Deviation inequalities for Markov chains, with applications to SGD and empirical risk minimization

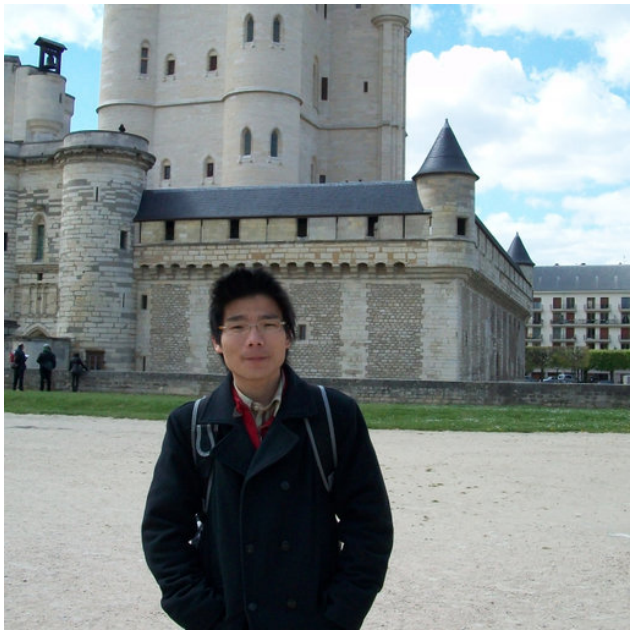
Pierre Alquier

RIKEN AIP
Center for Advanced Intelligence Project

High-Dimensional Statistical Modeling Team Seminar
March 1st, 2022

Co-authors

Fan, X. and Alquier, P. and Doukhan, P. (2021). *Deviation inequalities for stochastic approximation by averaging*. Preprint arXiv :2102.08685.



Xiequan Fan

Tianjin University

Paul Doukhan

CY Cergy Paris Université

Objective

General problem in probability and statistics

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right)\right| \geq x\right\} \leq ?$$

What can we expect ? (1/2)

Chebyshev's inequality

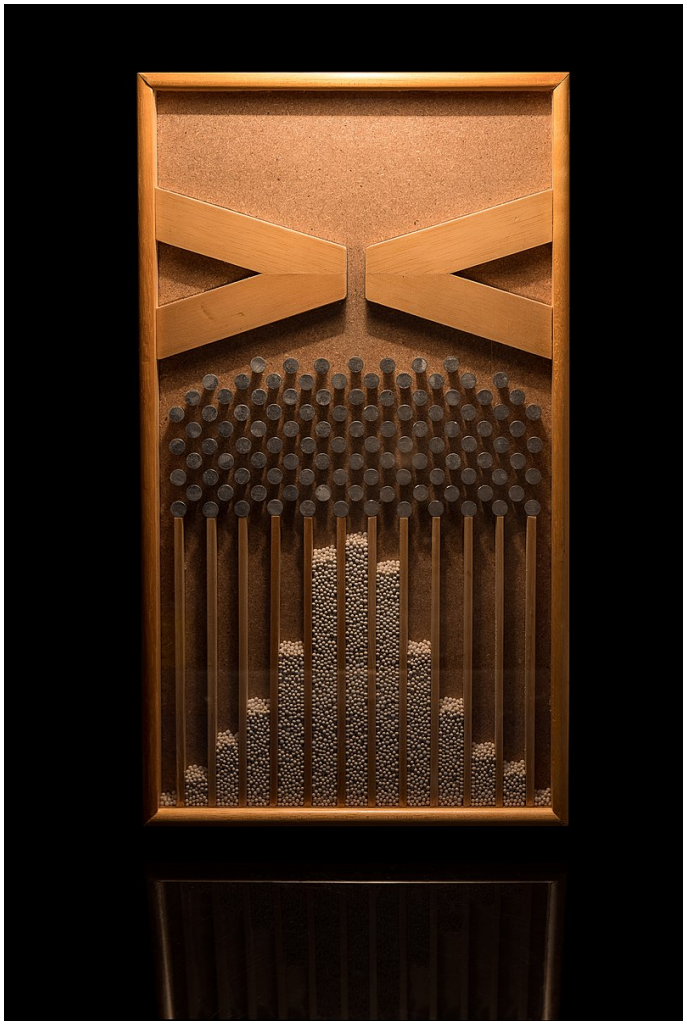
$$\mathbb{P}\Big\{|U - \mathbb{E}(U)| \geq x\Big\} \leq \frac{\mathrm{Var}(U)}{x^2}.$$

In a first time, assume the X_i's are independent, $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq x\right\} = \frac{\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right)}{n^2 x^2}$$

$$= \frac{\sigma^2}{nx^2}.$$

But...



(Photo : Wikipedia).

What can we expect ? (2/2)

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq x\right\} \leq \frac{\sigma^2}{nx^2}.$$

However, CLT :

$$\sqrt{\frac{n}{\sigma^2}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \rightsquigarrow \mathcal{N}(0,1).$$

So, we expect :

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq x\right\} \simeq 2\Phi\left(\frac{x\sqrt{n}}{\sigma}\right) \sim \frac{2\mathrm{e}^{-\frac{x^2 n}{2\sigma^2}}}{\frac{x\sqrt{n}}{\sigma}\sqrt{2\pi}}.$$

Chernoff bound

Chernoff bound

$$\mathbb{P}\Big\{ U - \mathbb{E}(U) \geq x \Big\} = \mathbb{P}\Big\{ \mathrm{e}^{s(U-\mathbb{E}(U))} \geq \mathrm{e}^{sx} \Big\} \leq \frac{\mathbb{E}\left(\mathrm{e}^{s(U-\mathbb{E}(U))}\right)}{\mathrm{e}^{sx}}.$$

$$\mathbb{P}\left\{ \frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq x \right\} \leq \frac{\mathbb{E}\left(\mathrm{e}^{\frac{s}{n}\sum_{i=1}^{n}(X_i-\mu)}\right)}{\mathrm{e}^{sx}}$$

$$= \mathrm{e}^{-sx} \prod_{i=1}^{n} \mathbb{E}\left(\mathrm{e}^{\frac{s}{n}(X_i-\mu)}\right).$$

