Deviation inequalities for stochastic approximation by averaging

Xiequan Fan, Pierre Alquier and Paul Doukhan

Center for Applied Mathematics, Tianjin University, Tianjin 300072, China.
E-mail: fanxiequan@hotmail.com.

RIKEN AIP, Japan.
E-mail: pierrealain.alquier@riken.jp.

CY University, Cergy-Pontoise, France.
E-mail: doukhan@cyu.fr.

Abstract

We introduce a class of Markov chains that includes models of stochastic approximation by averaging and non-averaging. Using a martingale approximation method, we establish various deviation inequalities for separately Lipschitz functions of such a chain, with different moment conditions on some dominating random variables of martingale differences. Finally, we apply these inequalities to stochastic approximation by averaging and empirical risk minimisation.

Keywords: Deviation inequalities; martingales; iterated random functions; stochastic approximation by averaging; empirical risk minimisation

2000 MSC: primary 60G42; 60J05; 60F10; secondary 60E15

1. Introduction

Markov chains, or iterated random functions, are of fundamental importance to model dependent phenomena. A nice reference on this topic is [10]. Probability inequalities for dependent variables were developed in [13], and more recently in [15, 17] as well as in [25, 26, 6, 7, 11, 12]. Most of these papers involve such inequalities for Markov chains. Recently, [8] provided such inequalities for contractive Markov chains thanks to a martingale based technique.

In these papers, only time homogeneous contractive Markov chains are considered. However, in many practical situations, such as stochastic approximation algorithms [20] and unit roots [21], the contraction coefficients are time-varying, and will tend either to 0 or to 1 as $n \to \infty$. In this paper, our objective is to provide results for such non-homogeneous Markov chains. Our framework is a large class of non-homogeneous models introduced in Section 1.2. Practical examples of chains fitting such conditions are considered in Section 1.3.

Using the martingale approximation method developed in [8], we establish various deviation inequalities for separately Lipschitz functions of such chains in Section 2. Our inequalities hold under various moment conditions on some dominating random variables of the martingale differences. Section 3 is dedicated to various classes of L^p-norm concentration inequalities, such as Bernstein type inequalities, semi-exponential bound, Fuk-Nagaev inequalities, as well as von Bahr-Esseen, McDiarmid and Hoeffding type bounds. Section 4 is devoted to moment inequalities: Marcinkiewicz-Zygmund and von Bahr-Esseen type bounds. Finally, in Section 5 we apply these inequalities to the stochastic approximation by averaging in Subsection 5.1 and to empirical risk minimisation (ERM) in Subsection 5.2.

1.1. Notations

In the paper, we adopt the convention that each $x \in \mathbb{R}^d$ is a column vector. The entries of x will be denoted by $x^{(1)}, \ldots, x^{(d)}$. The transpose of x will be denoted by x^T, thus $x^T = (x^{(1)}, \ldots, x^{(d)})$. The set of $d_1 \times d_2$ real-valued matrices will be denoted by $\mathbb{R}^{d_1 \times d_2}$, and I_d will denote the identity matrix in $\mathbb{R}^{d \times d}$. Let $\|\cdot\|$ denote a norm on $\mathbb{R}^d$. In most cases, we will use L^p norms. In this case, we will explicitely state that $\|\cdot\| = \|\cdot\|_p$, where

$$\|x\|_\infty = \max_{1 \le i \le d} |x^{(i)}| \quad \text{and} \quad \|x\|_p = \Big(\sum_{i=1}^d |x^{(i)}|^p \Big)^{1/p}, \quad p \in [1, \infty).$$

For any $M \in \mathbb{R}^{d \times d}$, we put

$$\lambda_{\min}^{(p)}(M) = \inf_{v \ne 0} \frac{\|Mx\|_p}{\|x\|_p} \quad \text{and} \quad \lambda_{\max}^{(p)}(M) = \sup_{x \ne 0} \frac{\|Mx\|_p}{\|x\|_p}.$$

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. All the random variables in the paper are defined over $(\Omega, \mathcal{A}, \mathbb{P})$. When V is a nonnegative real-valued random variable, we will let $\|V\|_\infty$ denote its essential supremum (note that there will be no ambiguity with the above). Finally, $(\mathcal{X}, d)$ and $(\mathcal{Y}, \delta)$ are two complete separable metric spaces. Our non-homogeneous Markov chains will take values in $\mathcal{X}$.

1.2. A class of iterated random functions

Let $(\varepsilon_i)_{i \ge 1}$ be a sequence of independent copies of a $\mathcal{Y}$-valued random variable ε. Let X_1 be a $\mathcal{X}$-valued random variable independent of $(\varepsilon_i)_{i \ge 2}$. We consider the Markov chain $(X_i)_{i \ge 1}$ such that

$$X_n = F_n(X_{n-1}, \varepsilon_n), \quad \text{for any } n \ge 2, \tag{1.1}$$

where $F_n : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ satisfies that there exists a positive number n_0 such that for any $n \ge n_0$,

$$\mathbb{E}\big[d\big(F_n(x, \varepsilon_1), F_n(x', \varepsilon_1)\big)\big] \le \rho_n d(x, x') \tag{1.2}$$

for some $\rho_n \in [0, 1)$, and

$$d(F_n(x, y), F_n(x, y')) \le \tau_n \delta(y, y') + \xi_n \tag{1.3}$$

for some $\tau_n \ge 0$, $\xi_n \ge 0$. The case $\xi_n \equiv 0$ corresponds to functions F_n that are Lipschitz with respect to ε_n, while the case $\tau_n \equiv 0$ corresponds to bounded chains. Note that when $\tau_n \equiv 0$, the metric δ is not involved in the properties of the chain.

The case where $F_n \equiv F$, $\rho_n \equiv \rho$, $\tau_n \equiv \tau$ and $\xi_n \equiv 0$ for two constants ρ and τ has been studied by Dedecker and Fan [8]. See also Dedecker, Doukhan and Fan [9] who weakened the condition in (1.3). In these papers, the authors have established very precise inequalities for Lipschitz functionals of the chain, by assuming various moment conditions. However, the conditions $F_n \equiv F$ and $\rho_n \equiv \rho$ are restrictive. They are not satisfied in many extremely useful models. For instance, the recursive algorithm of stochastic approximation in Polyak and Juditsky [20] returns a chain for which the conditions (1.1)–(1.3) are satisfied with $F_n(x, y) = (I_d - \tau_n A)x + \tau_n B - \tau_n y$ and $\rho_n = 1 - c\,\tau_n$ with $\tau_n \to 0$, where $A \in \mathbb{R}^{d\times d}$ is a positive-definite matrix, $x, y, B \in \mathbb{R}^d$ and $c, \tau_n > 0$. A special case of interest corresponds to $\rho_n = 1 - c_1/n^\alpha$ and $\tau_n = c_2/n^\alpha$ for three positive constants c_1, c_2 and $\alpha \in (0, 1)$. A second class of frequently used models which do not satisfy the condition $\rho_n \equiv \rho$ is that of time series auto-regressions with a unit root, see Phillips and Magdalinos [23]. In the model of Phillips and Magdalinos [23], the conditions (1.1)–(1.3) are satisfied with $\rho_n = 1 - c_1/n^\alpha$ and $\tau_n = c_2$. See also Phillips [21, 22] for the case $\rho_n = e^{-c_1/n}$ and $\tau_n = c_2$.

1.3. Examples

In this subsection, we give a non exhaustive list of models satisfying the conditions (1.1)–(1.3).

Example 1. In the case where $\mathcal{X}$ is a separable Banach space with norm $\|\cdot\|_{\mathcal{X}}$, and $d(x, x') = \|x - x'\|_{\mathcal{X}}$, let us consider the following functional auto-regressive model

$$X_n = f(X_{n-1}) + g(\varepsilon_n)\,, \tag{1.4}$$

where $f : \mathcal{X} \to \mathcal{X}$ and $g : \mathcal{Y} \to \mathcal{X}$ are such that

$$\|f(x) - f(x')\|_{\mathcal{X}} \le \rho\|x - x'\|_{\mathcal{X}} \quad \text{and} \quad \|g(y) - g(y')\|_{\mathcal{X}} \le \delta(y, y')$$

for some constant $\rho \in [0, 1)$. In this model, the conditions (1.1)–(1.3) are satisfied with

$$F_n(x, y) = f(x) + g(y), \quad \rho_n = \rho, \quad \tau_n = 1 \quad \text{and} \quad \xi_n \equiv 0 \tag{1.5}$$

for any $n \ge 1$. This model is a typical example considered in Dedecker and Fan [8]. We refer to the papers by Diaconis and Freedman [10] and Alquier et al. [1] for many other interesting examples.

Example 2. Consider the following auto-regression with a unit root model (see Phillips [21, 22] or Phillips and Magdalinos [23]): for any $n \ge 2$,

$$X_n = \frac{1}{1 + c/n^\alpha} X_{n-1} + \varepsilon_n \qquad \text{or} \tag{1.6}$$

$$X_n = (1 - \frac{c}{n^\alpha}) X_{n-1} + \varepsilon_n, \tag{1.7}$$

where $X_n, \varepsilon_n \in \mathbb{R}, \alpha \in (0, 1)$ and c is a positive constant. Let $d(x, x') = \delta(x, x') = |x - x'|$. In this model, the conditions (1.1)–(1.3) are satisfied with

$$\rho_n = 1 - \frac{c}{n^\alpha}, \quad \tau_n = 1 \quad \text{and} \quad \xi_n \equiv 0$$

for any n large enough. Moreover, if $c \in (0, 1)$, the conditions (1.1)–(1.3) are satisfied for any $n \ge 2$.

Example 3. Consider the following generalized linear problem. Set $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, and let d and δ be the L^p-norm on $\mathbb{R}^d$, that is $d(x, x') = \delta(x, x') = \|x - x'\|_p$ for $p \in [1, \infty]$. Assume that $(A_i)_{i\ge1}$ is a sequence of positive-definite i.i.d. random matrices such that $\mathbb{E}A_i = A \in \mathbb{R}^{d\times d}$, $\lambda^{(p)}_{\min}(A_1) \ge \lambda$ almost surely for some positive constant λ and $\|\lambda^{(p)}_{\max}(A_1)\|_\infty < \infty$, and that $(B_i)_{i\ge1}$ is a sequence of i.i.d. random vectors such that $\mathbb{E}B_i = B \in \mathbb{R}^d$. Here, for any given i, A_i and B_i may not be independent. These sequences are observed. We want to find x^*, which is solution of the following equation:

$$Ax = B. \tag{1.8}$$

To obtain the sequence of estimates $(\overline{X}_n)_{n\geq 1}$ of the solution x^*, the following recursive algorithm will be applied: for any $n \geq 2$,

$$X_n = X_{n-1} - \frac{\gamma}{n^\alpha} Y_n, \qquad Y_n = A_{n-1}X_{n-1} - B_{n-1} + \eta_n, \tag{1.9}$$

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{1.10}$$

where $\gamma \in (0,\infty)$ and $\alpha \in [0,1)$ are constants, and $X_1 \in \mathbb{R}^d$ can be an arbitrary deterministic point or a random point independent of $(A_n)_{n\geq 1}$ and $(B_n)_{n\geq 1}$. Here $A_{n-1}X_{n-1} - B_{n-1}$ is the prediction residual, and $\eta_n \in \mathbb{R}^d$ is a random disturbance independent of $(A_n)_{n\geq 1}$ and $(B_n)_{n\geq 1}$ (the distinction between B_i and η_i is kept because in applications, the user might add a random perturbation η_i to the noise of the gradient $B - B_i$).

In the special case $A_i \equiv A$ and $B_i \equiv B$ for any i, the generalized linear problem becomes the usual linear problem, see Polyak and Juditsky [20]. More generally, when B_i can be random and $A_i \equiv A$, this example matches our framework. Indeed, put $\varepsilon_i = \eta_i - B_{i-1}$, the conditions (1.1)–(1.3) are satisfied with

$$F_n(x,y) = F_n(x,y) = (I_d - \frac{\gamma}{n^\alpha}A)x - \frac{\gamma}{n^\alpha}y, \quad \rho_n = 1 - \frac{\gamma \lambda_{\min}^{(p)}(A)}{n^\alpha}, \quad \tau_n = \frac{\gamma}{n^\alpha} \text{ and } \xi_n \equiv 0$$

for any $n \geq 2$.

In the general case, A_i and B_i are random, so we will enrich the variable ε_i by $\varepsilon_i = (A_{i-1}, \eta_i - B_{i-1})$. The conditions might not be satisfied in this case. However, it is quite common to seek for the best approximation of x^* in the set $\mathcal{C} = \{x : \|x\|_2 \leq D\}$, that is $x^* \in \mathcal{C}$ and

$$\|Ax^* - B\|_2 = \min_{x\in\mathcal{C}} \|Ax - B\|_2. \tag{1.11}$$

In this case, it is natural to add a projection step on $\mathcal{C}$. We then focus on $p = 2$. Let $\Pi_{\mathcal{C}} : \mathbb{R}^d \to \mathcal{C}$ denote the orthogonal projection on $\mathcal{C}$. Note that $\Pi_{\mathcal{C}}$ is such that $\|\Pi_{\mathcal{C}}x - \Pi_{\mathcal{C}}y\|_2 \leq \|x - y\|_2$ for any $x, y \in \mathbb{R}^d$ and $\Pi_{\mathcal{C}}x = x$ for any $x \in \mathcal{C}$. Then, for any $n \geq 2$, take

$$X_n = \Pi_{\mathcal{C}}\Big[X_{n-1} - \frac{\gamma}{n^\alpha}Y_n\Big], \qquad Y_n = A_{n-1}X_{n-1} - B_{n-1} + \eta_n, \tag{1.12}$$

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{1.13}$$

The conditions (1.1)–(1.3) are satisfied, for $y = (M, u)$, with

$$F_n(x,y) = \Pi_{\mathcal{C}}\left[(I_d - \frac{\gamma}{n^\alpha}M)\Pi_{\mathcal{C}}[x] - \frac{\gamma}{n^\alpha}u\right], \quad \rho_n = 1 - \frac{\gamma\lambda}{n^\alpha}, \quad \tau_n = \frac{\gamma}{n^\alpha} \text{ and } \xi_n = \frac{2D\gamma\|\lambda_{\max}^{(2)}(A_1)\|_\infty}{n^\alpha}$$

for any $n \geq 3$ and $\alpha \in (0,1)$. For the case $\alpha = 0$, the conditions (1.1)–(1.3) are also satisfied, but with an additional assumption that $\gamma < 1/\lambda$.

Example 4. For the usual linear problem (cf. equation (1.8)), another recursive algorithm may be applied: for any $n \geq 2$,

$$X_n = X_{n-1} - \gamma Y_n, \qquad Y_n = AX_{n-1} - B + \frac{1}{n^\alpha}\varepsilon_n, \tag{1.14}$$

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \tag{1.15}$$

where $\alpha \in (0,1)$, $\gamma \in (0,\infty)$ is a constant such that $\gamma\lambda_{\min}^{(p)}(A) \in (0,1)$, and $X_1 \in \mathbb{R}^d$ can be an arbitrary deterministic point or a random point independent of $(\varepsilon_n)_{n\geq 2}$. Let d and δ be the L^p-norm on $\mathbb{R}^d$. In this recursive algorithm, the conditions (1.1)–(1.3) are satisfied with

$$F_n(x,y) = (I_d - \gamma A)x + \gamma B - \frac{\gamma}{n^\alpha}y, \quad \rho_n = 1 - \gamma\lambda_{\min}^{(p)}(A), \quad \tau_n = \frac{\gamma}{n^\alpha} \quad \text{and} \quad \xi_n \equiv 0$$

for any $n \geq 2$.

