

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

Pierre Alquier



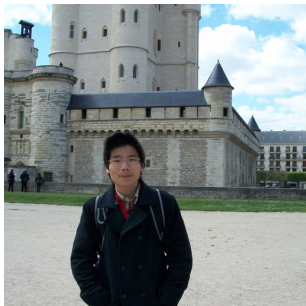
Center for
Advanced Intelligence Project

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)

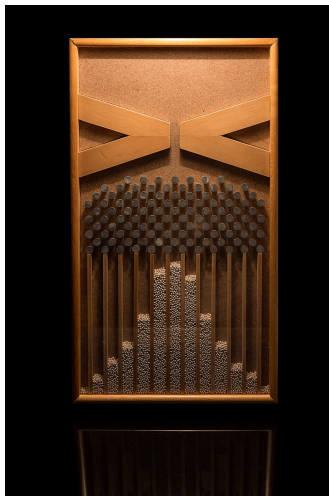
Chebyshev's inequality

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

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

$$\begin{aligned}\mathbb{P}\left\{\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| \geq x\right\} &= \frac{\text{Var}\left(\sum_{i=1}^n X_i\right)}{n^2 x^2} \\ &= \frac{\sigma^2}{n x^2}.\end{aligned}$$

But...



(Photo : Wikipedia).

What can we expect ? (2/2)

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

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 :

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

Chernoff bound

Chernoff bound

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

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

Hoeffding's inequality

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

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

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

Hoeffding's inequality

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

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

McDiarmid's inequality

McDiarmid's inequality

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

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

then

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

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

Assumptions on moments

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

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

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

Definition - sub-Gaussian random variable U

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

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

Contents

- 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

Objective of this talk

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

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

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

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

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

Objective of this talk

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

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

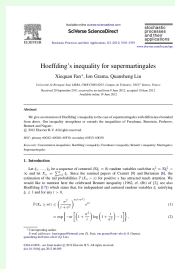
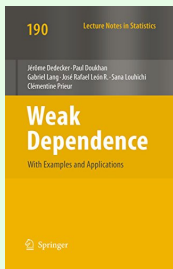
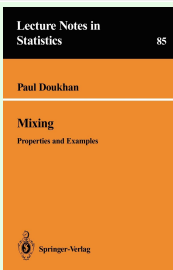
$$\mathbb{P} \left\{ \left| \frac{f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]}{n} \right| \geq x \right\} \leq ?$$



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

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

Deviation for time series : an active research field



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/sap

Deviation inequalities for separately Lipschitz functionals of iterated random functions

Jérôme Dedecker^{a,*}, Xiequan Fan^b

^a Université Paris Descartes, Sorbonne Paris Cité, Laboratoire MAP5 and CNRS UMR 8145, 75016 Paris, France

^b Regularity Team, Inria and MAS Laboratory, Ecole Centrale Paris – Grande Vitesse des Vigiers, 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 X -valued Markov chain X_1, X_2, \dots, X_n belonging to a class of iterated random functions, which is “one-step contracting” with respect to some distance d on X . If f is any separately Lipschitz function with respect to d , we use a well known decomposition of $S_n = f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]$ 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.

© 2014 Elsevier B.V. All rights reserved.

MSC: 60G42; 60B05; 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 (X, d) and (Y, δ) be two complete separable metric spaces. Let $(\eta_k)_{k \geq 1}$ 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), fanxqquan@hotmail.com (X. Fan).

<http://dx.doi.org/10.1016/j.sap.2014.08.001>
0304-4149/© 2014 Elsevier B.V. All rights reserved.

- study Markov chains of the form

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

- provide deviation inequalities when

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

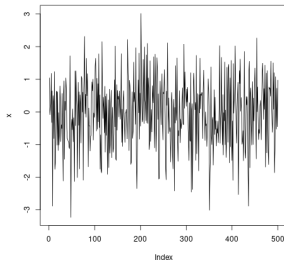
for some $\rho < 1$.