Hoeffding's inequality

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq x\right\} \leq \mathrm{e}^{-sx}\prod_{i=1}^{n}\mathbb{E}\left(\mathrm{e}^{\frac{s}{n}(X_i-\mu)}\right).$$

Hoeffding's lemma - U bounded : $a \leq U \leq b$

$$\mathbb{E}\left(\mathrm{e}^{s[U-\mathbb{E}(U)]}\right) \leq \mathrm{e}^{\frac{s^2(b-a)^2}{8}}.$$

Hoeffding's inequality

Assume the X_i's are independent and $a \leq X_i \leq b$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq x\right\} \leq 2\mathrm{e}^{-\frac{2nx^2}{(b-a)^2}}.$$

McDiarmid's inequality

McDiarmid's inequality

Assume the X_i's are independent and $f : \mathcal{X}^n \to \mathbb{R}$ such that

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c.$$

then

$$\mathbb{P}\left\{ \left| \frac{f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]}{n} \right| \geq x \right\} \leq 2\mathrm{e}^{-\frac{2x^2 n}{c^2}}.$$

We recover Hoeffding for $f(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$, $c = (b-a)$.

Assumptions on moments

Hoeffding's lemma - U bounded : $a \leq U \leq b$

$$\mathbb{E}\left(\mathrm{e}^{s[U-\mathbb{E}(U)]}\right) \leq \mathrm{e}^{\frac{s^2(b-a)^2}{8}}.$$

In general, why not assuming U satisfies such an inequality ?

Definition - sub-Gaussian random variable U

$$\mathbb{E}\left(\mathrm{e}^{s[U-\mathbb{E}(U)]}\right) \leq \mathrm{e}^{s^2 C_0^2}$$

$$U \text{ sub-Gaussian } \Leftrightarrow \forall k \in \mathbb{N},\ \mathbb{E}(|U|^{2k}) \leq k! C_1^k.$$

Contents

Objective of this talk

Objective : for some time series $\{X_t, t = 0, \ldots, \infty\}$

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{t=1}^{n} X_t - \frac{1}{n}\mathbb{E}\left(\sum_{t=1}^{n} X_t\right)\right| \geq x\right\} \leq ?$$

$$\mathbb{P}\left\{\left|\frac{f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]}{n}\right| \geq x\right\} \leq ?$$

$$\mathbb{P}\left\{\frac{1}{n}\sum_{t=1}^{n} X_t - \mu \geq x\right\} \leq \frac{\mathbb{E}\left(\mathrm{e}^{\frac{s}{n}\sum_{t=1}^{n}(X_t-\mu)}\right)}{\mathrm{e}^{sx}}$$

$$= \mathrm{e}^{-sx}\prod_{t=1}^{n}\mathbb{E}\left(\mathrm{e}^{\frac{s}{n}(X_t-\mu)}\right)$$

Objective of this talk

Objective : for some time series $\{X_t, t = 0, \ldots, \infty\}$

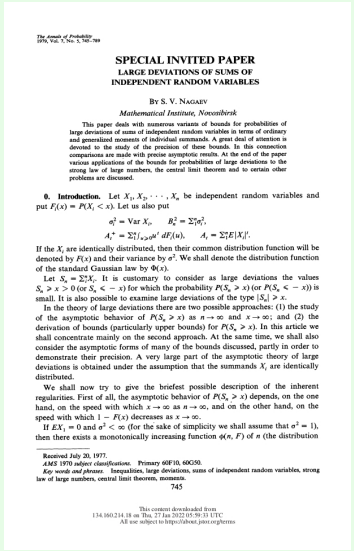
$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{t=1}^{n} X_t - \frac{1}{n}\mathbb{E}\left(\sum_{t=1}^{n} X_t\right)\right| \geq x\right\} \leq ?$$

$$\mathbb{P}\left\{\left|\frac{f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]}{n}\right| \geq x\right\} \leq ?$$

$$\mathbb{P}\left\{\frac{1}{n}\sum_{t=1}^{n} X_t \quad \right) \leq \frac{\mathbb{E}\left(e^{\frac{s}{n}\sum_{t=1}^{n}(X_t-\mu)}\right)}{e^{sx}}$$

$$= e^{-sx}\prod_{t=1}^{n}\mathbb{E}\left(e^{\frac{s}{n}(X_t-\mu)}\right)$$

Deviation for time series : an active research field



The Annals of Probability
1976, Vol. 4, No. 5, 745–789

SPECIAL INVITED PAPER

LARGE DEVIATIONS OF SUMS OF INDEPENDENT RANDOM VARIABLES

By S. V. Nagaev

Mathematical Institute, Novosibirsk

This paper deals with numerous variants of bounds for probabilities of large deviations of sums of independent random variables in terms of ordinary and generalized moments of individual summands. A great deal of attention is devoted to the study of the precision of these bounds. In this connection comparisons are made with precise asymptotic results. At the end of the paper various applications of the bounds for probabilities of large deviations to the strong law of large numbers, the central limit theorem and to certain other problems are discussed.

0. Introduction. Let $X_1, X_2, \cdots, X_n$ be independent random variables and put $F_j(x) = P(X_j < x)$. Let us also put

$$\sigma_j^2 = \operatorname{Var} X_j, \qquad B_n^2 = \sum_{j=1}^n \sigma_j^2,$$

$$A_n^{+} = \sum_{j=1}^n \int_{x \geq 0} x^t \, dF_j(x), \qquad A_t = \sum_{j=1}^n E|X_j|^t.$$

If the X_j are identically distributed, then their common distribution function will be denoted by $F(x)$ and their variance by σ^2. We shall denote the distribution function of the standard Gaussian law by $\Phi(x)$.