Example 5. A third recursive algorithm for the usual linear problem is given by: for any $n \geq 2$,

$$X_n = X_{n-1} - \frac{\gamma}{n^\alpha}Y_n + \varepsilon_n, \qquad Y_n = AX_{n-1} - B, \tag{1.16}$$

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \tag{1.17}$$

where $\alpha \in (0,1)$, $\gamma \in (0,\infty)$ are constants such that $\gamma\lambda_{\min}^{(p)}(A) \in (0,1)$, and $X_1 \in \mathbb{R}^d$ can be an arbitrary deterministic point or a random point independent of $(\varepsilon_n)_{n\geq 2}$. Let d and δ be the L^p-norm on $\mathbb{R}^d$. In this recursive algorithm, the conditions (1.1)–(1.3) are satisfied with

$$F_n(x,y) = (I_d - \frac{\gamma}{n^\alpha}A)x + \frac{\gamma}{n^\alpha}B - \gamma y, \quad \rho_n = 1 - \frac{\gamma\lambda_{\min}^{(p)}(A)}{n^\alpha}, \quad \tau_n \equiv \gamma \quad \text{and} \quad \xi_n \equiv 0$$

for any $n \geq 2$.

Example 6. We extend the previous examples to optimization of non-linear functions. We still consider $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and focus on the L^2 norm in this example. In machine learning, we need to minimize a function involving a large number of differentiable terms $L(x) = \sum_{i=1}^N \ell_i(x)$ on the set $\mathcal{C} = \{x : \|x\|_2 \leq D\}$. A popular strategy to this end is to use the projected stochastic gradient descent (SGD): for any $n \geq 2$,

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

where $\alpha \in (0,1]$, $\Pi_{\mathcal{C}} : \mathbb{R}^d \to \mathcal{C}$ denote the orthogonal projection on $\mathcal{C}$, J_n is drawn uniformly among all the subsets of $\{1,\ldots,N\}$ with cardinality M and

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

Note that $\mathbb{E}[\hat{\nabla}_{J_n}L(x)] = \nabla L(x)$ for any x. More generally, the stochastic gradient Langevin descent (SGLD) is given by

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

for some i.i.d. sequence η_n of random perturbations added by the user. For $y = (J, u)$, define

$$F_n(x, y) = \Pi_{\mathcal{C}}\Big[x - \frac{\gamma}{n^\alpha}\hat{\nabla}_J L(x) - \frac{\gamma}{n^\alpha}u\Big].$$

If we define $\varepsilon_i = (J_i, \eta_i)$, then this example fits (1.1). It is easy to see that

$$\begin{aligned}\|F_n(x, y) - F_n(x', y)\|_2^2 \ &\leq \ \|x - \frac{\gamma}{n^\alpha}\hat{\nabla}_{J_n} L(x) - x' - \frac{\gamma}{n^\alpha}\hat{\nabla}_{J_n} L(x')\|_2^2 \\ &= \|x - x'\|_2^2 + \left\|\frac{\gamma}{n^\alpha}\hat{\nabla}_{J_n} L(x) - \frac{\gamma}{n^\alpha}\hat{\nabla}_{J_n} L(x')\right\|_2^2 - 2(x - x')^T\left(\frac{\gamma}{n^\alpha}\hat{\nabla}_{J_n} L(x) - \frac{\gamma}{n^\alpha}\hat{\nabla}_{J_n} L(x')\right).\end{aligned}$$

Common assumptions are that the ℓ_i's are m-strongly convex, $m > 0$, which gives

$$(x - x')^T\left(\hat{\nabla}_{J_n} L(x) - \hat{\nabla}_{J_n} L(x')\right) \geq m\|x - x'\|_2^2$$

and that their gradients are ℓ-Lipschitz, that is,

$$\|\hat{\nabla}_{J_n} L(x) - \hat{\nabla}_{J_n} L(x')\|_2^2 \leq \ell^2\|x - x'\|_2^2.$$

We then obtain

$$\begin{aligned}\|F_n(x, y) - F_n(x', y)\|_2^2 &\leq \left(1 - \frac{2m\gamma}{n^\alpha} + \frac{\ell^2\gamma^2}{n^{2\alpha}}\right)\|x - x'\|_2^2 \\ &= \left[\left(1 - \frac{m\gamma}{n^\alpha}\right)^2 + \frac{(\ell^2 - m^2)\gamma^2}{n^{2\alpha}}\right]\|x - x'\|_2^2\end{aligned}$$

and thus the condition (1.2) is satisfied with

$$\rho_n = \sqrt{\left(1 - \frac{m\gamma}{n^\alpha}\right)^2 + \frac{(\ell^2 - m^2)\gamma^2}{n^{2\alpha}}}.$$

Note that $\rho_n - 1 \sim m\gamma/n^\alpha$. So for any n large enough, we have for example

$$\rho_n \leq 1 - \frac{m\gamma}{2n^\alpha}.$$

The condition also holds in the case $\alpha = 0$, but with an additional assumption that $2m\gamma - \ell^2\gamma^2 \in (0, 1)$, that can always be achieved with an adequate choice of γ. Finally, let us assume that $\|\nabla\ell_i(x)\|_2 \leq B$ for any $x \in \mathcal{C}$ and some $B > 0$. Condition (1.3) is satisfied with

$$\tau_n = \frac{\gamma}{n^\alpha} \ \text{ and } \ \xi_n = \frac{2B\gamma}{n^\alpha}$$

for any n large enough.

Example 7. Our final example illustrates that non-homogeneity can appear even in the context of a time homogeneous chain, if it is only observed at non evenly spaced dates $t_1, t_2, \dots$. Assume $(t_i)_{i\geq 1}$ is an increasing sequence, put $k_1 = t_1$ and $k_i = t_i - t_{i-1} > 0$ for any $i > 1$. Consider $F_n \equiv F$, $\rho_n \equiv \rho$, $\tau_n \equiv \tau$, $\xi_n \equiv 0$ and $(X_i)_{i\geq 1}$ the corresponding chain in (1.1)–(1.3). Assume that only the subsequence

$(X_{t_i})_{i\geq 1}$ is observed. Let $\varepsilon_i^{(Z)} = (\varepsilon_{j,i})_{j\geq 1}$ be an i.i.d. copy of the sequence $(\varepsilon_i)_{i\geq 1}$, and define, for any $n \geq 1$ and $y = (y_i)_{i\geq 1}$,

$$F_n^{(Z)}(x, y) = F(F(\dots F(F(x, y_1), y_2), \dots, y_{k_n-1}), y_{k_n})$$

(F_n only depends on the k_n first terms of the sequence y). It is clear that $Z_n = F_n^{(Z)}(Z_{n-1}, \varepsilon_n^{(Z)})$ admits the same distribution as X_{t_n}. Then with the notations above, $\rho_n^{(Z)} = \rho^{k_n}$, and this quantity tends to 0 as $n \to \infty$ if $k_n \to \infty$, which corresponds to a situation where sampling times become rarer and rarer. The expression of $\tau_n^{(Z)}$ is not clear in general. Let us now restrict our attention to the additive model in (1.4) (but note that even if (1.4) holds for $(X_i)_{i\geq 1}$, in general a similar expression does not hold for $(Z_i)_{i\geq 1}$). In this case,

$$|F_n^{(Z)}(x, y) - F_n^{(Z)}(x, y')| \leq \tau \sum_{i=1}^{n} \rho^{n-i} \delta(y_i, y_i'), \tag{1.18}$$

so, for example with the sup metric on $\mathcal{Y}^{\mathbb{N}_*}$, given by $\sup_{i\in\mathbb{N}_*} \delta(y_i, y_i')$, we obtain $\tau_n^{(Z)} = \frac{\tau(1-\rho^n)}{1-\rho} \leq \frac{\tau}{1-\rho}$.

2. Lipschitz functions of random vectors $X_1, \dots, X_n$

We remind that $\|\cdot\|$ is a norm on $\mathbb{R}^d$. Let $f : \mathcal{X}^n \mapsto \mathbb{R}^d$ be a separately Lipschitz function, that is

$$\|f(x_1, x_2, \dots, x_n) - f(x_1', x_2', \dots, x_n')\| \leq d(x_1, x_1') + d(x_2, x_2') + \dots + d(x_n, x_n'). \tag{2.1}$$

Let $\mathbb{P}_{X_1}$ and $\mathbb{P}_\varepsilon$ be the distributions of X_1 and ε, respectively. Assume that $\|\cdot\|$ satisfies

$$\Big\| \int h(x)\mathbb{P}_{X_1}(dx) \Big\| \leq \int \|h(x)\|\mathbb{P}_{X_1}(dx) \quad \text{and} \quad \Big\| \int h(x)\mathbb{P}_\varepsilon(dx) \Big\| \leq \int \|h(x)\|\mathbb{P}_\varepsilon(dx) \tag{2.2}$$

for any measurable function $h : \mathcal{X}^n \mapsto \mathbb{R}^d$. Clearly, if $\|\cdot\| = \|\cdot\|_p, p \in [1, \infty]$, then the condition (2.2) is satisfied.

Let

$$S_n := f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]\,. \tag{2.3}$$

Denote $(\mathcal{F}_k)_{k\geq 0}$ the natural filtration of the chain $(X_k)_{k\geq 1}$, that is $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and for any $k \in \mathbb{N}^*$, $\mathcal{F}_k = \sigma(X_1, X_2, \dots, X_k)$. For any $k \in [0, n]$, define

$$g_k(X_1, \dots, X_k) = \mathbb{E}[f(X_1, \dots, X_n)|\mathcal{F}_k] \tag{2.4}$$

and for any $k \in [1, n]$,

$$M_k = g_k(X_1, \dots, X_k) - g_{k-1}(X_1, \dots, X_{k-1})\,. \tag{2.5}$$

Then $(M_k, \mathcal{F}_k)_{1\leq k\leq n}$ is a finite sequence of martingale differences. For any $k \in [1, n-1]$, let then

$$S_k := M_1 + M_2 + \dots + M_k\,,$$

and note that S_n is already introduced in (2.3) and satisfies $S_n = M_1 + M_2 + \dots + M_n$. Then $(S_k, \mathcal{F}_k)_{1\leq k\leq n}$ is a martingale.

The following proposition gives some interesting properties of the functions $(g_k)_{1\leq k\leq n}$ and of the martingale differences $(M_k, \mathcal{F}_k)_{1\leq k\leq n}$. In this paper, we focus on the case $n_0 = 2$, where n_0 is given by the conditions (1.2) and (1.3).

Lemma 2.1. *For any $k \in [1, n]$ and ρ_k in $[0, 1)$, let*

$$K_{k,n} = 1 + \rho_{k+1} + \rho_{k+1}\rho_{k+2} + \cdots + \rho_{k+1}\rho_{k+2}\cdots\rho_n, \quad k \in [1, n-1] \ \text{ and } \ K_{n,n} = 1.$$

Let $(X_i)_{i\geq 1}$ be a Markov chain satisfying (1.1) for some functions $(F_n)_{n\geq 1}$ satisfying (1.2). We also assume that $\|\cdot\|$ satisfies (2.2). Let g_k and M_k be defined by (2.4) and (2.5), respectively.

1. *The function g_k is separately Lipschitz and such that*

$$\|g_k(x_1, x_2, \ldots, x_k) - g_k(x_1', x_2', \ldots, x_k')\| \leq d(x_1, x_1') + \cdots + d(x_{k-1}, x_{k-1}') + K_{k,n} d(x_k, x_k').$$

2. *Let G_{X_1} and $H_{k,\varepsilon}$ be functions defined by*

$$G_{X_1}(x) = \int d(x, x')\mathbb{P}_{X_1}(dx')$$

and

$$H_{k,\varepsilon}(x, y) = \int d(F_k(x, y), F_k(x, y'))\mathbb{P}_\varepsilon(dy') \,, \quad k \in [2, n],$$

respectively. Then, the martingale difference M_k satisfies that

$$\|M_1\| \leq K_{1,n} G_{X_1}(X_1)$$

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

$$\|M_k\| \leq K_{k,n} H_{k,\varepsilon}(X_{k-1}, \varepsilon_k) \,.$$

3. *Assume moreover that F_n satisfies (1.3), and let G_ε be the function defined by*

$$G_\varepsilon(y) = \int \delta(y, y')\mathbb{P}_\varepsilon(dy') \,.$$

Then $H_{k,\varepsilon}(x, y) \leq \tau_k G_\varepsilon(y)$, and consequently, for any $k \in [2, n]$,

$$\|M_k\| \leq K_{k,n}[\tau_k G_\varepsilon(\varepsilon_k) + \xi_k] \,.$$

4. *Assume moreover that there exist three constants $\alpha \in [0, 1)$, $\rho \in (0, 1)$ and $\eta \in (0, \infty)$ such that for any $n \geq 2$,*

$$\rho_n \leq 1 - \rho/n^\alpha \quad \text{and} \quad \max\{\xi_n, \tau_n\} \leq \eta/n^\alpha. \tag{2.6}$$

Then $K_{1,n} = O(1)$ and $(K_{k,n}[\tau_k + \xi_k])_{k\geq 1}$ is uniformly bounded for all k and n.

5. *Assume moreover that there exist three constants $\alpha \in (0, 1)$, $\rho \in (0, 1)$ and $\eta \in (0, \infty)$ such that for any $n \geq 2$,*

$$\rho_n \leq 1 - \rho/n^\alpha \quad \text{and} \quad \max\{\xi_n, \tau_n\} \leq \eta. \tag{2.7}$$

Then $K_{1,n} = O(1)$ and $K_{k,n}[\tau_k + \xi_k] = O(k^\alpha)$ as $k \to \infty$.

6. *Assume moreover that there exist three constants $\alpha \in (0, 1]$, $\rho \in (0, 1)$ and $\eta \in (0, \infty)$ such that for any $n \geq 2$,*

$$\rho_n \leq \rho \quad \text{and} \quad \max\{\xi_n, \tau_n\} \leq \eta/n^\alpha. \tag{2.8}$$

Then $K_{1,n} = O(1)$ and $K_{k,n}\tau_k = O(k^{-\alpha})$ as $k \to \infty$.

Remark 2.1. *Let us comment on the point 4 of Lemma 2.1. If* $\xi_n \equiv \alpha = 0$*, then* $K_{k,n} \leq \sum_{i=0}^{n-k}(1-\rho)^i < \frac{1}{\rho}$ *and* $\tau_k \leq \eta$ *for any* $k \in [1, n]$ *and* n*. Thus* $(K_{k,n}\tau_k)_{k\geq 1}$ *is uniformly bounded for all* k *and* n*, which has been proved by Proposition 2.1 of Dedecker and Fan [8].*

Remark 2.2. *Let us return to the examples in Subsection 1.3. It is easy to see that Examples 1 and 7 satisfy the condition (2.6) with* $\alpha = 0$*. Examples 2 and 5 satisfy the condition (2.7). Examples 3 and 6 satisfy the condition (2.6). Example 4 satisfies the condition (2.8).*

Proof. The first point will be proved by recurrence in the backward sense. The result is obvious for $k = n$, since $g_n = f$. Assume that it is true at step $k \in [2, n]$, and let us prove it at step $k-1$. By definition

$$g_{k-1}(X_1, \ldots, X_{k-1}) = \mathbb{E}[g_k(X_1, \ldots, X_k)|\mathcal{F}_{k-1}] = \int g_k(X_1, \ldots, X_{k-1}, F_k(X_{k-1}, y))\mathbb{P}_\varepsilon(dy)\,.$$

By assumption (2.2), it follows that

$$\begin{aligned}
&\|g_{k-1}(x_1, x_2, \ldots, x_{k-1}) - g_{k-1}(x'_1, x'_2, \ldots, x'_{k-1})\| \\
&\quad= \left\| \int g_k(x_1, x_2, \ldots, x_{k-1}, F_k(x_{k-1}, y)) - g_k(x'_1, x'_2, \ldots, x'_{k-1}, F_k(x'_{k-1}, y))\mathbb{P}_\varepsilon(dy) \right\| \\
&\quad\leq \int \|g_k(x_1, x_2, \ldots, x_{k-1}, F_k(x_{k-1}, y)) - g_k(x'_1, x'_2, \ldots, x'_{k-1}, F_k(x'_{k-1}, y))\|\mathbb{P}_\varepsilon(dy) \\
&\quad\leq d(x_1, x'_1) + \cdots + d(x_{k-1}, x'_{k-1}) + K_{k,n}\int d(F_k(x_{k-1}, y), F_k(x'_{k-1}, y))\mathbb{P}_\varepsilon(dy) \\
&\quad\leq d(x_1, x'_1) + \cdots + d(x_{k-2}, x'_{k-2}) + (1 + \rho_k K_{k,n})d(x_{k-1}, x'_{k-1}) \\
&\quad\leq d(x_1, x'_1) + \cdots + d(x_{k-2}, x'_{k-2}) + K_{k-1,n}d(x_{k-1}, x'_{k-1}),
\end{aligned} \tag{2.9}$$

which completes the proof of the point 1.