Example (1/2)

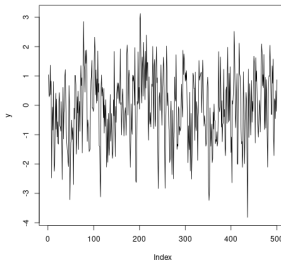
AR(1) process

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

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



$$\rho = 0$$



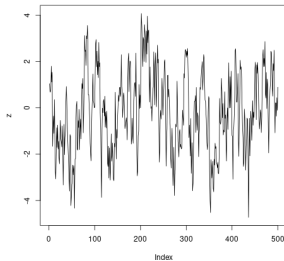
$$\rho = 0.5$$

Example (2/2)

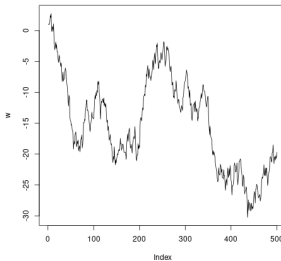
AR(1) process

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

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



$$\rho = 0.8$$

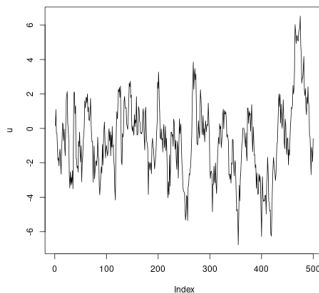


$$\rho = 1$$

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

- 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

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}, δ) .

Definition

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

VAR with varying 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 $\rho_n = \|A_n\|_{\text{op}} = \sup_{x \neq 0} \frac{\|A_n x\|}{\|x\|} \xrightarrow[n \rightarrow \infty]{<} 1$.
- 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 stochastic 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 $\xi = \text{diam}(\mathcal{C}).$

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 $\xi = \text{diam}(\mathcal{C}).$

Deviation inequality

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

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

$$|f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x'_i, \dots, x_n)| \leq d(x_i, x'_i),$$

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

$$\mathbb{P} \left\{ \left\| \frac{f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]}{n} \right\|_p \geq x \right\} \\ \leq \begin{cases} e^{-c_{p,d} n^x} & \rho_n \leq 1 - \rho < 1, \tau_n \leq \frac{\tau}{n^\alpha}, \alpha \in (0, 1] \\ e^{-c_{p,d} n(x \mathbf{1}_{x>1} + x^2 \mathbf{1}_{x \leq 1})} & \rho_n \leq 1 - \frac{\rho}{n^\alpha}, \tau_n \leq \frac{\tau}{n^\alpha}, \alpha \in [0, 1) \text{ or} \\ & \rho_n \leq 1 - \rho < 1, \xi > 0, \\ e^{-c_{p,d} n^{1-2\alpha} x^2} & \rho_n \leq 1 - \frac{\rho}{n^\alpha}, (\tau_n \leq \tau \text{ or } \xi > 0), \alpha \in (0, 1/2). \end{cases}$$

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 :

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

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

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

then

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

$$\mathbb{P} \left\{ \left| R(\theta) - R_n(\theta) \right| \geq x \right\} \leq \begin{cases} e^{-c n x} & \tau_n \leq \frac{\tau}{n^\alpha}, \alpha \in (0, 1] \\ e^{-c n (x 1_{x>1} + x^2 1_{x \leq 1})} & \rho_n \leq 1 - \frac{\rho}{n^\alpha}, \tau_n \leq \frac{\tau}{n^\alpha}, \alpha \in [0, 1) \\ e^{-c n^{1-2\alpha} x^2} & \rho_n \leq 1 - \frac{\rho}{n^\alpha}, \tau_n \leq \tau, \alpha \in (0, 1/2). \end{cases}$$

Empirical risk minimization (2/2)

ERM

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

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

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

Application to SGLD (1/2)

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

SGLD - $0 < \alpha < 1/2$, $\gamma > 0$, $\eta \geq 0$

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

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

$$\mathbb{P} \left\{ \left\| \bar{X}_n - \mathbb{E}(\bar{X}_n) \right\|_p \geq x \right\} \leq e^{-c_{p,d} n^{1-2\alpha} x^2}.$$

Application to SGLD (2/2)

Convergence of the expectation

Classical results on SGLD :

$$\|\mathbb{E}(\bar{X}_n - x^*)\|_2 \rightarrow 0$$

Combine this with our inequality

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