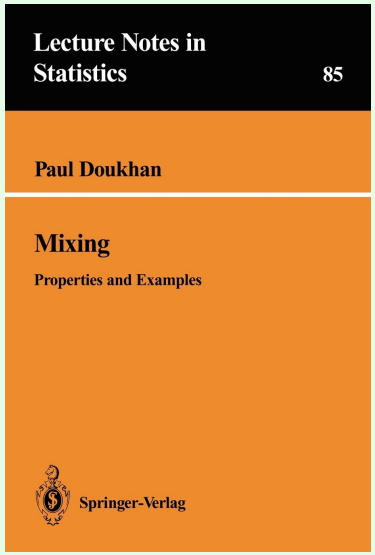
Let $S_n = \sum_{j=1}^n X_j$. It is customary to consider as large deviations the values $S_n \geq x > 0$ (or $S_n < -x$) for which the probability $P(S_n \geq x)$ (or $P(S_n < -x)$) is small. It is also possible to examine large deviations of the type $|S_n| \geq x$.

In the theory of large deviations there are two possible approaches: (1) the study of the asymptotic behavior of $P(S_n \geq x)$ as $n \to \infty$ and $x \to \infty$; and (2) the derivation of bounds (particulary upper bounds) for $P(S_n \geq x)$. In this article we shall concentrate mainly on the second approach. At the same time, we shall also consider the asymptotic forms of many of the bounds discussed, partly in order to demonstrate their precision. A very large part of the asymptotic theory of large deviations is obtained under the assumption that the summands X_j are identically distributed.

We shall now try to give the briefest possible description of the inherent regularities. First of all, the asymptotic behavior of $P(S_n \geq x)$ depends, on the one hand, on the speed with which $x \to \infty$ as $n \to \infty$, and on the other hand, on the speed with which $1 - F(x)$ decreases as $x \to \infty$.

If $EX_j = 0$ and $\sigma^2 < \infty$ (for the sake of simplicity we shall assume that $\sigma^2 = 1$), then there exists a monotonically increasing function $\phi(n, F)$ of n (the distribution

Received July 20, 1977.
AMS 1970 subject classifications. Primary 60F10, 60G50.
Key words and phrases. Inequalities, large deviations, sums of independent random variables, strong law of large numbers, central limit theorem, momenta.



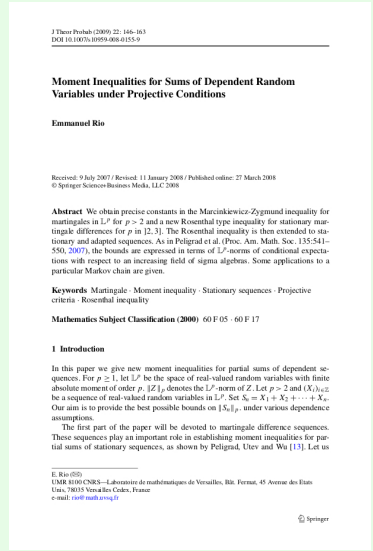
Lecture Notes in Statistics 85

Paul Doukhan

Mixing

Properties and Examples

Springer-Verlag



J Theor Probab (2009) 22: 146–163
DOI 10.1007/s10959-008-0155-9

Moment Inequalities for Sums of Dependent Random Variables under Projective Conditions

Emmanuel Rio

Received: 9 July 2007 / Revised: 11 January 2008 / Published online: 27 March 2008

Abstract We obtain precise constants in the Marcinkiewicz-Zygmund inequality for martingales in $\mathbb{L}^p$ for $p > 2$ and a new Rosenthal type inequality for stationary martingale differences for p in $[2, 3]$. The Rosenthal inequality is then extended to stationary and adapted sequences. As in Peligrad et al. (Proc. Am. Math. Soc. 135:541–550, 2007), the bounds are expressed in terms of $\mathbb{L}^p$-norms of conditional expectations with respect to an increasing field of sigma algebras. Some applications to a particular Markov chain are given.

Keywords Martingale · Moment inequality · Stationary sequences · Projective criteria · Rosenthal inequality

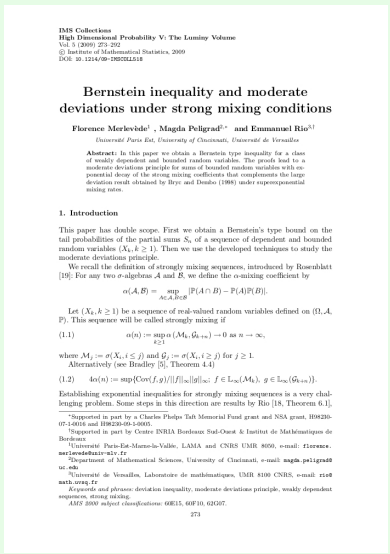
Mathematics Subject Classification (2000) 60 F 05 · 60 F 17

1 Introduction

In this paper we give new moment inequalities for partial sums of dependent sequences. For $p \geq 1$, let $\mathbb{L}^p$ be the space of real-valued random variables with finite absolute moment of order p. $\|Z\|_p$ denotes the $\mathbb{L}^p$-norm of Z. Let $p > 2$ and $(X_i)_{i \in \mathbb{Z}}$ be a sequence of real-valued random variables in $\mathbb{L}^p$. Set $S_n = X_1 + X_2 + \cdots + X_n$. Our aim is to provide the best possible bounds on $\|S_n\|_p$, under various dependence assumptions.