Let us give a proof for the point 2. First note that

$$\|M_1\| = \left\|g_1(X_1) - \int g_1(x)\mathbb{P}_{X_1}(dx)\right\| \leq K_{1,n}\int d(X_1, x)\mathbb{P}_{X_1}(dx) = K_{1,n}G_{X_1}(X_1)\,. \tag{2.10}$$

In the same way, for any $k \in [2, n]$,

$$\begin{aligned}
\|M_k\| &= \big\|g_k(X_1, \cdots, X_k) - \mathbb{E}[g_k(X_1, \cdots, X_k)|\mathcal{F}_{k-1}]\big\| \\
&\leq \int \big\|g_k(X_1, \cdots, F_k(X_{k-1}, \varepsilon_k)) - g_k(X_1, \cdots, F_k(X_{k-1}, y))\big\|\mathbb{P}_\varepsilon(dy) \\
&\leq K_{k,n}\int d(F_k(X_{k-1}, \varepsilon_k), F_k(X_{k-1}, y))\mathbb{P}_\varepsilon(dy) = K_{k,n}H_{k,\varepsilon}(X_{k-1}, \varepsilon_k)\,.
\end{aligned} \tag{2.11}$$

The point 3 is clear, since if (1.3) is true, then

$$H_{k,\varepsilon}(x, y) = \int d(F_k(x, y), F_k(x, y'))\mathbb{P}_\varepsilon(dy') \leq \int \tau_k\delta(y, y')\mathbb{P}_\varepsilon(dy') + \xi_k = \tau_k G_\varepsilon(y) + \xi_k.$$

Consequently, for any $k \in [2, n]$,

$$\|M_k\| \leq K_{k,n}[\tau_k G_\varepsilon(\varepsilon_k) + \xi_k].$$

Next we give a proof for the point 4. It is easy to see that

$$\begin{aligned}
\ln(\rho_{k+1}\cdots\rho_{k+l}) &= \ln\rho_{k+1} + \cdots + \ln\rho_{k+l} \\
&\leq -\rho\sum_{i=k+1}^{k+l} i^{-\alpha} = -\rho\, n^{1-\alpha}\sum_{i=k+1}^{k+l}(\frac{i}{n})^{-\alpha}n^{-1} \\
&\leq -\frac{\rho}{1-\alpha}n^{1-\alpha}\Big((\frac{k+l}{n})^{1-\alpha} - (\frac{k+1}{n})^{1-\alpha}\Big) \\
&\leq -\frac{\rho}{1-\alpha}n^{1-\alpha}\frac{1}{1-\alpha}\frac{l-1}{n}\frac{n^\alpha}{(k+l)^\alpha} \\
&= -\frac{\rho}{(1-\alpha)^2}\frac{l-1}{(k+l)^\alpha}.
\end{aligned} \tag{2.12}$$

Thus, we deduce that

$$\begin{aligned}
K_{k,n}\tau_k &\leq \frac{\eta}{k^\alpha}\sum_{l=1}^{n-k}\exp\Big\{-\frac{\rho}{(1-\alpha)^2}\frac{l-1}{(k+l)^\alpha}\Big\} \\
&\leq \frac{\eta}{k^\alpha}\Big(\sum_{l=1}^{k}\exp\Big\{-\frac{\rho}{(1-\alpha)^2}\frac{l-1}{(k+l)^\alpha}\Big\} + \sum_{l=k}^{\infty}\exp\Big\{-\frac{\rho}{(1-\alpha)^2}\frac{l-1}{(k+l)^\alpha}\Big\}\Big) \\
&=: \frac{\eta}{k^\alpha}(I_1 + I_2).
\end{aligned}$$

For I_1, we have the following estimation

$$\begin{aligned}
I_1 &\leq \exp\Big\{\frac{\rho}{(1-\alpha)^2}\frac{1}{(k+1)^\alpha}\Big\}\sum_{l=1}^{k}\exp\Big\{-\frac{\rho}{2^\alpha(1-\alpha)^2}\frac{l}{k}k^{1-\alpha}\Big\} \\
&\leq k\exp\Big\{\frac{\rho}{(1-\alpha)^2}\frac{1}{(k+1)^\alpha}\Big\}\int_0^1\exp\Big\{-\frac{\rho k^{1-\alpha}}{2^\alpha(1-\alpha)^2}x\Big\}dx \\
&= k\exp\Big\{\frac{\rho}{(1-\alpha)^2}\frac{1}{(k+1)^\alpha}\Big\}\frac{2^\alpha(1-\alpha)^2}{\rho k^{1-\alpha}}\Big(1-\exp\Big\{-\frac{\rho k^{1-\alpha}}{2^\alpha(1-\alpha)^2}\Big\}\Big) \\
&\leq k^\alpha\exp\Big\{\frac{\rho}{(1-\alpha)^2}\Big\}\frac{2^\alpha(1-\alpha)^2}{\rho}.
\end{aligned}$$

It is obvious that I_2 is bounded for $\alpha \in [0, 1)$. Hence $(K_{k,n}\tau_k)_{k\geq 1}$ is uniformly bounded for all k and n and $K_{1,n} = O(1)$.

For the point 5, from the proof of the point 4, it is easy to see that $K_{1,n} = O(1)$ and $K_{k,n}\tau_k = O(k^\alpha), k \to \infty$.

For the point 6, it is easy to see that $K_{k,n} \leq \sum_{i=0}^{n-k}\rho^i < \frac{1}{1-\rho}$ and $\tau_k \leq \eta/k^\alpha$ for all $k \in [1, n]$. Thus $K_{1,n} = O(1)$ and $K_{k,n}\tau_k = O(1/k^\alpha)$ as $k \to \infty$. The proof of the lemma is now complete. □

3. Deviation inequalities for S_n with L^p-norm

Let $n \geq 2$. In this section, we are interested in the concentration properties of S_n under the L^p-norm $\|\cdot\|_p$, where $(X_i)_{i\geq 1}$ is a Markov chain satisfying (1.1) for some functions $(F_n)_{n\geq 1}$ satisfying (1.2) and (1.3). Clearly, it holds for any $p \in [1, \infty]$,

$$\|x\|_p = \Big(\sum_{i=1}^{d} |x^{(i)}|^p\Big)^{1/p} \leq \Big(\sum_{i=1}^{d} \|x\|_\infty^p\Big)^{1/p} = d^{1/p}\|x\|_\infty. \tag{3.1}$$

When $x^{(i)} = x^{(1)}$ for all $i \in [1, d]$, the inequality is actually an equality. Set $S_{2,n} = S_n - M_1$. By (3.1), we have for any $p \in [1, \infty]$ and any $x > 0$,

$$\begin{aligned}\mathbb{P}\Big(\|S_n\|_p \geq d^{1/p}x\Big) &\leq \mathbb{P}\Big(\|M_1\|_p \geq d^{1/p}x/2\Big) + \mathbb{P}\Big(\|S_{2,n}\|_p \geq d^{1/p}x/2\Big)\\ &\leq \mathbb{P}\Big(\|M_1\|_\infty \geq x/2\Big) + \mathbb{P}\Big(\|S_{2,n}\|_\infty \geq x/2\Big)\\ &\leq \sum_{i=1}^{d} \mathbb{P}\Big(|M_1^{(i)}| \geq x/2\Big) + \sum_{i=1}^{d} \mathbb{P}\Big(|S_{2,n}^{(i)}| \geq x/2\Big)\\ &\leq d \max_{1\leq i\leq d} \mathbb{P}\Big(|M_1^{(i)}| \geq x/2\Big) + d \max_{1\leq i\leq d} \mathbb{P}\Big(|S_{2,n}^{(i)}| \geq x/2\Big). \end{aligned}\tag{3.2}$$

Using the inequality

$$|M_1^{(i)}| \leq \|M_1\|_p \leq K_{1,n} G_{X_1}(X_1) \tag{3.3}$$

(cf. the point 2 of Lemma 2.1), we have for any $p \in [1, \infty]$ and any $x > 0$,

$$\begin{aligned}\mathbb{P}\Big(\|S_n\|_p \geq d^{1/p}x\Big) &\leq d\, \mathbb{P}\Big(G_{X_1}(X_1) \geq \frac{x}{2K_{1,n}}\Big) + d \max_{1\leq i\leq d} \mathbb{P}\Big(|S_{2,n}^{(i)}| \geq \frac{x}{2}\Big)\\ &=: I_1(x) + I_2(x)\,. \end{aligned}\tag{3.4}$$

Hence, to dominate $\mathbb{P}\big(\|S_n\|_p \geq d^{1/p}x\big)$, we only need to establish deviation inequalities for $I_1(x)$ and $I_2(x)$. The term $I_1(x)$ represents the direct influence of the initial distribution of the chain, and it will be most of the time negligible. For instance, when $X_1 = x_1$ is a deterministic point, we have $G_{X_1}(X_1) = 0$ and $I_1(x) = 0$ for any $x > 0$. The main difficulty is to give an upper bound for $I_2(x)$, which is the purpose of the remaining of this section. By the point 3 of Lemma 2.1, the martingale differences $(M_k)_{k\in[2,n]}$ satisfy for all $i \in [1, d]$,

$$|M_k^{(i)}| \leq \|M_k\|_p \leq K_{k,n}[\tau_k G_\varepsilon(\varepsilon_k) + \xi_k], \quad k \in [2, n]. \tag{3.5}$$

Notice that $S_{2,n}^{(i)} = \sum_{k=2}^{n} M_k^{(i)}$. Since that $G_\varepsilon(\varepsilon_k), k \in [2, n]$, are i.i.d. random variables, (3.5) plays an important role for estimating $\mathbb{P}\big(|S_{2,n}^{(i)}| \geq x/2\big)$ and thus $I_2(x)$. Assuming various moment conditions on $G_\varepsilon(\varepsilon)$, we can obtain different upper bounds for $I_2(x)$.

In the sequel, denote by $c_{p,d}$ and $c'_{p,d}$ positive constants, which may depend on the constants p, d, α, ρ and η but do not depend on x and n.

3.1. Bernstein type bound

In this subsection, we are interested in establishing a deviation inequality for S_n under the Bernstein condition. We refer to de la Peña [7] for related inequalities: in this paper similar tight Bernstein type inequalities for martingales are proved. Using Lemma 2.1, we get the following proposition.

Proposition 3.1. *Let $p \in [1, \infty]$. Assume that there exist two positive constants H_1 and A_1 such that for any integer $k \geq 2$,*

$$\mathbb{E}\big[\big(G_\varepsilon(\varepsilon)\big)^k\big] \leq \frac{k!}{2} H_1^{k-2} A_1 \, . \tag{3.6}$$

Denote

$$V_n^2 = (1 + A_1) \sum_{k=2}^{n} \big(2K_{k,n}(\tau_k + \xi_k)\big)^2 \qquad \text{and} \qquad \delta_n = \max\{2K_{k,n}(\tau_k H_1 + \xi_k), k = 2, \dots, n\}.$$

Then for any $x > 0$,

$$\mathbb{P}\Big(\|S_n\|_p \geq d^{1/p} x\Big) \;\leq\; I_1(x) + 2d \exp\left\{ -\frac{(x/2)^2}{V_n^2 (1 + \sqrt{1 + x\delta_n / V_n^2}) + \delta_n x/2} \right\} \tag{3.7}$$

$$\leq\; I_1(x) + 2d \exp\left\{ -\frac{(x/2)^2}{2\, V_n^2 + \delta_n x} \right\} . \tag{3.8}$$

Assume moreover that there exits a positive constant c such that for any $x > 0$,

$$\mathbb{P}\big(G_{X_1}(X_1) \geq x\big) \leq c^{-1}\, e^{-c\,x}. \tag{3.9}$$

Then inequality (3.8) implies that:

[i] *If (2.6) is satisfied, then for any $x > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\left(\|S_n\|_p \geq nx\right) \;\leq\; -\,c_{p,d}\,\big(x\mathbf{1}_{\{x\geq 1\}} + x^2 \mathbf{1}_{\{0<x<1\}}\big). \tag{3.10}$$

[ii] *If (2.7) is satisfied with $\alpha \in (0, 1/2)$, then for any $x > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n^{1-2\alpha}} \ln \mathbb{P}\left(\|S_n\|_p \geq nx\right) \;\leq\; -c_{p,d}\, x^2. \tag{3.11}$$

[iii] *If (2.8) is satisfied, then for any $x > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\left(\|S_n\|_p \geq nx\right) \;\leq\; -\,c_{p,d}\, x. \tag{3.12}$$

If either (2.6) or (2.8) holds, then from (3.10) and (3.12), it is easy to see that S_n admits the classical large deviation convergence rate e^{-nc_x}, where $c_x > 0$. Moreover, as $x > x^2$ for $0 < x < 1$, the large deviation convergence rate in (3.12) is better than that in (3.10). Under the condition (2.7), the large deviation convergence rate for S_n becomes much worse, as shown by (3.11).

As mentioned above, when $X_1 = x_1$ a.s. is deterministic, it follows that $G_{X_1}(X_1) = 0$ a.s., and so the condition (3.9) holds for any constant $c \in (0, 1]$.

Since $G_\varepsilon(y) \leq \delta(y, y_0) + \mathbb{E}[\delta(y_0, \varepsilon)]$, it follows that $\mathbb{E}\big[\big(G_\varepsilon(\varepsilon)\big)^k\big] \leq 2^k \mathbb{E}\big[\big(\delta(\varepsilon, y_0)\big)^k\big]$. Hence, the following condition

$$\mathbb{E}\big[\big(\delta(\varepsilon, y_0)\big)^k\big] \leq \frac{k!}{2} A(y_0)^{k-2} B(y_0), \quad k \geq 2, \tag{3.13}$$

implies the condition (3.6) with $H_1 = 2A(y_0)$ and $A_1 = 4B(y_0)$.

In Examples 3, 4 and 5, when $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, we can take $f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^n x_i$, $\|\cdot\| = \|\cdot\|_p$ and $d(x, x') = \delta(x, x') = \|x - x'\|_p$. Then the condition (3.6) is satisfied, provided that

$$\mathbb{E}|\varepsilon^{(i)}|^k \leq \frac{k!}{2} H_1^{k-2} A_1, \quad i \in [1, d] \ \text{ and } \ k \geq 2. \tag{3.14}$$

To show this, by (3.13) with $y_0 = 0$, we only need to prove that

$$\mathbb{E}\|\varepsilon\|_p^k \leq \frac{k!}{2} (H_1 d^{(p+1)/p})^{k-2} A_1 d^{2(p+1)/p}, \quad k \geq 2. \tag{3.15}$$

Clearly, it holds

$$\|x\|_p^k \leq \left(d^{1/p} \sum_{i=1}^d |x^{(i)}| \right)^k \leq d^{k/p} d^{k-1} \sum_{i=1}^d |x^{(i)}|^k, \qquad k \geq 2.$$

Hence, by (3.14), we have

$$\mathbb{E}\|\varepsilon\|_p^k \leq d^{k/p} d^{k-1} \sum_{i=1}^d \mathbb{E}|\varepsilon^{(i)}|^k \leq \frac{k!}{2} (H_1 d^{(p+1)/p})^{k-2} A_1 d^{2(p+1)/p}, \quad k \geq 2,$$

which gives (3.15).

