Exponential inequalities for unbounded functions of geometrically ergodic Markov chains. Applications to quantitative error bounds for regenerative Metropolis algorithms

Olivier Wintenberger

Exponential inequalities for unbounded functions of geometrically ergodic Markov chains. Applications to quantitative error bounds for regenerative Metropolis algorithms.

Olivier Wintenberger*

Abstract

The aim of this note is to investigate the concentration properties of unbounded functions of geometrically ergodic Markov chains. We derive concentration properties of centered functions with respect to the square of the Lyapunov's function in the drift condition satisfied by the Markov chain. We apply the new exponential inequalities to derive confidence intervals for MCMC algorithms. Quantitative error bounds are providing for the regenerative Metropolis algorithm of [5].

Keywords: Markov chains, exponential inequalities, Metropolis algorithm, Confidence interval.

1 Introduction

At the conference in honor of Paul Doukhan, Jérôme Dedecker presented the new Hoeffding's inequality [6] for functions f of a geometric ergodic Markov chain (X_k), $1 \leq k \leq n$. Using a similar counter example than in [1], he showed that the boundedness assumption is necessary to obtain such exponential inequalities for functions of geometrically ergodic Markov chain.

In this note, we study different concentration properties for relaxing the boundedness condition. We extend the framework of [6] by considering concentration properties of f involving a second order term that depends also on the observations. Such exponential inequalities are empirical Bernstein's ones. As the second order term is an over estimator of the asymptotic variance, the new inequality (2.4) is closely related with the self normalized concentration inequalities studied in [7]. In this note, the second order term depends on the squares of teh Lyapunov's function V in the drift condition (2.1) satisfied by the

*olivier.wintenberger@upmc.fr, Sorbonne Universités, UPMC Univ Paris 06, LSTA, Case 158, 4 place Jussieu, 75005 Paris, FRANCE & Department of Mathematical Sciences, University of Copenhagen, DENMARK

Markov chain.

We apply our result to the construction of confidence intervals for some MCMC algorithms. Previous studies are based on a two steps reasoning: first some bounds are deriving with unknown constants, via the Chebyshev's and Hoeffding's inequalities, see [15] and [10] respectively. The second step consists in over-estimating the constants. The confidence level is obtained by an union bound on the confidence levels of the two steps. Our new empirical exponential inequalities provide concentration properties thanks to a second order term that is observed. We achieve a quantitative error analysis by a direct application of the techniques in [7] for the regenerative Metropolis algorithm of [5]. A similar one step procedure was developed in [14] under a more restrictive Ricci curvature condition. Our approach provides quantitative bounds that can be reasonable if the Lyapunov's function can be well-chosen. However, the confidence intervals are certainly over estimated due to the conservativeness of the coupling technique used in the proof.

The paper is organized as follows. The main result, the concentration properties for unbounded functions of Markov chains, is stated in Theorem 2.1 of Section 2. Follow its proof that relies on a coupling argument mixing the arguments of [11] and [26]. Then Section 3 is devoted to the construction of confidence intervals for MCMC algorithms. The case of the regenerative Metropolis algorithm of [5] is studied in details. Simulations and discussions are given in Section 4.

2 Concentration for unbounded functions of Markov chains under the drift condition.

We consider an exponentially ergodic Markov kernel P on some countably generated space E that satisfies the following drift and minorization conditions (2.1) and (2.2) respectively: it exists a Lypounov function $V : E \mapsto [1, \infty)$, a probability measure ν and four positive constants b, R_0 and $c < 1$, $\beta < 1$ such that

$$PV \leq \beta V + b, \tag{2.1}$$

$$P(x, \cdot) \geq c(R)\nu(\cdot), \qquad \text{if} \qquad V(x) \leq R, \qquad R \geq R_0. \tag{2.2}$$

These conditions are slightly stronger than the exponential ergodicity of the Markov chain. It is related to the Feller's property, see [19]. Note that $c(R) \to 0$ as $R \to \infty$. The conditions (2.1) and (2.2) are satisfied in many examples, such as random coefficient autoregressive processes, see [9], or the trajectories of the Random Walk Metropolis algorithm, see Section 3. Let us consider a function f on E^n satisfying

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, y_k, x_{k+1}, \ldots, x_n)| \leq L_k(V(x_k) + V(y_k)). \tag{2.3}$$