The first part of the paper will be devoted to martingale difference sequences. These sequences play an important role in establishing moment inequalities for partial sums of stationary sequences, as shown by Peligrad, Utev and Wu [13]. Let us

E. Rio (✉)
UMR 8100 CNRS—Laboratoire de mathématiques de Versailles, Bât. Fermat, 45 Avenue des Etats Unis, 78035 Versailles Cedex, France
e-mail: rio@ math.uvsq.fr



IMS Collections
High Dimensional Probability V: The Luminy Volume
Vol. 5 (2009) 273–292
© Institute of Mathematical Statistics, 2009
DOI: 10.1214/09-IMSCOLL518

Bernstein inequality and moderate deviations under strong mixing conditions

Florence Merlevède[1], Magda Peligrad[2,*] and Emmanuel Rio[3,†]

Université Paris Est, University of Cincinnati, Université de Versailles

Abstract: In this paper we obtain a Bernstein type inequality for a class of weakly dependent and bounded random variables. The proofs lead to a moderate deviations principle for sums of bounded random variables with exponential decay of the strong mixing coefficients that complements the large deviation result obtained by Bryc and Dembo (1988) under supexponential mixing rates.

1. Introduction

This paper has double scope. First we obtain a Bernstein's type bound on the tail probabilities of the partial sums S_n of a sequence of dependent and bounded random variables $(X_k, k \geq 1)$. Then we use the developed techniques to study the moderate deviations principle.

We recall the definition of strongly mixing sequences, introduced by Rosenblatt [19]: For any two σ-algebras $\mathcal{A}$ and $\mathcal{B}$, we define the α-mixing coefficient by

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

Let $(X_k, k \geq 1)$ be a sequence of real-valued random variables defined on $(\Omega, \mathcal{A}, \mathbb{P})$. This sequence will be called strongly mixing if

$$\alpha(n) := \sup_{k \geq 1} \alpha\left(\mathcal{M}_k, \mathcal{G}_{k+n}\right) \to 0 \text{ as } n \to \infty, \tag{1.1}$$

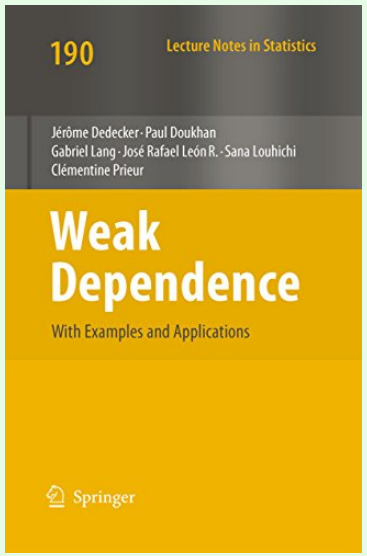
where $\mathcal{M}_j := \sigma(X_i, i \leq j)$ and $\mathcal{G}_j := \sigma(X_i, i \geq j)$ for $j \geq 1$.

Alternatively (see Bradley [5], Theorem 4.4)

$$4\alpha(n) := \sup\{\operatorname{Cov}(f, g)/\|f\|_\infty \|g\|_\infty : \ f \in \mathbb{L}_\infty(\mathcal{M}_k), \ g \in \mathbb{L}_\infty(\mathcal{G}_{k+n})\}. \tag{1.2}$$

Establishing exponential inequalities for strongly mixing sequences is a very challenging problem. Some steps in this direction are results by Rio [18, Theorem 6.1],

*Supported in part by a Charles Phelps Taft Memorial Fund grant and NSA grant, H98230-07-1-0016 and H98230-09-1-0005.
†Supported in part by Centre INRIA Bordeaux Sud-Ouest & Institut de Mathématiques de Bordeaux
[1]Université Paris-Est-Marne-la-Vallée, LAMA and CNRS UMR 8050, e-mail: florence.merlevede@univ-mlv.fr
[2]Department of Mathematical Sciences, University of Cincinnati, e-mail: magda.peligrad@uc.edu
[3]Université de Versailles, Laboratoire de mathématiques, UMR 8100 CNRS, e-mail: rio@math.uvsq.fr
Keywords and phrases: deviation inequality, moderate deviations principle, weakly dependent sequences, strong mixing.
AMS 2000 subject classifications: 60E15, 60F10, 62G07.



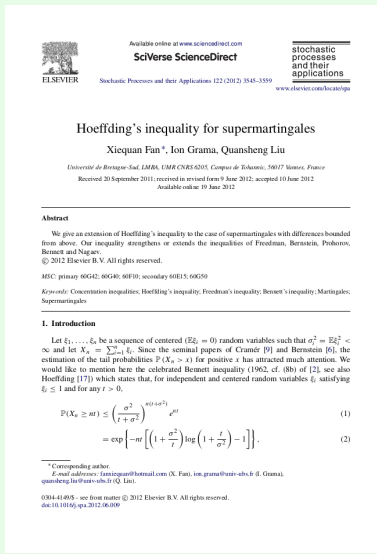
190 Lecture Notes in Statistics

Jérôme Dedecker · Paul Doukhan
Gabriel Lang · José Rafael León R. · Sana Louhichi
Clémentine Prieur

Weak Dependence

With Examples and Applications

Springer



Available online at www.sciencedirect.com
SciVerse ScienceDirect
Stochastic Processes and their Applications 122 (2012) 3545–3559
www.elsevier.com/locate/spa

Hoeffding's inequality for supermartingales

Xiequan Fan*, Ion Grama, Quansheng Liu

Université de Bretagne-Sud, LMBA, UMR CNRS 6205, Campus de Tohannic, 56017 Vannes, France

Received 20 September 2011; received in revised form 9 June 2012; accepted 10 June 2012
Available online 19 June 2012

Abstract

We give an extension of Hoeffding's inequality to the case of supermartingales with differences bounded from above. Our inequality strengthens or extends the inequalities of Freedman, Bernstein, Prohorov, Bennett and Nagaev.