Proof. Notice that $S_{2,n}^{(i)} = \sum_{k=2}^n M_k^{(i)}$ is a sum of martingale differences. By (3.5), we have for any $k \in [2, n]$,

$$|M_k^{(i)}|^j \leq K_{k,n}^j [\tau_k G_\varepsilon(\varepsilon_k) + \xi_k]^j \leq 2^{j-1} K_{k,n}^j [\tau_k^j (G_\varepsilon(\varepsilon_k))^j + \xi_k^j] \leq 2^j K_{k,n}^j [\tau_k^j (G_\varepsilon(\varepsilon_k))^j + \xi_k^j]$$

and so for any $t > 0$,

$$\mathbb{E}\,[e^{tM_k^{(i)}}] \leq 1 + \sum_{j=2}^\infty \frac{t^j}{j!} \mathbb{E}\,[|M_k^{(i)}|^j] \leq 1 + \sum_{j=2}^\infty \frac{t^j}{j!} (2K_{k,n})^j \left[\tau_k^j \mathbb{E}\left[\big(G_\varepsilon(\varepsilon_k)\big)^j\right] + \xi_k^j \right].$$

Using the condition (3.6), we deduce that for any $k \in [2, n]$ and any $t \in [0, \delta_n^{-1})$,

$$\begin{aligned} \mathbb{E}\,[e^{tM_k^{(i)}}] &\leq 1 + \sum_{j=2}^\infty \frac{t^j}{j!} (2K_{k,n})^j \left[\tau_k^j \frac{j!}{2} H_1^{j-2} A_1 + \xi_k^j \right] \\ &= 1 + \frac{1}{2} \sum_{j=2}^\infty t^j (2K_{k,n})^j \left[\tau_k^j H_1^{j-2} A_1 + \frac{2}{j!} \xi_k^j \right] \\ &\leq 1 + \frac{1}{2} t^2 (2K_{k,n})^2 \sum_{j=2}^\infty t^{j-2} \, (2K_{k,n})^{j-2} (A_1 + 1) [\tau_k H_1 + \xi_k]^{j-2} (\tau_k + \xi_k)^2 \\ &\leq 1 + \frac{t^2 (A_1 + 1) \big(2K_{k,n}(\tau_k + \xi_k)\big)^2}{2(1 - t\,\delta_n)}. \end{aligned} \tag{3.16}$$

Applying the inequality $1+u \leq e^u$ for $u \geq 0$ to (3.16), we deduce that for any $k \in [2, n]$ and any $t \in [0, \delta_n^{-1})$,

$$\mathbb{E}\,[e^{tM_k^{(i)}}|\mathcal{F}_{k-1}] \leq \exp\left\{\frac{t^2(1+A_1)\big(2K_{k,n}(\tau_k+\xi_k)\big)^2}{2(1-t\,\delta_n)}\right\}. \tag{3.17}$$

By the tower property of conditional expectation and the last inequality, it is easy to see that for any $n \geq 2$ and any $t \in [0, \delta_n^{-1})$,

$$\begin{aligned}
\mathbb{E}\left[e^{tS_{2,n}^{(i)}}\right] &= \mathbb{E}\left[\mathbb{E}\,[e^{tS_{2,n}^{(i)}}|\mathcal{F}_{n-1}]\right] && (3.18)\\
&= \mathbb{E}\left[e^{tS_{2,n-1}^{(i)}}\mathbb{E}\,[e^{tM_n^{(i)}}|\mathcal{F}_{n-1}]\right]\\
&\leq \mathbb{E}\left[e^{tS_{2,n-1}^{(i)}}\right]\exp\left\{\frac{t^2(1+A_1)\big(2K_{k,n}(\tau_k+\xi_k)\big)^2}{2(1-t\,\delta_n)}\right\}\\
&\leq \exp\left\{\frac{t^2V_n^2}{2(1-t\,\delta_n)}\right\}. && (3.19)
\end{aligned}$$

Clearly, the same bound holds for $\mathbb{E}\left[e^{-tS_{2,n}^{(i)}}\right]$. Returning to (3.4), by the exponential Markov inequality, we have for any $x > 0$ and any $t \in [0, \delta_n^{-1})$,

$$\begin{aligned}
I_2(x) &\leq d\max_{1\leq i\leq d}\mathbb{E}\left[\exp\left\{tS_{2,n}^{(i)}-\frac{1}{2}tx)\right\}+\exp\left\{-tS_{2,n}^{(i)}-\frac{1}{2}tx\right\}\right]\\
&\leq 2d\exp\left\{-\frac{1}{2}t\,x+\frac{t^2V_n^2}{2(1-t\,\delta_n)}\right\}. && (3.20)
\end{aligned}$$

The last bound reaches its minimum at

$$t = t(x) := \frac{x/V_n^2}{x\delta_n/V_n^2+1+\sqrt{1+x\delta_n/V_n^2}}\,.$$

Substituting $t = t(x)$ in (3.20), we obtain for any $x > 0$,

$$\begin{aligned}
I_2(x) &\leq 2d\exp\left\{-\frac{(x/2)^2}{V_n^2(1+\sqrt{1+x\delta_n/V_n^2})+x\delta_n/2}\right\}\\
&\leq 2d\exp\left\{-\frac{(x/2)^2}{2V_n^2+x\delta_n}\right\},
\end{aligned}$$

where the last line follows by the inequality $\sqrt{1+x\,\delta_n/V_n^2} \leq 1+x\,\delta_n/(2V_n^2)$. Applying the upper bounds to (3.4), we obtain the inequalities (3.7) and (3.8).

Condition (3.9) implies that

$$I_1(x) \leq dc^{-1}\exp\left\{-\frac{c\,x}{2K_{1,n}}\right\}. \tag{3.21}$$

If the condition (2.6) is satisfied, then we have $K_{1,n} = O(1)$, $V_n^2 = O(n)$ and $\delta_n = O(1)$ as $n \to \infty$, by the point 4 of Lemma 2.1. Applying (3.21) to (3.8), we deduce that for any $x > 0$,

$$\mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq dc^{-1} \exp\left\{-c'_{p,d}nx\right\} + 2d\exp\left\{-\frac{c'_{p,d}(nx)^2}{c_{p,d}(n+nx)}\right\}.$$

This last inequalitiy implies (3.10).

If the condition (2.7) is satisfied, from the point 5 of Lemma 2.1, then we have $K_{1,n} = O(1)$,

$$V_n^2 = O(1)\sum_{k=2}^{n} k^{2\alpha} = O(1)n^{1+2\alpha}\sum_{k=2}^{n}(\frac{k}{n})^{2\alpha}\frac{1}{n} = O(n^{1+2\alpha})$$

and $\delta_n = O(n^\alpha)$ as $n \to \infty$. Applying (3.21) to (3.8), we deduce that for any $\alpha \in (0, 1/2)$ and any $x > 0$,

$$\mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq dc^{-1} \exp\left\{-c'_{p,d}nx\right\} + 2d\exp\left\{-\frac{c'_{p,d}(nx)^2}{c_{p,d}(n^{1+2\alpha}+nxn^\alpha)}\right\}.$$

The last inequality implies (3.11).

If the condition (2.8) is satisfied, by the point 6 of Lemma 2.1, then we have $K_{1,n} = O(1)$,

$$V_n^2 = O(1)\sum_{k=1}^{n}\frac{1}{k^{2\alpha}} = \begin{cases} O(n^{1-2\alpha}) & \text{if } 0 \leq \alpha < 1/2 \\ O(\ln n) & \text{if } \alpha = 1/2 \\ O(1) & \text{if } 1/2 < \alpha \leq 1 \end{cases} \tag{3.22}$$

and $\delta_n = O(1)$ as $n \to \infty$. Applying (3.21) to (3.8), we deduce that for any $x > 0$,

$$\mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq dc^{-1} \exp\left\{-c'_{p,d}nx\right\} + 2d\exp\left\{-\frac{c'_{p,d}(nx)^2}{c_{p,d}(n^{1-2\alpha}\vee \ln n+nx)}\right\}.$$

The last line implies (3.12). This completes the proof of Proposition 3.1. □

3.2. Semi-exponential bound

If both $G_{X_1}(X_1)$ and $G_\varepsilon(\varepsilon)$ have semi-exponential moments, the following proposition holds. It can be compared to the corresponding results in Borovkov [5] for partial sums of independent random variables, Merlevède *et al.* [17] for partial sums of weakly dependent sequences, Lesigne and Volný [15] and Fan *et al.* [12] for martingales.

Proposition 3.2. *Let $p \in [1, \infty]$ and $q \in (0, 1)$. Assume that there exists a positive constant A_1 such that*

$$\mathbb{E}\left[\left(G_\varepsilon(\varepsilon)\right)^2 \exp\left\{\left(G_\varepsilon(\varepsilon)\right)^q\right\}\right] \leq A_1\,. \tag{3.23}$$

Denote

$$V_n^2 = 2e\sum_{k=2}^{n} K_{k,n}^2\Big(\tau_k^2 A_1 + \xi_k^2\mathbb{E}[e^{|G_\varepsilon(\varepsilon)|^q}]\Big) \quad \textit{and} \quad \delta_n = \max\Big\{K_{k,n}\tau_k, K_{k,n}\xi_k, k = 2, \ldots, n\Big\}.$$

If $V_n \geq 1$, then for any $x > 0$,

$$\mathbb{P}\Big(\|S_n\|_p \geq d^{1/p}x\Big) \ \leq \ I_1(x) + 4d\exp\bigg\{ - \frac{(x/2)^2}{2(V_n^2 + (x/2)^{2-q}\delta_n^q)} \bigg\}. \tag{3.24}$$

Assume moreover that there exits a positive constant c such that for any $x > 0$,

$$\mathbb{P}\big(G_{X_1}(X_1) \geq x\big) \leq c^{-1}\, e^{-c\, x^q}. \tag{3.25}$$

Then inequality (3.24) implies that:

[i] *If (2.6) or (2.8) is satisfied, then for any $x > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n^q} \ln \mathbb{P}\left(\|S_n\|_p \geq nx\right) \ \leq \ -c_{p,d}x^q. \tag{3.26}$$

[ii] *Assume that (2.7) is satisfied with $\alpha \in (0, 1/2)$. If $0 < \alpha < \frac{1-q}{2-q}$, then for any $x > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n^{q(1-\alpha)}} \ln \mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq -c_{p,d}x^q.$$

If $\alpha = \frac{1-q}{2-q}$, then for any $x > 0$,

$$\limsup_{n\to\infty} \frac{1}{n^{q/(2-q)}} \ln \mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq -c_{p,d}\big(x^q \mathbf{1}_{\{x\geq 1\}} + x^2\mathbf{1}_{\{0<x<1\}}\big).$$

If $\frac{1-q}{2-q} < \alpha < \frac{1}{2}$, then for any $x > 0$,

$$\limsup_{n\to\infty} \frac{1}{n^{1-2\alpha}} \ln \mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq -c_{p,d}x^2.$$

If either (2.6) or (2.8) is satisfied, from (3.26), it is easy to see that the large deviation convergence rate is the same as the classical one. On the contrary under the condition (2.7), this convergence rate becomes worsen as α increases.

Proof. Notice that $S_{2,n}^{(i)}/\delta_n = \sum_{k=2}^n M_k^{(i)}/\delta_n$ is a sum of martingale differences. By (3.5) and the condition (3.23), it is easy to see that for any $k \in [2, n]$,

$$\begin{aligned}
\mathbb{E}\,[(M_k^{(i)}/\delta_n)^2 e^{|M_k^{(i)}/\delta_n|^q}] \ &\leq \ \delta_n^{-2}\mathbb{E}\,[(K_{k,n}[\tau_k G_\varepsilon(\varepsilon_k) + \xi_k])^2 e^{|K_{k,n}[\tau_k G_\varepsilon(\varepsilon_k)+\xi_k]/\delta_n|^q}] \\
&\leq \ 2(K_{k,n}\delta_n^{-1})^2\mathbb{E}\,[((\tau_k G_\varepsilon(\varepsilon))^2 + \xi_k^2)e^{|G_\varepsilon(\varepsilon)|^q+1}] \\
&\leq \ 2e(K_{k,n}\delta_n^{-1})^2\Big(\tau_k^2\mathbb{E}\,[(G_\varepsilon(\varepsilon))^2 e^{|G_\varepsilon(\varepsilon)|^q}] + \xi_k^2\mathbb{E}[e^{|G_\varepsilon(\varepsilon)|^q}]\Big) \\
&\leq \ 2e(K_{k,n}\delta_n^{-1})^2\Big(\tau_k^2 A_1 + \xi_k^2\mathbb{E}[e^{|G_\varepsilon(\varepsilon)|^q}]\Big).
\end{aligned}$$

If $V_n\delta_n^{-1} \geq 1$, using inequality (2.7) of Fan *et al.* [12], then we have for any $t > 0$,

$$\mathbb{P}\Big(|S_{2,n}^{(i)}/\delta_n| \geq t\Big) \ \leq \ 4\exp\bigg\{ - \frac{t^2}{2(V_n^2\delta_n^{-2} + t^{2-q})} \bigg\}. \tag{3.27}$$

Substituting $t = x/(2\delta_n)$ in (3.27), we get for any $x > 0$,

$$\mathbb{P}\left(|S_{2,n}^{(i)}| \geq x/2\right) \;\leq\; 4\exp\left\{-\frac{(x/2)^2}{2(V_n^2 + (x/2)^{2-q}\delta_n^q)}\right\}.$$

From (3.4) and the last inequality, we obtain the desired inequality (3.24).

Condition (3.25) implies that

$$I_1(x) \leq dc^{-1}\exp\left\{-c\left(\frac{x}{2K_{1,n}}\right)^q\right\}. \tag{3.28}$$

If the condition (2.6) is satisfied, then by the point 4 of Lemma 2.1, we obtain $K_{1,n} = O(1)$, $V_n^2 = O(n)$ and $\delta_n = O(1)$ as $n \to \infty$. Applying inequality (3.28) together with (3.24), we deduce that for any $x > 0$,

$$\mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq dc^{-1}\exp\left\{-c'_{p,d}(nx)^q\right\} + 4d\exp\left\{-\frac{c'_{p,d}(nx)^2}{c_{p,d}(n + (nx)^{2-q})}\right\}.$$

From the last inequality, we get (3.26).

If the condition (2.7) is satisfied, then by the point 5 of Lemma 2.1, we have $K_{1,n} = O(1)$, $V_n^2 = O(n^{1+2\alpha})$ and $\delta_n = O(n^\alpha)$ as $n \to \infty$. Applying (3.28) to inequality (3.24), we deduce that for any $\alpha \in (0, 1/2)$ and any $x > 0$,

$$\mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq dc^{-1}\exp\left\{-c'_{p,d}(nx)^q\right\} + 4d\exp\left\{-\frac{c'_{p,d}(nx)^2}{c_{p,d}(n^{1+2\alpha} + (nx)^{2-q}n^{q\alpha})}\right\}.$$

The last inequality implies the point [ii] of Proposition 3.2. Note that when $0 < \alpha < \frac{1-q}{2-q}$, we have $n^{2-q+q\alpha} > n^{1+2\alpha}$, while when $\frac{1-q}{2-q} < \alpha < \frac{1}{2}$, then $n^{2-q+q\alpha} < n^{1+2\alpha}$.

If the condition (2.8) is satisfied, then by the point 6 of Lemma 2.1, we have $K_{1,n} = O(1)$, (3.22) and $\delta_n = O(1)$ as $n \to \infty$. Applying again inequalities (3.28) and (3.24), we deduce that for any $x > 0$,

$$\mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq dc^{-1}\exp\left\{-c'_{p,d}(nx)^q\right\} + 4d\exp\left\{-\frac{c'_{p,d}(nx)^2}{c_{p,d}(n^{1-2\alpha} \vee \ln n + (nx)^{2-q})}\right\}.$$

Inequality (3.26) is an easy consequence of this last inequality. □

3.3. Fuk-Nagaev type bound

If the martingale differences $(M_i)_{i\geq 2}$ admit finite qth order moments ($q \geq 2$), then we have the following Fuk-Nagaev type inequality (cf. Corollary 3′ of Fuk [13] and Nagaev [19]).