Dedecker and Gouezel [6] extended the Hoeffding's inequality to the trajectory $(X_1, \ldots, X_n)$ of the Markov chain P starting from $X_0 = x$ and denoted P_x. They proved the existence

of a constant $K_R > 0$ independent on n such that

$$\mathbb{E}_x[\exp(f - \mathbb{E}_x[f])] \leq e^{K_R \sum_{k=1}^n L_k^2}, \qquad x \in \{V \leq R\}.$$

We prove the following result:

Theorem 2.1. *Assume that P satisfies the drift conditions (2.1) and the minorization condition (2.2) with $R \geq R_0$ satisfying $\bar{\beta}(R) := \beta + 2b/(1+R) < 1$. Assume that $PV_k^2 := \mathbb{E}[V^2(X_k) \mid X_{k-1}]$ is well defined and denote $V_k := V(X_k)$. If f satisfies (2.3) we have, for any $x \in E$,*

$$\mathbb{E}_x\Big[\exp\Big(\lambda(f - \mathbb{E}_x[f]) - \frac{1 + \bar{\beta}(R)((R-1)/c(R) - R)}{1 - \bar{\beta}(R)} \sum_{k=1}^n \frac{(\lambda L_k)^2}{2}(PV_k^2 + V_k^2)\Big)\Big] \leq 1. \quad (2.4)$$

Eq (2.4) holds true in the stationary case with $\mathbb{E}$ replacing $\mathbb{E}_x$ and $\mathbb{E}[V(X_1)^2]$ replacing PV_1^2.

Remark 2.1. Such inequalities implies exponential inequalities for the normalized process. Applying Theorem 2.1 of [7], we obtain the subgaussian inequality $\mathbb{E}[\exp(xY)] \leq \sqrt{2}\exp(Cx^2)$, $x > 0$, of the process

$$Y := \frac{f - \mathbb{E}_x[f]}{\sqrt{\sum_{k=1}^n L_k^2(PV_k^2 + V_k^2 + 2\mathbb{E}_x[V_k^2])}}$$

for some constant $C > 0$. One recognizes the self normalized process when $f = \sum_{k=1}^n L_k V_k$. Such bounds cannot be obtained using the approach of [6] because the bounded differences properties [18] of the self normalized process are growing as $\sqrt{n}$.

Remark 2.2. For bounded function f one can compare (2.4) with the result of Dedecker and Gouezel [6]. The limitation of the result in Theorem 2.1 is that considering $V = 1$ constrains the Markov chain to be uniformly ergodic. In such restrictive case, the classical Bernstein's inequality was extended by Samson in [24]. So our approach, providing less accurate exponential inequalities, is useless in the case $V = 1$. The advantage of our approach is that we have explicit constants depending on V and that x can be taken arbitrary in E. For instance, when $f = \sum_{k=1}^n g(X_k)$ with g bounded, considering the Markov chain $(g(X_k))_{1 \leq k \leq n}$ and the Lyapunov function Vog^{-1} we obtain

$$\mathbb{E}_x\Big[\exp\Big(\frac{\lambda}{\sqrt{n}} \sum_{k=1}^n (g(X_k) - \mathbb{E}_x[g(X_k)])\Big)\Big] \leq e^{K\lambda^2}, \qquad x \in E,$$

with $K = (1 + \bar{\beta}((R-1)/c(R) - R))/(1 - \bar{\beta}(R)) \sup_y V^2 og^{-1}(y)$ for y in the range of g. Then we extend the Hoeffding's inequality of Dedecker and Gouezel [6] to $x \in E$ with an explicit constant when $f = \sum_{k=1}^n g(X_k)$ with g bounded.

Proof of Theorem 2.1. The proof is based on a new coupling argument applied to the coupling scheme $(X_k, X_k')_{1 \leq k \leq n}$ of [23], where $(X_k')_{1 \leq k \leq n}$ is a copy of $(X_k)_{1 \leq k \leq n}$. Let us

first recall the construction of the coupling scheme for completeness. Any Markov chain $\bar{P}$ on E^2 with common margins P also satisfies

$$\bar{P}\bar{V}(x,x') \leq \beta \bar{V}(x,x') + 2b,$$

for the drift function $\bar{V}(x,x') = V(x) + V(x')$. Moreover, there exists a coupling kernel $\bar{P}$, see [23] for details, with common margin P such that

$$\bar{P}((x,x'),\cdot\times\cdot) \geq c(R)\nu(\cdot), \qquad (x,x') \in \{V \leq R\}^2.$$