MSC: primary 60G42; 60G40; 60F10; secondary 60E15; 60G50

Keywords: Concentration inequalities; Hoeffding's inequality; Freedman's inequality; Bennett's inequality; Martingales; Supermartingales

1. Introduction

Let $\xi_1, \ldots, \xi_n$ be a sequence of centered ($\mathbb{E}\xi_i = 0$) random variables such that $\sigma_i^2 = \mathbb{E}\xi_i^2 < \infty$ and let $X_n = \sum_{i=1}^n \xi_i$. Since the seminal papers of Cramér [9] and Bernstein [6], the estimation of the tail probabilities $\mathbb{P}(X_n > x)$ for positive x has attracted much attention. We would like to mention here the celebrated Bennett inequality (1962, cf. (8b) of [2], see also Hoeffding [17]) which states that, for independent and centered random variables ξ_i satisfying $\xi_i \leq 1$ and for any $t > 0$,

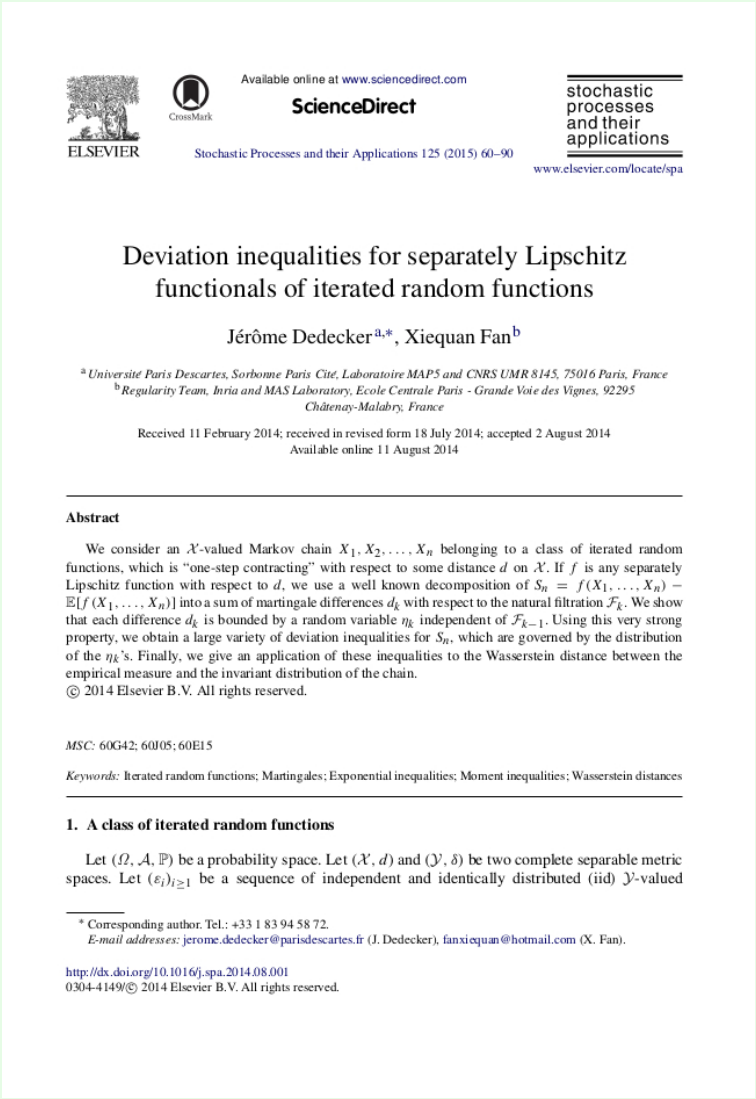
$$\mathbb{P}(X_n \geq nt) \leq \left(\frac{\sigma^2}{t + \sigma^2}\right)^{n(t+\sigma^2)} e^{nt} \tag{1}$$

$$= \exp\left\{-nt\left[\left(1 + \frac{\sigma^2}{t}\right)\log\left(1 + \frac{t}{\sigma^2}\right) - 1\right]\right\}. \tag{2}$$

* Corresponding author.
E-mail addresses: fanxiequan@hotmail.com (X. Fan), ion.grama@univ-ubs.fr (I. Grama), quansheng.liu@univ-ubs.fr (Q. Liu).

0304-4149/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.spa.2012.06.009

A remarkable result for Markov chains



Available online at www.sciencedirect.com
ScienceDirect
Stochastic Processes and their Applications 125 (2015) 60–90
stochastic processes and their applications
www.elsevier.com/locate/spa

Deviation inequalities for separately Lipschitz functionals of iterated random functions

Jérôme Dedecker[a,*], Xiequan Fan[b]

[a] Universite Paris Descartes, Sorbonne Paris Cite, Laboratoire MAP5 and CNRS UMR 8145, 75016 Paris, France
[b] Regularity Team, Inria and MAS Laboratory, Ecole Centrale Paris - Grande Voie des Vignes, 92295 Châtenay-Malabry, France

Received 11 February 2014; received in revised form 18 July 2014; accepted 2 August 2014
Available online 11 August 2014

Abstract

We consider an $\mathcal{X}$-valued Markov chain $X_1, X_2, \ldots, X_n$ belonging to a class of iterated random functions, which is "one-step contracting" with respect to some distance d on $\mathcal{X}$. If f is any separately Lipschitz function with respect to d, we use a well known decomposition of $S_n = f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]$ into a sum of martingale differences d_k with respect to the natural filtration $\mathcal{F}_k$. We show that each difference d_k is bounded by a random variable η_k independent of $\mathcal{F}_{k-1}$. Using this very strong property, we obtain a large variety of deviation inequalities for S_n, which are governed by the distribution of the η_k's. Finally, we give an application of these inequalities to the Wasserstein distance between the empirical measure and the invariant distribution of the chain.