Proposition 3.3. *Let $p \in [1, \infty]$ and $q \in [2, \infty)$. Assume that there exist two positive constants A_1 and $B_1(q)$ such that*

$$\mathbb{E}\big[\big(G_\varepsilon(\varepsilon)\big)^2\big] \leq A_1 \quad \textit{and} \quad \mathbb{E}\big[\big(G_\varepsilon(\varepsilon)\big)^q\big] \leq B_1(q)\,. \tag{3.29}$$

Denote

$$V_n^2 = 2\sum_{k=2}^n K_{k,n}^2\Big(\tau_k^2 A_1 + \xi_k^2\Big) \qquad \textit{and} \qquad H_n(q) = 2^{q-1}\sum_{k=2}^n K_{k,n}^q\Big(\tau_k^q B_1(q) + \xi_k^q\Big).$$

Then for any $x > 0$,

$$\mathbb{P}\Big(\|S_n\|_p \geq d^{1/p}x\Big) \leq I_1(x) + 2^{q+1}d\Big(1+\frac{2}{q}\Big)^q\frac{H_n(q)}{x^q} + 2d\exp\left\{-\frac{x^2}{2(q+2)^2e^qV_n^2}\right\}. \tag{3.30}$$

Assume moreover that there exists a positive constant c *such that for any* $x > 0$,

$$\mathbb{P}\big(G_{X_1}(X_1) \geq x\big) \leq c\,x^{-q}, \tag{3.31}$$

then inequality (3.30) implies that:

[i] *If (2.6) is satisfied, then for any* $x > 0$,

$$\mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq \frac{c_{p,d}}{x^q}\cdot\frac{1}{n^{q-1}}. \tag{3.32}$$

[ii] *If (2.7) is satisfied with* $0 < \alpha < \frac{1}{2}$, *then for any* $x > 0$,

$$\mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq \frac{c_{p,d}}{x^q}\cdot\frac{1}{n^{q-1-\alpha q}}. \tag{3.33}$$

[iii] *If (2.8) is satisfied, then for any* $x > 0$,

$$\mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq \begin{cases} \dfrac{c_{p,d}}{x^q}\cdot\dfrac{1}{n^{q-1+\alpha q}}, & \textit{if } 0 < \alpha < \frac{1}{q}, \\[2ex] \dfrac{c_{p,d}}{x^q}\cdot\dfrac{\ln n}{n^q}, & \textit{if } \alpha = \frac{1}{q}, \\[2ex] \dfrac{c_{p,d}}{x^q}\cdot\dfrac{1}{n^q}, & \textit{if } \frac{1}{q} < \alpha < 1. \end{cases} \tag{3.34}$$

Under the condition (2.6), then from (3.32), it is easy to see that the large deviation convergence rate is the same as the classical one, which is of order n^{1-q} as $n \to \infty$. When the condition (2.7) is satisfied, from (3.33), we find that the large deviation convergence rate becomes worse when α tends to $1/2$. Moreover, the large deviation convergence rate is slower than the classical one. Now, if the condition (2.8) is satisfied, then the inequalities (3.34) imply that this convergence rate is much better than the classical one.

Proof. By (3.5) and the condition (3.29), it follows that

$$\begin{aligned}\sum_{k=2}^n \mathbb{E}[|M_k^{(i)}|^q|\mathcal{F}_{k-1}] &\leq \sum_{k=2}^n \mathbb{E}[\big(K_{k,n}(\tau_k G_\varepsilon(\varepsilon_k)+\xi_k)\big)^q] = 2^{q-1}\sum_{k=2}^n\Big(\big(K_{k,n}\tau_k\big)^q\mathbb{E}[(G_\varepsilon(\varepsilon))^q] + K_{k,n}^q\xi_k^q\Big) \\ &\leq 2^{q-1}\sum_{k=2}^n K_{k,n}^q\Big(\tau_k^qB_1(q)+\xi_k^q\Big) = H_n(q).\end{aligned}$$

Notice that $H_n(2) = V_n^2$. Using the Corollary 3′ in Fuk [13], we have for any $x > 0$,

$$\mathbb{P}(|S_{2,n}^{(i)}| \geq x/2) \leq 2^{q+1}\Big(1+\frac{2}{q}\Big)^q \cdot \frac{H_n(q)}{x^q} + 2\exp\left\{-\frac{x^2}{2(q+2)^2 e^q V_n^2}\right\}. \tag{3.35}$$

Applying the last inequality to (3.4), we get the first desired inequality.

The condition (3.31) implies that for any $x > 0$,

$$I_1(x) \leq c\, d\Big(\frac{x}{2K_{1,n}}\Big)^{-q}. \tag{3.36}$$

If the condition (2.6) is satisfied, then from the point 4 of Lemma 2.1, we have $K_{1,n} = O(1)$, $V_n^2 = O(n)$ and $H_n(q) = O(n)$ as $n \to \infty$. Applying (3.36) to (3.30), we get for any $x > 0$,

$$\mathbb{P}(\|S_n\|_p \geq nx) \leq \frac{c_{p,d}}{(nx)^q} + \frac{c_{p,d}\, n}{(nx)^q} + 2d\exp\left\{-c'_{p,d}\frac{(nx)^2}{n}\right\}.$$

The inequality (3.32) follows from the last inequality.

If the condition (2.7) is satisfied, by the point 5 of Lemma 2.1, then we have $K_{1,n} = O(1)$, $V_n^2 = O(n^{1+2\alpha})$ and $H_n(q) = O(1)\sum_{k=2}^n k^{\alpha q} = O(n^{1+\alpha q})$ as $n \to \infty$. Applying (3.36) to (3.30), we get for any $x > 0$,

$$\mathbb{P}(\|S_n\|_p \geq nx) \leq \frac{c_{p,d}}{(nx)^q} + \frac{c_{p,d}\, n^{1+\alpha q}}{(nx)^q} + 2d\exp\left\{-c'_{p,d}\frac{(nx)^2}{n^{1+2\alpha}}\right\}.$$

The inequality (3.33) follows from this last inequality. Note that if $\alpha < 1/2$, then the third term in the right hand side of the last inequality tends to 0 as $n \to \infty$.

If the condition (2.8) is satisfied, then by the point 6 of Lemma 2.1, we have $K_{1,n} = O(1)$, (3.22) and

$$H_n(q) = O(1)\sum_{k=1}^n k^{-\alpha q} = \begin{cases} O(n^{1-\alpha q}), & \text{if } 0 \leq \alpha < 1/q, \\ O(\ln n), & \text{if } \alpha = 1/q, \\ O(1), & \text{if } 1/q < \alpha < 1 \end{cases}$$

as $n \to \infty$. Similarly, we prove that the inequality (3.30) implies (3.34). □

3.4. von Bahr-Esseen type bound

If the dominating random variables $G_{X_1}(X_1)$ and $G_\varepsilon(\varepsilon)$ admit only a finite moment with order $q \in [1,2]$, we have the following von Bahr-Esseen type deviation bound.

Proposition 3.4. *Let $p \in [1,\infty]$ and $q \in [1,2]$. Assume that there exists a positive constant $A_1(q)$ such that*

$$\mathbb{E}\big[\big(G_\varepsilon(\varepsilon)\big)^q\big] \leq A_1(q). \tag{3.37}$$

Denote

$$V_n(q) = 2^{q-1}\Big[K_{2,n}^q(\tau_2^q A_1(q) + \xi_2^q) + 2^{2-q}\sum_{k=3}^n K_{k,n}^q\big(\tau_k^q A_1(q) + \xi_k^q\big)\Big].$$

Then for any $x > 0$,

$$\mathbb{P}\Big(\|S_n\|_p \geq d^{1/p}x\Big) \leq I_1(x) + 2^q d\frac{V_n(q)}{x^q}. \tag{3.38}$$

Assume moreover (3.31). Then inequality (3.38) implies that:

[i] *If (2.6) is satisfied, then (3.32) holds.*

[ii] *If (2.7) is satisfied with $0 < \alpha < 1 - \frac{1}{q}$, then (3.33) holds.*

[iii] *If (2.8) is satisfied, then (3.34) holds.*

Remark 3.1. *The constant 2^{2-q} in $V_n(q)$ can be replaced by the more precise constant $\tilde{C}_q$ described in Proposition 1.8 of Pinelis [24].*

Proof. Notice that $S_{2,n}^{(i)} = \sum_{k=2}^n M_k^{(i)}$ is a sum of martingale differences. Using a refinement of the von Bahr-Esseen inequality (cf. Proposition 1.8 of Pinelis [24]), we get for any $q \in [1, 2]$,

$$\mathbb{E}|S_{2,n}^{(i)}|^q \le \mathbb{E}|M_2^{(i)}|^q + 2^{2-q} \sum_{k=3}^n \mathbb{E}|M_k^{(i)}|^q .$$

By (3.5) and (3.37), we deduce that for any $q \in [1, 2]$,

$$\begin{aligned} \mathbb{E}|S_{2,n}^{(i)}|^q &\le K_{2,n}^q \mathbb{E}\big[\big(\tau_2 G_\varepsilon(\varepsilon) + \xi_2\big)^q\big] + 2^{2-q} \sum_{k=3}^n K_{k,n}^q \mathbb{E}\big[\big(\tau_k G_\varepsilon(\varepsilon) + \xi_k\big)^q\big] \\ &\le 2^{q-1}\left[K_{2,n}^q \mathbb{E}\big[\big(\tau_2 G_\varepsilon(\varepsilon)\big)^q + \xi_2^q\big] + 2^{2-q} \sum_{k=3}^n K_{k,n}^q \mathbb{E}\big[\big(\tau_k G_\varepsilon(\varepsilon)\big)^q + \xi_k^q\big]\right] \\ &\le 2^{q-1}\left[K_{2,n}^q \big(\tau_2^q A_1(q) + \xi_2^q\big) + 2^{2-q} \sum_{k=3}^n K_{k,n}^q \big(\tau_k^q A_1(q) + \xi_k^q\big)\right] \\ &= V_n(q). \end{aligned} \tag{3.39}$$

By Markov's inequality, we get for any $x > 0$,

$$\begin{aligned} \mathbb{P}(|S_{2,n}^{(i)}| \ge x/2) &\le 2^q \frac{\mathbb{E}|S_{2,n}^{(i)}|^q}{x^q} \\ &\le 2^q \frac{V_n(q)}{x^q}. \end{aligned} \tag{3.40}$$

Applying the last inequality to (3.4), we get the desired inequality (3.38). The remaining of the proof is similar to the proof of Proposition 3.3. □

Next, we consider the case where the random variables $G_{X_1}(X_1)$ and $G_\varepsilon(\varepsilon)$ have only a weak moment. Recall that for any real-valued random variable Z and any $q \ge 1$, the weak moment of order q is defined by

$$\|Z\|_{w,q}^q = \sup_{x>0} x^q \mathbb{P}(|Z| > x) . \tag{3.41}$$

When the variables $G_{X_1}(X_1)$ and $G_\varepsilon(\varepsilon)$ have only a weak moment of order $q \in (1, 2)$, we have the following deviation inequality.

Proposition 3.5. *Let $p \in [1, \infty]$ and $q \in (1, 2)$. Assume that there exists a positive constant $A_1(q)$ such that*

$$\big\|G_\varepsilon(\varepsilon)\big\|_{w,q}^q \le A_1(q) . \tag{3.42}$$

Then for any $x > 0$,

$$\mathbb{P}\Big(\|S_n\|_p \geq d^{1/p}x\Big) \leq I_1(x) + C_{d,q}\frac{B(n,q)}{x^q}\,, \tag{3.43}$$

where

$$C_{d,q} = 2^{2+q}d\Big(\frac{q}{q-1} + \frac{2}{2-q}\Big) \qquad \text{and} \qquad B(n,q) = \sum_{k=2}^{n}(2K_{k,n})^q(\tau_k^q A_1(q) + \xi_k^q).$$

Assume moreover $\big\|G_{X_1}(X_1)\big\|_{w,q}^q < \infty$. *Then inequality (3.43) implies that:*

[i] *If (2.6) is satisfied, then (3.32) holds.*

[ii] *If (2.7) is satisfied with* $0 < \alpha < 1 - \frac{1}{q}$, *then (3.33) holds.*

[iii] *If (2.8) is satisfied, then (3.34) holds.*

Proof. By Proposition 3.3 of Cuny, Dedecker and Merlevède [6], we have for any $x > 0$,

$$\mathbb{P}\big(|S_{2,n}^{(i)}| \geq x/2\big) \leq \frac{C_q}{x^q}\sum_{k=2}^{n}\|M_k^{(i)}\|_{w,q}^q,$$

where $C_q = 2^{2+q}(\frac{q}{q-1} + \frac{2}{2-q})$. By (3.4), we have for any $x > 0$,

$$\mathbb{P}\big(\|S_n\|_p \geq d^{1/p}x\big) \leq I_1(x) + \frac{C_{d,q}}{x^q}\max_{1\leq i\leq d}\sum_{k=2}^{n}\|M_k^{(i)}\|_{w,q}^q. \tag{3.44}$$

Using (3.5), we have for any $k \in [2,n]$,

$$\begin{aligned}
\|M_k^{(i)}\|_{w,q}^q &\leq \sup_{x>0} x^q\mathbb{P}(K_{k,n}\tau_k G_\varepsilon(\varepsilon) + K_{k,n}\xi_k > x) \\
&\leq (2K_{k,n}\xi_k)^q + \sup_{x>2K_{k,n}} x^q\mathbb{P}(K_{k,n}\tau_k G_\varepsilon(\varepsilon) > x/2) \\
&\leq (2K_{k,n}\xi_k)^q + \sup_{x>0}(2K_{k,n}\tau_k x)^q\mathbb{P}(G_\varepsilon(\varepsilon) > x) \\
&\leq (2K_{k,n}\xi_k)^q + (2K_{k,n}\tau_k)^q\big\|G_\varepsilon(\varepsilon)\big\|_{w,q}^q \\
&\leq (2K_{k,n}\xi_k)^q + (2K_{k,n}\tau_k)^q A_1(q).
\end{aligned}$$

Returning to (3.44), we get for any $x > 0$,

$$\mathbb{P}\big(\|S_n\|_p \geq d^{1/p}x\big) \leq I_1(x) + \frac{C_{d,q}}{x^q}\sum_{k=2}^{n}(2K_{k,n})^q(\xi_k^q + \tau_k^q A_1(q)),$$

which is exactly the first desired inequality. The remaining of the proof is similar to the proof of Proposition 3.3. □

3.5. McDiarmid type bound

In this section, we consider the case where the increments M_k are bounded in L_∞-norm. We shall make use of a refinement of the well-known McDiarmid inequality, which has been recently established by Rio [26]. Following the notations in Rio [26], denote

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

and let

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

be the Young transform of $\ell(t)$. As quoted by Rio [26], for any $x \in [0, 1)$, it holds

$$\ell^*(x) \geq (x^2 - 2x) \ln(1 - x) \geq 2x^2. \tag{3.45}$$

Denote by ε' an independent copy of ε, and X_1' an independent copy of X_1.

