail, P. and Clémençon, S. (2010). Sharp bounds for the tails of functionals of Markov chains. *Theory of Probability & Its Applications*, 54(3):505–515.

Bertail, P. and Portier, F. (2018). Rademacher complexity for Markov chains: applications to kernel smoothing and Metropolis-Hasting. *Bernoulli*, to appear.

Birg é, L. and Massart, P. (2006). Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 138(1–2):33–73.

Blanchard, G. and Zadorozhnyi, O. (2017). Concentration of weakly dependent Banach-valued sums and applications to kernel learning methods. *arXiv preprint arXiv:1712.01934*.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Catoni, O. (2003). Laplace transform estimates and deviation inequalities. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 39(1):1–26.

Collet, P., Martinez, S., and Schmitt, B. (2002). Exponential inequalities for dynamical measures of expanding maps of the interval. *Probability Theory and Related Fields*, 123(3):301–322.

Dahlhaus, R. (1996). On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic processes and their applications*, 62(1):139–168.

de la Pena, V. (1999). A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1):537–564.

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

Dedecker, J. and Fan, X. (2015). Deviation inequalities for separately Lipschitz functionals of iterated random functions. *Stochastic Processes and their Applications*, 125(1):60–90.

Diaconis, P. and Freedman, D. (1999). Iterated random functions. *SIAM review*, 41(1):45–76.

Doukhan, P. (2018). *Stochastic models for time series*. Springer.

Doukhan, P. and Neumann, M. H. (2007). Probability and moment inequalities for sums of weakly dependent random variables, with applications. *Stochastic Processes and their Applications*, 117:878–903.

Fan, J., Jiang, B., and Sun, Q. (2018). Hoeffding's lemma for Markov Chains and its applications to statistical learning. *arXiv preprint arXiv:1802.00211*.

Hang, H. and Steinwart, I. (2014). Fast learning from α-mixing observations. *Journal of Multivariate Analysis*, 127:184–199.

Hang, H. and Steinwart, I. (2017). A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics*, 45(2):708–743.

Joulin, A. and Ollivier, Y. (2010). Curvature, concentration and error estimates for Markov chain Monte Carlo. *The Annals of Probability*, 38(6):2418–2442.

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

Kuznetsov, V. and Mohri, M. (2015). Learning theory and algorithms for forecasting non-stationary time series. In *Advances in neural information processing systems*, 541–549.

Lerasle, M. (2011). Optimal model selection for density estimation of stationary data under various mixing conditions. *The Annals of Statistics*, 39(4):1852–1877.

London, B., Huang, B., Taskar, B., and Getoor, L. (2014). PAC-Bayesian collective stability. In *Artificial Intelligence and Statistics*, 585–594.

Massart, P. (2007). *Concentration inequalities and model selection, Ecole d'été de Probabilités de Saint-Flour 2003*. Lecture Notes in Mathematics 1896, Springer.

McDonald, D. J., Shalizi, C. R., and Schervish, M. (2017). Nonparametric risk bounds for time-series forecasting. *Journal of Machine Learning Research*, 18(32):1–40.

Meir, R. (2000). Nonparametric time series prediction through adaptive model selection. *Machine learning*, 39(1):5–34.

Merlevède, F., Peligrad, M., and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, pages 273–292. Institute of Mathematical Statistics.

Merlevède, F., Peligrad, M., and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4):435–474.

Modha, D. S. and Masry, E. (2002). Minimum complexity regression estimation with weakly dependent observations *IEEE Transactions on Information Theory*, 42:2133–2145.

Paulin, D. (2018). Concentration inequalities for Markov chains by Marton couplings and spectral methods *Electronic Journal of Probability*, 20(79):1–32.

Rio, E. (2013a). Extensions of the Hoeffding-Azuma inequalities. *Electronic Commununications in Probability*, 18(54):1–6.

Rio, E. (2013b). On McDiarmids concentration inequality. *Electronic Commununications in Probability*, 18(44):1–11.

Rio, E. (2017). *Asymptotic theory of weakly dependent random processes*, Springer, Berlin.

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

Sanchez-Perez, A. (2015). Time series prediction via aggregation: an oracle bound including numerical cost. In *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Antoniadis, A., Poggi, J.-M. and Brossat, X. (Eds.), Springer, 243–265.

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

Shalizi, C. and Kontorovich, A. (2013). Predictive PAC learning and process decompositions. In *Advances in neural information processing systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Weinberger (Eds.), 26:1619–1627.

Steinwart, I. and Christmann, A. (2009). Fast learning from non i.i.d observations. In *Advances in neural information processing systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams and A. Culotta (Eds.), 22:1768–1776.

Steinwart, I., Hush, D., and Scovel, C. (2009). Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194.

van de Geer, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics*, 23(5):1779–1801.

Vapnik, V. (1998). *Statistical learning theory.* Wiley, New York.

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

Wintenberger, O. (2017). Exponential inequalities for unbounded functions of geometrically ergodic Markov chains: applications to quantitative error bounds for regenerative Metropolis algorithms. *Statistics*, 51(1):222–234.