MSC: 60G42; 60J05; 60E15

Keywords: Iterated random functions; Martingales; Exponential inequalities; Moment inequalities; Wasserstein distances

1. A class of iterated random functions

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Let $(\mathcal{X}, d)$ and $(\mathcal{Y}, \delta)$ be two complete separable metric spaces. Let $(\varepsilon_i)_{i\geq 1}$ be a sequence of independent and identically distributed (iid) $\mathcal{Y}$-valued

* Corresponding author. Tel.: +33 1 83 94 58 72.
E-mail addresses: jerome.dedecker@parisdescartes.fr (J. Dedecker), fanxiequan@hotmail.com (X. Fan).

http://dx.doi.org/10.1016/j.spa.2014.08.001
0304-4149/

- study Markov chains of the form

$$X_n = F(X_{n-1}, \varepsilon_n)$$

- provide deviation inequalities when

$$\mathbb{E}\left\{ d\Big(F(x, \varepsilon_n), F(x', \varepsilon_n) \Big) \right\} \leq \rho d(x, x')$$

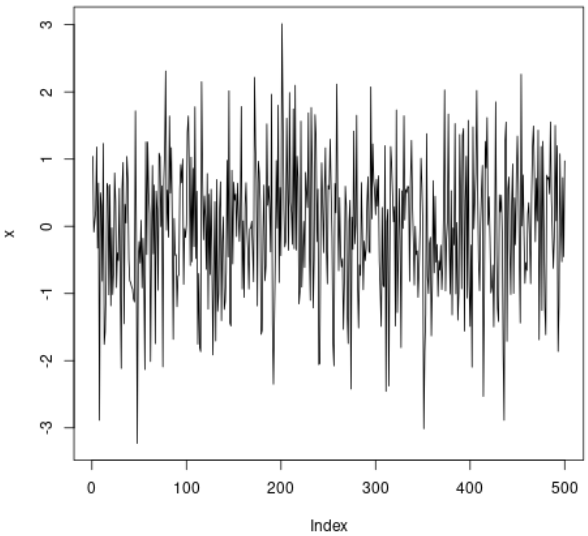
for some $\rho < 1$.

Example (1/2)

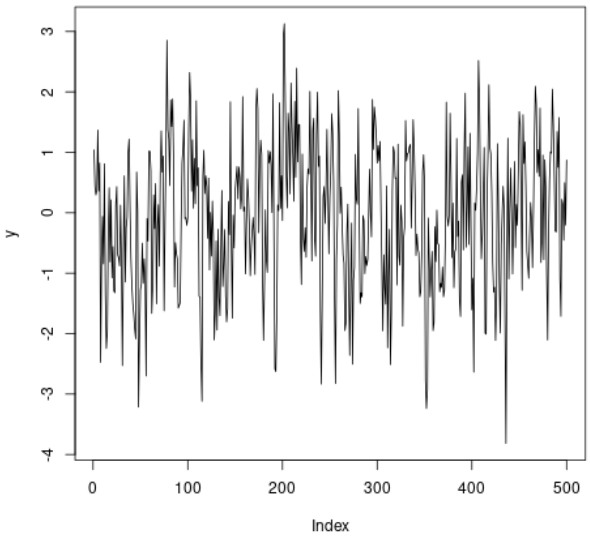
AR(1) process

$$X_n = F(X_{n-1}, \varepsilon_n) := \rho X_{n-1} + \varepsilon_n$$

$$|F(x, \varepsilon_n) - F(x', \varepsilon_n)| \leq \rho |x - x'|$$



$\rho = 0$



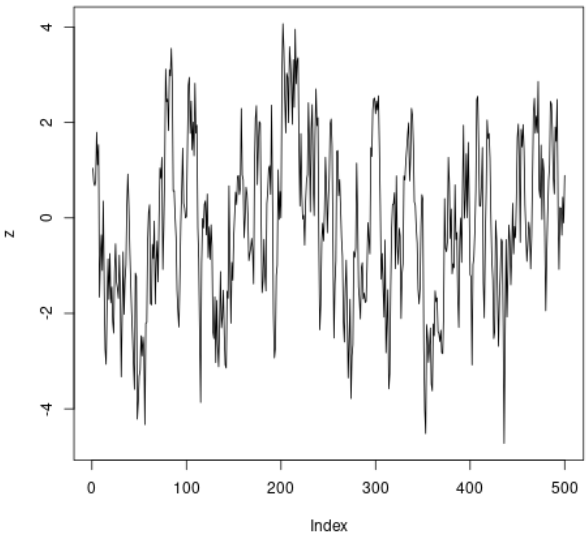
$\rho = 0.5$

Example (2/2)