Proposition 3.6. *Let $p \in [1, \infty]$. Assume that there exists a positive constant T_1 such that*

$$\|\delta(\varepsilon, \varepsilon')\|_\infty \leq T_1. \tag{3.46}$$

Let

$$V_n^2 = \sum_{k=2}^n K_{k,n}^2 (\tau_k T_1 + \xi_k)^2 \qquad \text{and} \qquad D_n = \sum_{k=2}^n K_{k,n} (\tau_k T_1 + \xi_k).$$

Then, for any $x \in [0, 2D_n]$,

$$\mathbb{P}\Big(\|S_n\|_p \geq d^{1/p} x \Big) \;\leq\; I_1(x) + 2d \exp\left\{ -\frac{D_n^2}{V_n^2}\, \ell^*\Big(\frac{x}{2D_n} \Big) \right\}. \tag{3.47}$$

Consequently, for any $x \in [0, 2D_n]$,

$$\mathbb{P}\Big(\|S_n\|_p \geq d^{1/p} x \Big) \;\leq\; I_1(x) + 2d \left(\frac{D_n - x/2}{D_n} \right)^{\frac{D_n x - (x/2)^2}{V_n^2}} \tag{3.48}$$

$$\leq\; I_1(x) + 2d \exp\left\{ -\frac{x^2}{2V_n^2} \right\}. \tag{3.49}$$

Assume moreover $\|d(X_1, X_1')\|_\infty < \infty$. Then we have:

[i] *If (2.6) is satisfied, then for any $x > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\left(\|S_n\|_p \geq nx \right) \leq -c_{p,d}\, x^2. \tag{3.50}$$

[ii] *If (2.7) is satisfied with $\alpha \in (0, 1/2)$, then for any $x > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n^{1-2\alpha}} \ln \mathbb{P}\left(\|S_n\|_p \geq nx \right) \leq -c_{p,d}\, x^2. \tag{3.51}$$

[iii] *Assume the condition (2.8). If $0 < \alpha < 1/2$, then for any $x > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n^{1+2\alpha}} \mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq -c_{p,d}\, x^2 \,.$$

If $\alpha = 1/2$, then for any $x > 0$,

$$\limsup_{n\to\infty} \frac{\ln n}{n^2} \mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq -c_{p,d}\, x^2 \,.$$

If $1/2 < \alpha < 1$, then for any $x > 0$,

$$\limsup_{n\to\infty} \frac{1}{n^2} \mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq -c_{p,d}\, x^2 \,.$$

Proof. For any $k \in [2, n]$, let

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

and

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

From the proof of Proposition 2.1, it follows that for any $k \in [2, n]$,

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

By the proof of Proposition 2.1 and the condition (3.46), we have

$$v_{k-1}^{(i)}(X_1, \ldots, X_{k-1}) - u_{k-1}^{(i)}(X_1, \ldots, X_{k-1}) \leq K_{k,n}(\tau_k T_1 + \xi_k)\,, \quad k \in [2, n].$$

Now, with an argument similar to the proof of Theorem 3.1 of Rio [26] with $\Delta_k = K_{k,n}(\tau_k T_1 + \xi_k), k \in [2, n]$, we get for any $x \in [0, 2D_n]$,

$$\mathbb{P}\big(|S_{2,n}^{(i)}| \geq x/2\big) \;\leq\; 2\exp\left\{-\frac{D_n^2}{V_n^2}\, \ell^*\Big(\frac{x}{2D_n}\Big)\right\}\,. \tag{3.52}$$

Applying this last inequality to (3.4), we obtain (3.47). By the inequality $\ell^*(x) \geq (x^2 - 2x)\ln(1-x), x \in [0, 1)$, inequality (3.48) follows from (3.47). Since for any $x \in [0, 1)$, $(x^2 - 2x)\ln(1-x) \geq 2\,x^2$, it follows that for any $x \in [0, 2D_n]$,

$$\left(\frac{D_n - x/2}{D_n}\right)^{\frac{D_n x - (x/2)^2}{V_n^2}} \;\leq\; \exp\left\{-\frac{x^2}{2V_n^2}\right\},$$

which gives (3.49).

If $\|d(X_1, X_1')\|_\infty < \infty$, then we have for any $x > 0$,

$$I_1(x) \leq d\,\mathbb{P}\bigg(\|d(X_1, X_1')\|_\infty \geq \frac{x}{2K_{1,n}}\bigg).$$

The last inequality implies that $I_1(x) = 0$ for $x > 2K_{1,n}\|d(X_1, X_1')\|_\infty$.

If the condition (2.6) is satisfied, by the point 4 of Lemma 2.1, then it holds $V_n^2 = O(n)$. Thus, inequality (3.49) implies that for any $x > 0$,

$$\mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq I_1(nxd^{-1/p}) + 2d\exp\left\{-c_{p,d}\frac{(xn)^2}{n}\right\}. \tag{3.53}$$

The last inequality implies (3.50).

If the condition (2.7) is satisfied, by the point 5 of Lemma 2.1, then we have $V_n^2 = O(n^{1+2\alpha})$ as $n \to \infty$. Thus, inequality (3.49) implies that for any $x > 0$,

$$\mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq I_1(nxd^{-1/p}) + 2d\exp\left\{-c_{p,d}\,\frac{(xn)^2}{n^{1+2\alpha}}\right\}.$$

From the last inequality, we get (3.51).

If the condition (2.8) is satisfied, by the point 6 of Lemma 2.1, then we have (3.22). Thus, inequality (3.49) implies that for any $x > 0$,

$$\mathbb{P}\left(\|S_n\|_p \geq nx\right) \leq \begin{cases} I_1(nxd^{-1/p}) + 2d\exp\left\{-c_{p,d}\,\dfrac{(xn)^2}{n^{1-2\alpha}}\,,\right\}, & \text{if } 0 \leq \alpha < 1/2, \\[2ex] I_1(nxd^{-1/p}) + 2d\exp\left\{-c_{p,d}\,\dfrac{(xn)^2}{\ln n}\right\}, & \text{if } \alpha = 1/2, \\[2ex] I_1(nxd^{-1/p}) + 2d\exp\Big\{-c_{p,d}(xn)^2\Big\}, & \text{if } 1/2 < \alpha \leq 1. \end{cases} \tag{3.54}$$

From (3.54), we obtain the point [iii] of the property. □

3.6. Hoeffding type bound

The next proposition is an application of Corollary 2.3 of Fan *et al.* [11], which is an extension of Hoeffding's inequality for super-martingales.

Proposition 3.7. *Assume that there exists a positive constant A_1 such that*

$$\mathbb{E}\big[\big(G_\varepsilon(\varepsilon)\big)^2\big] \leq A_1\,.$$

Put

$$V_n^2 = 2\sum_{k=2}^{n} K_{k,n}^2(\tau_k^2 A_1 + \xi_k^2) \quad \text{ and } \quad \delta_n = \max\{K_{k,n}\tau_k, K_{k,n}\xi_k, k = 2, \ldots, n\}. \tag{3.55}$$

Then for any $x, y > 0$,

$$\mathbb{P}\Big(\|S_n\|_p \geq d^{1/p}x\Big) \leq I_1(x) + 2d\,H_n\left(\frac{x}{2(y+1)\delta_n}, \frac{V_n}{(y+1)\delta_n}\right) + 2d\,\mathbb{P}\left(\max_{2\leq k\leq n} G_\varepsilon(\varepsilon_k) > y\right), \tag{3.56}$$

where

$$H_n(x, v) = \left\{\left(\frac{v^2}{x+v^2}\right)^{x+v^2}\left(\frac{n}{n-x}\right)^{n-x}\right\}^{\frac{n}{n+v^2}} \mathbf{1}_{\{x\leq n\}}\ , \tag{3.57}$$

with the convention that $(+\infty)^0 = 1$ *(which applies when* $x = n$*). In particular, if*

$$G_\varepsilon(\varepsilon) \le T \quad a.s.,$$

for a positive constant T*, then (3.56) implies that for any* $x > 0$*,*

$$\mathbb{P}\Big(\|S_n\|_p \ge d^{1/p} x\Big) \le I_1(x) + 2d\, H_n\left(\frac{x}{2(T+1)\delta_n}, \frac{V_n}{(T+1)\delta_n}\right). \tag{3.58}$$

Proof. We adapt the Corollary 2.3 of Fan *et al.* [11] with the truncature level $(y+1)\delta_n$. By (3.5), we obtain $|M_k^{(i)}| \le \delta_n(G_\varepsilon(\varepsilon_k)+1)$ for $k \in [2,n]$ and $i \in [1,d]$. Hence, for any $k \in [2,n]$,

$$\begin{aligned}\mathbb{E}\big[(M_k^{(i)})^2 \mathbf{1}_{\{M_k^{(i)} \le y\delta_n\}}|\mathcal{F}_{k-1}\big] &\le 2(K_{k,n}\tau_k)^2 \mathbb{E}\big[\big(G_\varepsilon(\varepsilon)\big)^2\big] + 2(K_{k,n}\xi_k)^2 \\ &\le 2K_{k,n}^2(\tau_k^2 A_1 + \xi_k^2).\end{aligned}$$

By Corollary 2.3 of Fan *et al.* [11], it follows that

$$\begin{aligned}\mathbb{P}(S_n^{(i)} \ge x/2) &\le H_n\left(\frac{x}{2(y+1)\delta_n}, \frac{V_n}{(y+1)\delta_n}\right) + \mathbb{P}\Big(\max_{2\le k\le n} M_k^{(i)} \ge (y+1)\delta_n\Big) \\ &\le H_n\left(\frac{x}{2(y+1)\delta_n}, \frac{V_n}{(y+1)\delta_n}\right) + \mathbb{P}\left(\max_{2\le k\le n} G_\varepsilon(\varepsilon_k) > y\right). \end{aligned} \tag{3.59}$$

Moreover, the same bound holds for $\mathbb{P}(-S_n^{(i)} \ge x/2)$. Applying (3.59) to (3.4), we obtain the desired inequality. □

Remark 3.2. *Using the Remark 2.1 of Fan et al. [11], we have for any* $x, v > 0$*,*

$$H_n(x,v) \le B(x,v) := \left(\frac{v^2}{x+v^2}\right)^{x+v^2} e^x \tag{3.60}$$

$$\le B_1(x,v) := \exp\left\{-\frac{x^2}{2(v^2+\frac{1}{3}x)}\right\}. \tag{3.61}$$

Note that $B(x,v)$ *and* $B_1(x,v)$ *are respectively known as Bennett's and Bernstein's bounds. Then, inequality (3.58) also implies the following Bennett's and Bernstein's bounds. For any* $x > 0$*, we have*

$$\begin{aligned}\mathbb{P}\Big(\|S_n\|_p \ge d^{1/p}x\Big) &\le I_1(x) + 2d\, B\left(\frac{x}{2(T+1)\delta_n}, \frac{V_n}{(T+1)\delta_n}\right) \\ &\le I_1(x) + 2d\, B_1\left(\frac{x}{2(T+1)\delta_n}, \frac{V_n}{(T+1)\delta_n}\right).\end{aligned}$$

4. Moment inequalities

4.1. Marcinkiewicz-Zygmund type bound

If the martingale differences $(M_i)_{i\ge1}$ have finite qth moments ($q \ge 2$), then we have the following Marcinkiewicz-Zygmund type inequality (cf. Rio [25]).

Proposition 4.1. *Let $q \geq 2$. Assume that there exist two positive constants $B_1(q)$ and $B_2(q)$ such that*

$$\mathbb{E}\big[\big(G_{X_1}(X_1)\big)^q\big] \leq B_1(q) \quad \textit{and} \quad \mathbb{E}\big[\big(G_\varepsilon(\varepsilon)\big)^q\big] \leq B_2(q)\,. \tag{4.1}$$

Denote

$$T_n(q) = K_{1,n}^2\big(B_1(q)\big)^{2/q} + (q-1)2^{2-2/q}\sum_{k=2}^{n} K_{k,n}^2\big(\tau_k^q B_2(q) + \xi_k^q\big)^{2/q}.$$

Then

$$\mathbb{E}\|S_n\|_q \leq d^{1/q}\sqrt{T_n(q)}. \tag{4.2}$$

Proof. Using Theorem 2.1 of Rio [25], we have for any $q \geq 2$,

$$(\mathbb{E}|S_n^{(i)}|^q)^{2/q} \quad \leq \quad (\mathbb{E}|M_1^{(i)}|^q)^{2/q} + (q-1)\sum_{k=2}^{n}(\mathbb{E}|M_k^{(i)}|^q)^{2/q}.$$

Again by (3.3), (3.5) and (4.1), we deduce that

$$\begin{aligned}
(\mathbb{E}|S_n^{(i)}|^q)^{2/q} &\leq \Big(\mathbb{E}[(K_{1,n}G_{X_1}(X_1))^q]\Big)^{2/q} + (q-1)\sum_{k=2}^{n}\Big(\mathbb{E}[(K_{k,n}[\tau_k G_\varepsilon(\varepsilon_k)+\xi_k])^q]\Big)^{2/q} \\
&\leq \Big(\mathbb{E}[(K_{1,n}G_{X_1}(X_1))^q]\Big)^{2/q} + (q-1)\sum_{k=2}^{n}\Big(2^{q-1}(K_{k,n})^q\mathbb{E}[(\tau_k G_\varepsilon(\varepsilon_k))^q + \xi_k^q]\Big)^{2/q} \\
&\leq K_{1,n}^2\big(B_1(q)\big)^{2/q} + (q-1)2^{2-2/q}\sum_{k=2}^{n}K_{k,n}^2\big(\tau_k^q B_2(q) + \xi_k^q\big)^{2/q} \\
&= T_n(q).
\end{aligned}$$

Using Jensen's inequality and the last inequality, we get

$$\begin{aligned}
\mathbb{E}\|S_n\|_q &= \mathbb{E}\Big(\sum_{i=1}^{d}|S_n^{(i)}|^q\Big)^{1/q} \leq \Big(\mathbb{E}\sum_{i=1}^{d}|S_n^{(i)}|^q\Big)^{1/q} = \Big(\sum_{i=1}^{d}\mathbb{E}|S_n^{(i)}|^q\Big)^{1/q} \\
&\leq d^{1/q}\sqrt{T_n(q)}\,,
\end{aligned}$$

which is the desired inequality. □

4.2. von Bahr-Esseen type bound

When the dominating random variables $G_{X_1}(X_1)$ and $G_\varepsilon(\varepsilon)$ have a moment of order $q \in [1,2]$, we have the following von Bahr-Esseen type bound.

Proposition 4.2. *Let $q \in [1,2]$. Assume that*

$$\mathbb{E}\big[\big(G_{X_1}(X_1)\big)^q\big] \leq A_1(q) \quad \textit{and} \quad \mathbb{E}\big[\big(G_\varepsilon(\varepsilon)\big)^q\big] \leq A_2(q)\,. \tag{4.3}$$

Let

$$V_n(q) = K_{1,n}^q A_1(q) + 2\sum_{k=2}^{n} K_{k,n}^q\big(\tau_k^q A_2(q) + \xi_k^q\big).$$

Then

$$\mathbb{E}\|S_n\|_q \le \big(d\, V_n(q)\big)^{1/q}. \tag{4.4}$$

Proof. By an argument similar to the proof of (3.39), we have

$$\begin{aligned}
\mathbb{E}|S_{2,n}^{(i)}|^q &\le \mathbb{E}\big[\big(K_{1,n}G_{X_1}(X_1)\big)^q\big] + 2^{2-q}\sum_{k=2}^{n} K_{k,n}^q \mathbb{E}\big[\big(\tau_k G_\varepsilon(\varepsilon) + \xi_k\big)^q\big] \\
&\le K_{1,n}^q \mathbb{E}\big[\big(G_{X_1}(X_1)\big)^q\big] + 2^{q-1}2^{2-q}\sum_{k=2}^{n} K_{k,n}^q \mathbb{E}\big[\big(\tau_k G_\varepsilon(\varepsilon)\big)^q + \xi_k^q\big] \\
&\le K_{1,n}^q A_1(q) + 2\sum_{k=2}^{n} K_{k,n}^q\big(\tau_k^q A_2(q) + \xi_k^q\big) \\
&= V_n(q).
\end{aligned}$$

It is easy to see that

$$\mathbb{E}\|S_n\|_q = \mathbb{E}\Big(\sum_{i=1}^{d} |S_n^{(i)}|^q\Big)^{1/q} \le \Big(\mathbb{E}\sum_{i=1}^{d} |S_n^{(i)}|^q\Big)^{1/q} = \Big(\sum_{i=1}^{d} \mathbb{E}|S_n^{(i)}|^q\Big)^{1/q} \le (d\, V_n(q))^{1/q},$$

which gives (4.4). □

5. Applications

5.1. Stochastic approximation by averaging

Let us return to the general linear problem of Example 3 in Subsection 1.3. In Subsection 5.1, we fix some $p \in [1, \infty]$ and $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ equipped with $d(x, x') = \delta(x, x') = \|x - x'\|_p$. For linear problem, the central limit theorems for $\overline{X}_n - x^*$ have been well studied by Polyak and Juditsky [20]. In this subsection, we focus on deviation inequalities for $\overline{X}_n - x^*$.

We first consider the case of Example 3 where $A_i \equiv A$ is deterministic. Recall that $\alpha \in [0, 1)$,

$$Ax^* = B \qquad \text{and} \qquad X_n = X_{n-1} - \frac{\gamma}{n^\alpha}(AX_{n-1} + \underbrace{\eta_n - B_{n-1}}_{=\varepsilon_n}).$$

From the last line, we deduce that

$$\begin{aligned}
\mathbb{E}X_n - x^* &= \mathbb{E}X_{n-1} - \frac{\gamma}{n^\alpha}A\ \mathbb{E}X_{n-1} + \frac{\gamma}{n^\alpha}B - x^* \\
&= \mathbb{E}X_{n-1} - \frac{\gamma}{n^\alpha}A\ \mathbb{E}X_{n-1} + \frac{\gamma}{n^\alpha}Ax^* - x^* \\
&= (I_d - \frac{\gamma}{n^\alpha}A)(\mathbb{E}X_{n-1} - x^*) \\
&= \prod_{k=2}^{n}(I_d - \frac{\gamma}{k^\alpha}A)(\mathbb{E}X_1 - x^*).
\end{aligned}$$

Thus, we have

$$\|\mathbb{E}X_n - x^*\|_p \leq \|\mathbb{E}X_1 - x^*\|_p \prod_{k=2}^{n} \rho_k,$$

where $\rho_n = 1 - \frac{\gamma \lambda_{\min}^{(p)}(A)}{n^\alpha}$. Using inequality (2.12), we have $\prod_{k=2}^n \rho_k \leq \exp\Big\{ - \frac{\gamma \lambda_{\min}^{(p)}(A)}{(1-\alpha)^2} \frac{n-2}{n^\alpha} \Big\}$, which leads to

$$\|\mathbb{E}X_n - x^*\|_p \;\leq\; \|\mathbb{E}X_1 - x^*\|_p \exp\Big\{ - \frac{\gamma \lambda_{\min}^{(p)}(A)}{(1-\alpha)^2} \frac{n-2}{n^\alpha} \Big\} \tag{5.1}$$

and

$$\|\mathbb{E}\overline{X}_n - x^*\|_p \leq \frac{1}{n} \sum_{k=1}^{n} \|\mathbb{E}X_k - x^*\|_p \;\leq\; \frac{C_0}{n},$$

where

$$C_0 = \|\mathbb{E}X_1 - x^*\|_p \bigg(1 + \sum_{k=2}^{\infty} \exp\Big\{ - \frac{\gamma \lambda_{\min}^{(p)}(A)}{(1-\alpha)^2} \frac{k-2}{k^\alpha} \Big\} \bigg). \tag{5.2}$$

Taking $f(X_1, X_2, \ldots, X_n) = n\,\overline{X}_n$, we can see that (2.1). Clearly, it holds

$$\overline{X}_n - x^* = \overline{X}_n - \mathbb{E}\overline{X}_n + \mathbb{E}\overline{X}_n - x^*,$$

which implies that

$$\|\overline{X}_n - x^*\|_p \leq \|\overline{X}_n - \mathbb{E}\overline{X}_n\|_p + \|\mathbb{E}\overline{X}_n - x^*\|_p \leq \|\overline{X}_n - \mathbb{E}\overline{X}_n\|_p + \frac{C_0}{n}.$$

Hence, we have

$$\begin{aligned} \mathbb{P}\Big(\big\|\overline{X}_n - x^*\big\|_p \geq x \Big) \;&\leq\; \mathbb{P}\Big(\|\overline{X}_n - \mathbb{E}\overline{X}_n\|_p \geq x - \frac{C_0}{n} \Big) \\ &=\; \mathbb{P}\Big(\big\| f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \big\|_p \geq nx - C_0 \Big) \\ &=\; \mathbb{P}\Big(\big\| S_n \big\|_p \geq nx - C_0 \Big). \end{aligned}$$

Notice that the condition (2.6) is satisfied in Example 3. Thus, the following qualitative inequalities are consequences of our deviation inequalities.

- If (3.6) and (3.9) hold, then there exist some positive constants $c_{1,p,d}$ and $c_{2,p,d}$, such that

$$\limsup_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\Big(\big\|\overline{X}_n - x^*\big\|_p \geq x \Big) \leq \begin{cases} -c_{1,p,d}\, x & \text{if } \ x \in (1, \infty) \\ -c_{2,p,d} x^2 & \text{if } \ x \in (0, 1]. \end{cases} \tag{5.3}$$

  This follows from the point [i] in Proposition 3.1.

- If (3.23) and (3.25) hold for some $q \in (0,1)$, then there exists a positive constant $c_{p,d}$ such that for any $x > 0$,

$$\limsup_{n\to\infty} \frac{1}{n^q} \ln \mathbb{P}\Big(\big\|\overline{X}_n - x^*\big\|_p \geq x \Big) \leq -c_{p,d}\, x^q. \tag{5.4}$$

  This follows from the point [i] in Proposition 3.2.

- If (3.31) and (3.37) hold for some $q \geq 1$, then there exists a positive constant $c_{p,d}$ such that for any $x > 0$,

$$\limsup_{n\to\infty} n^{q-1} \mathbb{P}\Big(\big\|\overline{X}_n - x^*\big\|_p \geq x \Big) \leq \frac{c_{p,d}}{x^q}. \tag{5.5}$$

  This follows from the points [i] in Proposition 3.3 (case $q \geq 2$) and Proposition 3.4 (case $q \in [1,2)$).

And for the moment bounds of S_n:

- If (4.1) holds for some $q \geq 2$, then, by (4.2) and the point 4 of Lemma 2.1,

$$\mathbb{E}\|\overline{X}_n - x^*\|_q \leq \frac{1}{n}\mathbb{E}\|S_n\|_q + \frac{C_0}{n} \leq \frac{c_{p,d}}{\sqrt{n}}. \tag{5.6}$$

- If (4.3) holds for some $q \in [1,2]$, then, by (4.4) and $V_n(q) = O(n)$,

$$\mathbb{E}\|\overline{X}_n - x^*\|_q \leq \mathbb{E}\|\overline{X}_n - \mathbb{E}\overline{X}_n\|_q + \frac{C_0}{n} \leq \frac{1}{n}\mathbb{E}\|S_n\|_q + \frac{C_0}{n} \leq \frac{c_{p,d}}{n^{1-1/q}}. \tag{5.7}$$

Remark 5.1. *Let us make some comments on the performances of uniform averaging $\overline{X}_n$, final iterate X_n and suffix averaging $\hat{X}_n = \frac{2}{n}\sum_{i=[n/2]}^{n} X_i$.*

1. *Clearly, if $\overline{X}_n - x^*$ is replaced by $X_n - x^*$, then inequalities (5.3)-(5.6) hold true, with C_0 replaced by*

$$C_n = n\|\mathbb{E}X_1 - x^*\|_p \exp\Big\{ -\frac{\lambda_{\min}^{(p)}(A)}{(1-\alpha)^2}\frac{n-2}{n^\alpha} \Big\}$$

   which is smaller than C_0 defined by (5.2) for any n large enough. Hoverer, this does not improve the convergence rates for the bounds (5.3)-(5.6). Thus uniform averaging $\overline{X}_n$ and final iterate X_n have almost the same performance for estimating x^* in a long time view.
2. *When uniform averaging $\overline{X}_n$ is replaced by suffix averaging $\hat{X}_n$, the inequalities (5.3)-(5.6) remain valid.*

Let us now focus on the case where A_i is stochastic. Recall that $\alpha \in [0,1)$,

$$x^* = \arg\min_{x\in\mathcal{C}} \|Ax - B\|_2 \qquad \text{and} \qquad X_n = \Pi_{\mathcal{C}}\Big[X_{n-1} - \frac{\gamma}{n^\alpha}(A_{n-1}X_{n-1} + \eta_n - B_{n-1}) \Big].$$

For the sake of simplicity, assume that $X_1 \in \mathcal{C}$, and that the condition (3.6) is satisfied. Under our assumptions, with moreover $\alpha \in [1/2, 1)$, and assuming that η_n and B_{n-1} have moments of order 4, Theorem 3 of [18] leads to

$$\mathbb{E}\|\overline{X}_n - x^*\|_2 = \frac{C_0}{\sqrt{n}}$$

for some $C_0 > 0$. Combining this with the point [i] in Proposition 3.1, we obtain for any $x > 0$,

$$\limsup_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\left(\|\overline{X}_n - x^*\|_2 \geq x\right) \;\leq\; -c_{2,d}\left(x\mathbf{1}_{\{x\geq 1\}} + x^2\mathbf{1}_{\{0<x<1\}}\right), \tag{5.8}$$

from which we get for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(\|\overline{X}_n - x^*\|_2 \;\leq\; \frac{C_0 + \sqrt{\frac{2}{c_{2,d}} \ln \frac{1}{\delta}}}{\sqrt{n}}\right) \geq 1 - \delta$$

for all n large enough.

5.2. Empirical risk minimization

It has been shown in the past how Bernstein type inequalities allow to control the error of the empirical risk minimizer, and thus to perform model selection for time series [16, 4, 3, 14, 2, 1]. These results are available under restrictive assumptions. For example, [4, 3, 1] focus on stationary series. In [2], nonstationary Markov chains as in (1.1) are considered, under the restriction that $\rho_n \leq \rho < 1$ and $\tau_n \leq \eta$ in the conditions (1.2) and (1.3). Our new Bernstein type bound, Proposition 3.1, allows to extend these results to a more general setting.

The context is as follows. For simplicity, here, $\mathcal{X}$ will be a Banach space with norm $\|\cdot\|_{\mathcal{X}}$. Assume we have a parameter set Θ and a family of functions $f_n(\theta, x)$ of $\theta \in \Theta$ and $x \in \mathcal{X}$, with $f_n(\theta, 0) = 0$. We observe that $X_1, \ldots, X_n$ satisfy (1.1), with $F_n(x, y) = f_n(\theta^0, x) + y$ for some unknown $\theta^0 \in \Theta$. Of course, the distributions of X_1 and of the ε_n's are also unknown.

Remark 5.2. *We review here some examples studied in the aforementioned references. In the case $\mathcal{X} = \mathbb{R}^d$, [1] studied functions of the form*

$$f_n(\theta, x) = \theta x$$

where θ is some $d \times d$ matrix. Note that in this case, the model does actually not depend on n. On the other hand, [2] considered a T-periodic version of these functions: for $\theta = (A_1|\ldots|A_T)$ where each A_t is a $d \times d$ matrix, they used

$$f_n(\theta, x) = A_{n(\mathrm{mod}\, T)} x.$$

Other examples include nonlinear autoregression with neural networks [4].

Let $\ell : \mathcal{X} \to [0, +\infty)$ be a function with $\ell(0) = 0$, it is usually refered to as the loss function. We will measure the performance of a predictor through its risk:

$$R_n(\theta) = \frac{1}{n-1} \sum_{k=2}^{n} \mathbb{E}\big[\ell\big(X_k - f_k(\theta, X_{k-1})\big)\big].$$

A classical loss function is simply given by the norm $\ell(x) = \|x\|_{\mathcal{X}}$ but other examples can be used, for example [3] used quantile losses in the case $\mathcal{X} = \mathbb{R}$. Our objective will be to estimate the minimizer θ^* of R_n. Under suitable assumptions, $\theta^* = \theta^0$: this is for example the case when $\mathcal{X}$ is actually a Hilbert space, $\ell(x) = \|x\|^2_{\mathcal{X}}$ and the ε_n are centered with $\mathbb{E}\|\varepsilon_n\|^2_{\mathcal{X}} < \infty$. However, this has no reason to be true in general, and it is important to note that if the objective is to minimize the loss of the predictions, to estimate θ^* is more important than to estimate θ^0. We define the ERM estimator $\hat{\theta}$ (for Empirical Risk Minimizer) by

$$\hat{\theta} = \arg\min_{\theta\in\Theta} r_n(\theta), \qquad \text{where } r_n(\theta) = \frac{1}{n-1}\sum_{k=2}^{n} \ell\big(X_k - f_k(\theta, X_{k-1})\big).$$

Definition 5.1. *Define the covering number $\mathcal{N}(\Theta, \epsilon)$ as the cardinality of the smallest set $\Theta_\epsilon \subset \Theta$ such that for any $\theta \in \Theta$, there exists a $\theta_\epsilon \in \Theta_\epsilon$ such that*

$$\sup_{k\in\{2,\ldots,n\}} \sup_{x\in\mathcal{X}} \frac{\|f_k(\theta, x) - f_k(\theta_\epsilon, x)\|_{\mathcal{X}}}{\|x\|_{\mathcal{X}}} \leq \epsilon.$$

Define the entropy of Θ by $\mathcal{H}(\Theta, \epsilon) = 1 \vee \ln \mathcal{N}(\Theta, \epsilon)$.

Examples of computation of $\mathcal{H}(\Theta, \epsilon)$ for some models can be found in references [2, 1]. In most classical examples, $\mathcal{H}(\Theta, \epsilon)$ is roughly in $1 \vee [D\ln(1 + C/\epsilon)]$, where D is the dimension of Θ and $C > 0$ is some constant.

Proposition 5.1. *Let $X_1, \ldots, X_n$ satisfy (1.1), (1.2) and (1.3) with (τ_n) and (ρ_n) satisfying (2.7) with $\alpha \in (0, 1/2)$ and $d(x, y) = \delta(x, y) = \|x - y\|_{\mathcal{X}}$ (note that $\xi_n \equiv 0$ in this case). Assume that (3.6) is satisfied, and that ℓ is L-Lipschitz, that is for any $(x, y) \in \mathcal{X}^2$,*

$$|\ell(x) - \ell(y)| \leq L\|x - y\|_{\mathcal{X}}.$$

Assume also that all the functions in the model are λ-Lipschitz: for any $(x, y) \in \mathcal{X}^2$, any $k \in \mathbb{N}$ and any $\theta \in \Theta$,

$$\|f_k(\theta, x) - f_k(\theta, y)\|_{\mathcal{X}} \leq \lambda\|x - y\|_{\mathcal{X}},$$

and that $\mathcal{H}(\Theta, 1/(Ln)) \leq D\ln n$ for some constant D. For n large enough, we have for any $\eta \in (0, 1)$,

$$R_n(\hat{\theta}) \leq \min_{\theta\in\Theta} R_n(\theta) + C_1\sqrt{\frac{D\ln n}{n^{1-2\alpha}}} + C_2\frac{1 + \ln\left(\frac{1}{\eta}\right)}{\sqrt{n^{1-2\alpha}}}$$

with probability at least $1 - \eta$, where C_1 and C_2 are constants that depend only on λ, L, and the constant $c_{p,d}$ in the proof of Proposition 3.1.

In particular, for $\alpha = 0$, we recover bounds that are similar to the ones in [2, 1].

In the case where several models are available and one doesn't know which one contains the truth, the previous result can be used to perform model selection. We refer the reader to [2] for example for details on this classical construction.