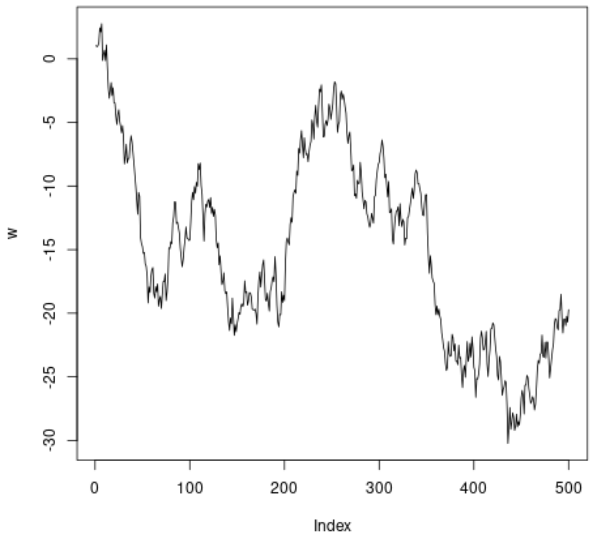
AR(1) process

$$X_n = F(X_{n-1}, \varepsilon_n) := \rho X_{n-1} + \varepsilon_n$$

$$|F(x, \varepsilon_n) - F(x', \varepsilon_n)| \leq \rho |x - x'|$$



$\rho = 0.8$

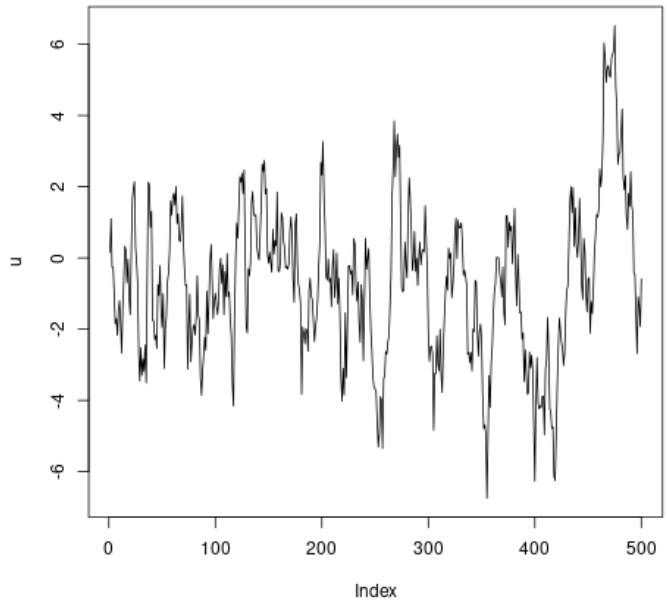


$\rho = 1$

What happens for non-homogeneous chains ?

AR(1) process with varying coefficients

$$X_n = F_n(X_{n-1}, \varepsilon_n) := \rho_n X_{n-1} + \varepsilon_n$$



$$\rho_n = \left(1 - \frac{1}{\sqrt{n}}\right).$$

Inequalities for non-homogeneous Markov chains

A class of non-homogeneous Markov chains

- X_n takes values in $(\mathcal{X}, d)$. Example : $\mathcal{X} = \mathbb{R}^d$, d large.
- (ε_n) are i.i.d random variables in $(\mathcal{Y}, \delta)$.

Definition

1. $X_n = F_n(X_{n-1}, \varepsilon_n)$.
2. $\mathbb{E}\left\{ d\Big(F_n(x, \varepsilon_n), F_n(x', \varepsilon_n)\Big) \right\} \leq \rho_n d(x, x')$.
3. $d\Big(F_n(x, y), F_n(x, y')\Big) \leq \tau_n \delta(y, y') + \xi$, $\xi = 0$ or $\tau_n \equiv 0$.

VAR with variying coefficients

Phillips, P.C.B. (1988). Regression theory for near integrated time series. *Econometrica*.

- $X_n \in \mathbb{R}^d$.
- (ε_n) are i.i.d $\mathcal{N}(0, \sigma^2 I_d)$.

1. $X_n = F_n(X_{n-1}, \varepsilon_n) = A_n X_{n-1} + \varepsilon_n$.
2. $\rho_n = \|A_n\|_{\mathrm{op}} = \sup_{x \neq 0} \frac{\|A_n x\|}{\|x\|} \xrightarrow[n \to \infty]{<} 1$.
3. $\tau_n = 1$.

Example : stochastic optimization

$$\text{Minimize } L(x) = \sum_{i=1}^{N} \ell_i(x)$$

For I drawn uniformly in $\{1, \dots, N\}$ with M elements,

$$\hat{\nabla}_n L(x) := \frac{1}{M} \sum_{i \in I} \nabla \ell_i(x).$$

- Projected tochastic gradient descent (SGD) :

$$X_n = \Pi_{\mathcal{C}} \left[X_{n-1} - \frac{\gamma}{n^\alpha} \hat{\nabla}_n L(x) \right]$$

- Projected stochastic gradient Langevin descent (SGLD) :

$$X_n = \Pi_{\mathcal{C}} \left[X_{n-1} - \frac{\gamma}{n^\alpha} \hat{\nabla}_n L(x) + \frac{\eta}{n^\beta} \varepsilon_n \right]$$

Example : SGD

Assume L is m-strongly convex and ∇L is ℓ-Lipschitz.

SGD - $\alpha \in [0,1]$, $\gamma > 0$

1. $X_n = F_n(X_{n-1}, \varepsilon_n) = \Pi_{\mathcal{C}} \left[X_{n-1} - \frac{\gamma}{n^\alpha} \hat{\nabla}_n L(x) \right]$.
2.
   - $\rho_n \sim 1 - \frac{m\gamma}{n^\alpha}$ for $\alpha > 0$,
   - $\rho_n = 1 - 2m\gamma + \ell^2\gamma^2$ if $\alpha = 0$.
3. $\xi = \mathrm{diam}(\mathcal{C})$.

Example : SGLD

Assume L is m-strongly convex and ∇L is ℓ-Lipschitz.

SGLD - $\alpha, \beta \in [0,1]$, $\gamma, \eta > 0$, $\varepsilon_n \sim \mathcal{N}(0,1)$

1. $X_n = F_n(X_{n-1}, \varepsilon_n) = \Pi_{\mathcal{C}} \left[X_{n-1} - \frac{\gamma}{n^\alpha} \hat{\nabla}_n L(x) + \frac{\eta}{n^\beta} \varepsilon_n \right].$
2.
   - $\rho_n \sim 1 - \frac{m\gamma}{n^\alpha}$ for $\alpha > 0$,
   - $\rho_n = 1 - 2m\gamma + \ell^2\gamma^2$ if $\alpha = 0$.
3. $\xi = \operatorname{diam}(\mathcal{C})$.

Deviation inequality

Theorem (Proposition 3.1 in the paper) - $p \in [1, +\infty], d \in \mathbb{N}$

Assume $f : \mathcal{X}^n \to \mathbb{R}^d$ such that

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq d(x_i, x_i'),$$

$\mathbb{E}_{\varepsilon_n}([\mathbb{E}_{\varepsilon_n'}\delta(\varepsilon_n, \varepsilon_n')]^k) \leq C_1^k k!$ and a similar condition for X_1,

$$\mathbb{P}\left\{\left\|\frac{f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]}{n}\right\|_p \geq x\right\}$$

$$\leq \begin{cases} e^{-c_{p,d} n x} & \rho_n \leq \rho < 1,\ \tau_n \leq \frac{\tau}{n^\alpha},\ \alpha \in (0, 1] \\ e^{-c_{p,d} n (x 1_{x>1} + x^2 1_{x \leq 1})} & \rho_n \leq 1 - \frac{\rho}{n^\alpha},\ \tau_n \leq \frac{\tau}{n^\alpha},\ \alpha \in [0, 1) \text{ or} \\ & \rho_n \leq 1 - \frac{\rho}{n},\ \xi > 0, \\ e^{-c_{p,d} n^{1-2\alpha} x^2} & \rho_n \leq 1 - \frac{\rho}{n^\alpha},\ (\tau_n \leq \tau \text{ or } \xi > 0), \alpha \in (0, 1/2). \end{cases}$$