Proof. Fix $\theta \in \Theta$ and consider the random variable $S_n = g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)]$, where

$$g(x_1, \ldots, x_n) = \frac{1}{L(\lambda+1)} \sum_{k=2}^{n} \ell\left(x_k - f_k(\theta, x_{k-1})\right).$$

Note that

$$\begin{aligned}
|g(x_1, \ldots, x_n) - g(x_1, \ldots, x'_k, \ldots, x_n)| &\leq \frac{|\ell\left(x_{k+1} - f_{k+1}(\theta, x_k)\right) - \ell\left(x_{k+1} - f_{k+1}(\theta, x'_k)\right)|}{L(\lambda+1)} \\
&\quad + \frac{|\ell\left(x_k - f_k(\theta, x_{k-1})\right) - \ell\left(x'_k - f_k(\theta, x_{k-1})\right)|}{L(\lambda+1)} \\
&\leq \frac{\|f_{k+1}(\theta, x_k) - f_{k+1}(\theta, x'_k)\|_{\mathcal{X}} + \|x_k - x'_k\|_{\mathcal{X}}}{\lambda+1} \\
&\leq \|x_k - x'_k\|_{\mathcal{X}},
\end{aligned}$$

which means that g is separately Lipschitz. So we apply (3.19) in the proof of Proposition 3.1, that is, for any $t \in [0, \delta_n^{-1})$,

$$\mathbb{E}[\exp\{\pm t S_n\}] \leq \exp\left\{\frac{t^2 V_n}{2 - 2t\delta_n}\right\}.$$

Note that $S_n = \frac{n-1}{L(1+\lambda)}\left(r_n(\theta) - \mathbb{E}[r_n(\theta)]\right)$, and $R_n(\theta) = \mathbb{E}[r_n(\theta)]$. Set $s = t(n-1)/L(1+\lambda)$. We obtain that, for any $s \in [0, \delta_n^{-1}(n-1)/L(1+\lambda))$,

$$\mathbb{E}[\exp\{\pm s(r_n(\theta) - R_n(\theta))\}] \leq \exp\left\{\frac{s^2(1+\lambda)^2 L^2 \frac{V_n}{n-1}}{2(n-1) - 2s(1+\lambda)\delta_n L}\right\}. \tag{5.9}$$

Fix now $\epsilon > 0$ and a set $\Theta_\epsilon \subset \Theta$ as in Definition 5.1. For any $\theta \in \Theta_\epsilon$, we have $\theta \in \Theta$ and so (5.9) holds. Then, for any $s \in [0, \delta_n^{-1}(n-1)/L(1+\lambda))$ and any $x > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(\sup_{\theta \in \Theta_\epsilon} |r_n(\theta) - R_n(\theta)| > x\right) &\leq \sum_{\theta \in \Theta_\epsilon} \mathbb{P}\left(|r_n(\theta) - R_n(\theta)| > x\right) \\
&\leq \sum_{\theta \in \Theta_\epsilon} \mathbb{E}[\exp\{s|r_n(\theta) - R_n(\theta)| - sx\}] \\
&\leq 2\mathcal{N}(\Theta, \epsilon) \exp\left\{\frac{s^2(1+\lambda)^2 L^2 \frac{V_n}{n-1}}{2(n-1) - 2s(1+\lambda)\delta_n L} - sx\right\}.
\end{aligned} \tag{5.10}$$

Thanks to the definition of Θ_ϵ, for any $\theta \in \Theta$ there is a θ_ϵ such that

$$\sup_{i \in \{2, \ldots, n\}} \sup_{x \in \mathcal{X}} \frac{\|f_i(\theta, x) - f_i(\theta_\epsilon, x)\|_{\mathcal{X}}}{\|x\|_{\mathcal{X}}} \leq \epsilon.$$

So

$$|\ell(X_k - f_k(\theta_\epsilon, X_{k-1})) - \ell(X_k - f_k(\theta, X_{k-1}))| \leq L\|f_k(\theta_\epsilon, X_{k-1}) - f_k(\theta, X_{k-1})\|_{\mathcal{X}} \leq L\epsilon\|X_{k-1}\|_{\mathcal{X}}$$

and as a consequence, we obtain

$$|r_n(\theta) - r_n(\theta_\epsilon)| \leq L\epsilon \cdot \frac{1}{n-1}\sum_{k=1}^{n-1} \|X_k\|_{\mathcal{X}} \tag{5.11}$$

and

$$|R_n(\theta) - R_n(\theta_\epsilon)| \leq L\epsilon \cdot \frac{1}{n-1}\sum_{k=1}^{n-1} \mathbb{E}\|X_k\|_{\mathcal{X}}. \tag{5.12}$$

Using Proposition 3.1 with $f(X_1, \dots, X_n) = \sum_{k=1}^{n-1} \|X_k\|_{\mathcal{X}}$, we get for any $y > 0$ and any $u \in [0, \delta_n^{-1})$,

$$\begin{aligned}\mathbb{P}\Big(\sum_{k=1}^{n-1} \|X_k\|_{\mathcal{X}} > \sum_{k=1}^{n-1} \mathbb{E}\|X_k\|_{\mathcal{X}} + y\Big) &\leq \mathbb{E}\exp\left\{u\Big(\sum_{k=1}^{n-1} \|X_k\|_{\mathcal{X}} - \sum_{k=1}^{n-1} \mathbb{E}\|X_k\|_{\mathcal{X}} - y\Big)\right\} \\ &\leq \exp\left\{\frac{u^2 V_n}{2\,(1-u\delta_n)} - uy\right\}. \end{aligned} \tag{5.13}$$

From now, let us use the short notation $z_n = \sum_{k=1}^{n-1} \mathbb{E}\|X_k\|_{\mathcal{X}}$ and consider the "favorable" event

$$\mathcal{E} = \left\{\sum_{k=1}^{n-1} \|X_k\|_{\mathcal{X}} \leq z_n + y\right\} \bigcap \left\{\sup_{\theta\in\Theta_\epsilon} |r_n(\theta) - R_n(\theta)| \leq x\right\}.$$

On $\mathcal{E}$, by (5.11) and (5.12), we have

$$\begin{aligned} R_n(\hat\theta) &\leq R_n(\hat\theta_\epsilon) + \epsilon L \frac{z_n}{n-1} \leq r_n(\hat\theta_\epsilon) + x + \epsilon L \frac{z_n}{n-1} \\ &\leq r_n(\hat\theta) + x + \epsilon L\left[2\frac{z_n}{n-1} + \frac{y}{n-1}\right] \\ &= \min_{\theta\in\Theta} r_n(\theta) + x + \epsilon L \frac{2z_n + y}{n-1} \\ &\leq \min_{\theta\in\Theta_\varepsilon} r_n(\theta) + x + \epsilon L \frac{2z_n + y}{n-1} \\ &\leq \min_{\theta\in\Theta_\epsilon} R_n(\theta) + 2x + \epsilon L \frac{2z_n + y}{n-1} \\ &\leq \min_{\theta\in\Theta} R_n(\theta) + 2x + \epsilon L \frac{3z_n + y}{n-1}. \end{aligned}$$

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

$$R_n(\hat\theta) \leq \min_{\theta\in\Theta} R_n(\theta) + 2x + \frac{3z_n + y}{n(n-1)}. \tag{5.14}$$

Inequalities (5.10) and (5.13) lead to

$$\mathbb{P}\left(\mathcal{E}^c\right) \leq \exp\left\{\frac{u^2 V_n}{2\,(1-u\delta_n)} - uy\right\} + 2\mathcal{N}(\Theta, \frac{1}{Ln}) \exp\left\{\frac{s^2(1+\lambda)^2 L^2 \frac{V_n}{n-1}}{2(n-1) - 2s(1+\lambda)\delta_n L} - sx\right\}. \tag{5.15}$$

As it was explained in the proof of Proposition 3.1, we get $\delta_n = O(n^\alpha)$ and $V_n = O(n^{1+2\alpha})$. So, letting $c_{p,d}$ be as in the proof of Proposition 3.1, we get

$$\mathbb{P}\left(\mathcal{E}^c\right) \leq \exp\left\{\frac{u^2 c_{p,d} n^{1+2\alpha}}{2\left(1-u c_{p,d} n^\alpha\right)} - uy\right\} + 2\mathcal{N}(\Theta, \frac{1}{Ln}) \exp\left\{\frac{s^2(1+\lambda)^2 L^2 c_{p,d} n^{2\alpha}}{2(n-1) - 2s(1+\lambda)c_{p,d} n^\alpha L} - sx\right\}. \quad (5.16)$$

Fix $\eta \in (0,1)$ and put

$$x = \frac{s(1+\lambda)^2 L^2 c_{p,d} n^{2\alpha}}{2(n-1) - 2s(1+\lambda)c_{p,d} n^\alpha L} + \frac{\mathcal{H}(\Theta, \frac{1}{Ln}) + \ln\left(\frac{4}{\eta}\right)}{s}$$

and $y = \frac{\ln\left(\frac{2}{\eta}\right)}{u} + \frac{u c_{p,d} n^{1+2\alpha}}{2(1-u c_{p,d} n^\alpha)}$. Note that, plugging x and y into (5.16), those choices ensure $\mathbb{P}(\mathcal{E}^c) \leq \eta/2 + \eta/2 = \eta$, while (5.14) becomes:

$$\begin{aligned} R_n(\hat{\theta}) \quad &\leq \quad \min_{\theta\in\Theta} R_n(\theta) + \frac{s(1+\lambda)^2 L^2 c_{p,d} n^{2\alpha}}{(n-1) - 2s(1+\lambda)c_{p,d} n^\alpha L} + 2\frac{D\ln(n) + \ln\left(\frac{4}{\eta}\right)}{s} \\ &\quad + \frac{3z_n}{n(n-1)} + \frac{\ln\left(\frac{2}{\eta}\right)}{un(n-1)} + \frac{u c_{p,d} n^{1+2\alpha}}{2(1-u c_{p,d} n^\alpha) n(n-1)}. \end{aligned}$$

The final steps are to choose $u \in [0, \delta_n^{-1})$, $s \in [0, \delta_n^{-1}(n-1)/L(1+\lambda))$ and to provide an upper bound on z_n. First, put $u = 1/(2c_{p,d} n^\alpha)$ and

$$s = \frac{n-1}{(1+\lambda)L}\sqrt{\frac{2D\ln(n)}{n^{1+2\alpha}}}.$$

Note that we always have $u < \delta_n^{-1}$. Moreover, we have $sn^\alpha = o(n)$, so for n large enough, the condition on s is satisfied too. Thus, for n large enough, there are constants C_1 and C_2 such that

$$R_n(\hat{\theta}) \leq \min_{\theta\in\Theta} R_n(\theta) + C_1\sqrt{\frac{D\ln n}{n^{1-2\alpha}}} + C_2\frac{1+\ln\left(\frac{1}{\eta}\right)}{\sqrt{n^{1-2\alpha}}} + \frac{3z_n}{n(n-1)}. \quad (5.17)$$

Let us now present an upper bound of $z_n = \sum_{k=1}^{n-1} \mathbb{E}\|X_k\|_{\mathcal{X}}$. Recall that $\mathbb{E}\|X_1\|_{\mathcal{X}} = \int \|x - 0\|_{\mathcal{X}} \mathbb{P}_{X_1}(dx) = G_{X_1}(0)$ and $\mathbb{E}\|\varepsilon\|_{\mathcal{X}} = \int \|x-0\|_{\mathcal{X}} \mathbb{P}_\varepsilon(dx) = G_\varepsilon(0)$. Then, by the point 5 of Lemma 2.1 and inequality (2.12), we have

$$\begin{aligned} \mathbb{E}\|X_n\|_{\mathcal{X}} &= \mathbb{E}\|f_{n-1}(\theta, X_{n-1}) + \varepsilon_n\|_{\mathcal{X}} \leq \mathbb{E}\|f_{n-1}(\theta, X_{n-1}) - f_{n-1}(\theta, 0)\|_{\mathcal{X}} + \mathbb{E}\|\varepsilon_n\|_{\mathcal{X}} \\ &\leq \rho_n \mathbb{E}\|X_{n-1}\|_{\mathcal{X}} + G_\varepsilon(0) \leq \cdots \leq \rho_n \cdots \rho_2 \mathbb{E}\|X_1\|_{\mathcal{X}} + K_{2,n}\mathbb{E}\|\varepsilon\|_{\mathcal{X}} \\ &\leq C_3 n^\alpha G_\varepsilon(0) + \exp\left\{-\frac{(n-1)\rho}{(1-\alpha)^2 n^\alpha}\right\} G_{X_1}(0). \end{aligned}$$

Therefore, it holds

$$z_n \leq C_4 n^{1+\alpha}.$$

Applying the last inequality to (5.17), we get

$$
\begin{aligned}
R_n(\hat{\theta}) &\le \min_{\theta\in\Theta} R_n(\theta) + C_1\sqrt{\frac{D\ln n}{n^{1-2\alpha}}} + C_2\frac{1+\ln\left(\frac{1}{\eta}\right)}{\sqrt{n^{1-2\alpha}}} + \frac{3C_4 n^{1+\alpha}}{n(n-1)} \\
&\le \min_{\theta\in\Theta} R_n(\theta) + C_5\sqrt{\frac{D\ln n}{n^{1-2\alpha}}} + C_2\frac{1+\ln\left(\frac{1}{\eta}\right)}{\sqrt{n^{1-2\alpha}}},
\end{aligned}
$$

which gives the desired result. □

Acknowledgements

This work has been partially supported by the National Natural Science Foundation of China (Grant nos. 11601375 and 11971063). This work has also been funded by CY Initiative of Excellence (grant "Investissements d'Avenir" ANR-16-IDEX-0008), Project "EcoDep" PSI-AAP2020-0000000013.

References

[1] Alquier, P., Bertin, K., Doukhan, P., Garnier, R. 2020. High-dimensional VAR with low-rank transition. *Statist. Computing* 30(4), 1139-1153.

[2] Alquier P., Doukhan P., Fan X. 2019. Exponential inequalities for nonstationary Markov chains. *Depend. Model.* 7(1), 150–168.

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

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

[5] Borovkov, A.A., 2000. Estimates for the distribution of sums and maxima of sums of random variables when the Cramér condition is not satisfied. *Sib. Math. J.* 41, 811–848.

[6] Cuny, C., Dedecker, J., Merlevède, F., 2017. Large and moderate deviations for the left random walk on $GL_d(R)$. *ALEA* 14, 503-527.

[7] De la Peña, V.H., 1999. A general class of exponential inequalities for martingales and ratios. *Ann. Probab.* 27, 537–564.

[8] Dedecker, J., Fan, X., 2015. Deviation inequalities for separately Lipschitz functionals of iterated random functions. *Stochastic Process. Appl.* 125, 60–90.

[9] Dedecker, J., Doukhan P., Fan X. 2019. Deviation inequalities for separately Lipschitz functionals of composition of random functions. *J. Math. Anal. Appl.* 479(2), 1549–1568.

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

[11] Fan, X., Grama, I., Liu. Q., 2012. Hoeffding's inequality for supermartingales. *Stochastic Process. Appl.* 122, 3545–3559.

[12] Fan, X., Grama, I., Liu, Q., 2017. Deviation inequalities for martingales with applications. *J. Math. Anal. Appl.* 448(1), 538–566.

[13] Fuk, D. Kh., 1973. Some probabilistic inequalities for martingales. *Sib. Math. J.* 14, 185–193.

[14] Hang, H., Steinwart, I., 2014. Fast learning from α-mixing observations. *J. Multivariate Anal.* 127, 184-199.

[15] Lesigne, E., Volný, D., 2001. Large deviations for martingales. *Stochastic Process. Appl.* 96, 143–159.

[16] Meir, R., 2000. Nonparametric time series prediction through adaptive model selection. *Mach. Learn.* 39(1), 5-34.

[17] Merlevède, F., Peligrad, M., Rio, E., 2011. A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probab. Theory Relat. Fields* 151, 435–474.

[18] Moulines, E., Bach, F., 2011. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems* 24, 451–459.

[19] Nagaev, S.V. 1979. Large deviations of sums of independent random variables. *Ann. Probab.* 7, 745–789.

[20] Polyak, B.T., Juditsky, A.B., 1992. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* 30(4), 838–855.

[21] Phillips, P.C.B., 1987. Towards a unified asymptotic theory for autoregression. *Biometrika* 74(3), 535–547.

[22] Phillips, P.C.B., 1988. Regression theory for near integrated time series. *Econometrica* 56, 1021–1044.

[23] Phillips, P.C.B., Magdalinos, T., 2007. Limit theory for moderate deviations from a unit root under weak dependence. *J. Econometrics* 136(1), 115–130.

[24] Pinelis, I., 2015. Best possible bounds of the von Bahr–Esseen type. *Ann. Funct. Anal.* 6(4), 1-29.

[25] Rio, E., 2009. Moment inequalities for sums of dependent random variables under projective conditions. *J. Theor. Probab.* 22, 146-163.

[26] Rio, E., 2013. On McDiarmids concentration inequality. *Electron. Commun. Probab.* 18(44), 1–11.