Proof technique

The proof technique relies on martingale decomposition :

$$f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] = \sum_{t=1}^{n} M_t$$

where

$$M_t = \mathbb{E}[f(X_1, \dots, X_n)|X_1, \dots, X_t] - \mathbb{E}[f(X_1, \dots, X_n)|X_1, \dots, X_{t-1}].$$

Conditional Chernoff :

$$\frac{\mathbb{E}\left(\mathrm{e}^{\frac{s}{n}\sum_{t=1}^{n} M_t}\right)}{\mathrm{e}^{sx}} = \frac{\mathbb{E}\left[\mathrm{e}^{\frac{s}{n}\sum_{t=1}^{n-1} M_t}\mathbb{E}\left(\mathrm{e}^{\frac{s}{n}M_n}|X_1, \dots, X_{n-1}\right)\right]}{\mathrm{e}^{sx}}.$$

Here the study of $\mathbb{E}\left(\mathrm{e}^{\frac{s}{n}M_n}|X_1, \dots, X_{n-1}\right)$ requires some care...

Shameless name-dropping

In the paper, we provide an exhaustive list of inequalities, under various moment assumptions :

- exponential inequalities :
  - McDiarmid,
  - Hoeffding,
  - Bernstein.
- semi-exponential inequalities :
  - Fuk-Nagaev,
  - von Bahr-Esseen.
- moment inequalities :
  - Marcinkiewicz-Zygmund,
  - von Bahr-Esseen.

Applications

1. Deviation inequalities for time series : introduction
   - Why deviation inequalities ?
   - Deviation inequalities for time series
2. Non-homogeneous Markov chains
   - Inequalities for non-homogeneous Markov chains
   - Applications in machine learning

Empirical risk minimization (1/2)

In the stationary case,

$$f(X_1, \dots, X_n) = \frac{1}{n} \sum_{t=1}^{n} \ell(\theta, X_i) = R_n(\theta)$$

then

$$\mathbb{E}\left[f(X_1, \dots, X_n)\right] = \mathbb{E}\left[\ell(\theta, X)\right] = R(\theta).$$

$$\mathbb{P}\left\{ \left| R(\theta) - R_n(\theta) \right| \geq x \right\}$$

$$\leq \begin{cases} \mathrm{e}^{-cnx} & \tau_n \leq \frac{\tau}{n^\alpha},\ \alpha \in (0,1] \\ \mathrm{e}^{-cn(x 1_{x>1} + x^2 1_{x \leq 1})} & \rho_n \leq 1 - \frac{\rho}{n^\alpha},\ \tau_n \leq \frac{\tau}{n^\alpha},\ \alpha \in [0,1) \\ \mathrm{e}^{-cn^{1-2\alpha}x^2} & \rho_n \leq 1 - \frac{\rho}{n^\alpha},\ \tau_n \leq \tau,\ \alpha \in (0,1/2). \end{cases}$$

Empirical risk minimization (2/2)

ERM

$$\hat{\theta} = \arg\min_{\theta \in \Theta} R_n(\theta).$$

Say $\mathrm{Card}(\Theta) = N$ is finite,

$$\mathbb{P}\left\{ R(\hat{\theta}) \geq R_n(\hat{\theta}) + x \right\} \leq \begin{cases} N\mathrm{e}^{-cnx}, \\ N\mathrm{e}^{-cn(x1_{x>1} + x^2 1_{x\leq 1})}, \\ N\mathrm{e}^{-cn^{1-2\alpha}x^2}. \end{cases}$$

Application to SGLD (1/2)

L is m-strongly convex and ∇L is ℓ-Lipschitz.

SGLD - $\alpha, \beta \in [0,1]$, $\gamma, \eta \geq 0$

$$X_n = \Pi_{\mathcal{C}} \left[X_{n-1} - \frac{\gamma}{n^\alpha} \hat{\nabla}_n L(x) + \frac{\eta}{n^\beta} \varepsilon_n \right], \quad \bar{X}_n = \frac{1}{n} \sum_{t=1}^{n} X_t.$$

For some $c_{p,d} = c_{p,d}(\ell, m)$,

$$\mathbb{P}\left\{ \left\| \bar{X}_n - \mathbb{E}(\bar{X}_n) \right\|_p \geq x \right\}$$

$$\leq \begin{cases} \mathrm{e}^{-c_{p,d} n (x 1_{x>1} + x^2 1_{x \leq 1})} & [0 \leq \alpha \leq \beta] \text{ or } [0 \leq \alpha, \eta = 0] \\ \mathrm{e}^{-c_{p,d} n^{1-2\alpha} x^2} & 0 < \alpha < \frac{1}{2},\ \beta = 0. \end{cases}$$

Application to SGLD (2/2)

Assume in addition that L is Lipschitz.

Theorem - $\alpha \in [1/2, 1], \eta = 0$

$$\mathbb{E}\left(\|\bar{X}_n - x^*\|_2^2 \right) \leq \frac{c}{n}, \quad c = c(d, L, m).$$

Bach, F. and Moulines, E. (2011). Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. *NIPS*.

Combine Bach & Moulines (2011) with our inequality

$$\mathbb{P}\left\{ \left\| \bar{X}_n - x^* \right\|_2 \leq \sqrt{\frac{c + \frac{1}{c_{2,d}} \log\left(\frac{1}{\delta}\right)}{n}} \right\} \geq 1 - \delta.$$