In particular, $\bar{P}((x,x'),\cdot)$, $(x,x') \in \{V \leq R\}^2$, has a mass at least equal to $c(R)$ on the diagonal. As $\bar{V} \geq 1 + R$ when $(x,x') \notin \{V \leq R\}^2$, we also have

$$\bar{P}\bar{V}(x,x') \leq \left(\beta + \frac{2b}{1+R}\right)\bar{V}(x,x'), \qquad (x,x') \notin \{V \leq R\}^2.$$

We have $\bar{\beta} = \beta + 2b/(1+R) < 1$ by assumption. Then one can apply the Nummelin's splitting scheme on the Markov chain (X_t, X'_t) driven by $\bar{P}$. There exists an enlargement (X_t, X'_t, B_t) with $B_t \in \{0,1\}$ such that it admits an atom $A = \{V \leq R\}^2 \times \{1\}$ and $\mathbb{P}(B_t = 1 \mid (X_t, X'_t) \in \{V \leq R\}^2) = c(R)$. Let τ_A denotes the first hitting time to the atom A. From the Dynkin's formula, denoting $\bar{V}_k = \bar{V}(X_k, X'_k)$ we have for any stopping time τ

$$\bar{\mathbb{E}}_{x,x'}[\bar{V}_\tau] = \bar{V}(x,x') + \bar{\mathbb{E}}_{x,x'}\Big[\sum_{k=1}^{\tau} \bar{P}\bar{V}_k - \bar{V}_{k-1}\Big].$$

We consider the stopping time τ as the first hitting time to $\{V \leq R\}^2$. Plugging the drift condition in the Dynkin's formula, we obtain

$$\begin{aligned}\bar{\mathbb{E}}_{x,x'}[\bar{V}_\tau] &= \bar{V}(x,x') + \bar{\mathbb{E}}_{x,x'}\Big[\sum_{k=1}^{\tau} \bar{P}\bar{V}_k - \bar{V}_{k-1}\Big] \\ &\leq \bar{V}(x,x') + (\bar{\beta}(R) - 1)\bar{\mathbb{E}}_{x,x'}\Big[\sum_{k=1}^{\tau} \bar{V}_{k-1}\Big].\end{aligned}$$

Then we obtain

$$\mathbb{E}_{x,x'}\Big[\sum_{k=0}^{\tau} \bar{V}_k\Big] \leq \frac{\bar{V}(x,x') - \bar{\beta}(R)\bar{\mathbb{E}}_{x,x'}[\bar{V}_\tau]}{1 - \bar{\beta}(R)} \leq \frac{\bar{V}(x,x') - 2\bar{\beta}(R)}{1 - \bar{\beta}(R)}. \tag{2.5}$$

Denoting $\tau(j)$ the successive hitting times to $\{V \leq R\}^2$, we have

$$\begin{aligned}\bar{\mathbb{E}}_{x,x'}\Big[\sum_{k=0}^{\infty} \bar{V}_k\Big] &= \mathbb{E}_{x,x'}\Big[\sum_{k=0}^{\tau} \bar{V}_k\Big] + \bar{\mathbb{E}}_{x,x'}\left[\sum_{j=1}^{\infty}\sum_{k=\tau(j)+1}^{\tau(j+1)} \bar{V}_k 1_{B_1=\cdots B_j=0}\right] \\ &\leq \frac{\bar{V}(x,x') - 2\bar{\beta}(R)}{1 - \bar{\beta}(R)} + \bar{\mathbb{E}}_{x,x'}\left[\sum_{j=1}^{\infty}(1-c(R))^j \bar{E}_{(X_{\tau(j)},X_{\tau(j)'})}\left[\sum_{k=\tau(j)+1}^{\tau(j+1)} \bar{V}_k\right]\right].\end{aligned}$$

using the strong Markov property to assert the last identity. Using (2.5) and $\sup_{\{V\le R\}^2}\bar{V}\le 2R$ we obtain

$$\bar{E}_{(X_{\tau(j)},X_{\tau(j)'})}\left[\sum_{k=\tau(j)+1}^{\tau(j+1)}\bar{V}_k\right]\le\sup_{\{V\le R\}^2}\mathbb{E}_{x,x'}\Big[\sum_{k=1}^{\tau}\bar{V}_k\Big]\le\frac{2\bar{\beta}(R)}{1-\bar{\beta}(R)}(R-1)$$

Collecting those bounds, we derive

$$\begin{aligned}\bar{\mathbb{E}}_{x,x'}\Big[\sum_{k=0}^{\infty}\bar{V}_k\Big]&\le\frac{\bar{V}(x,x')-2\bar{\beta}(R)}{1-\bar{\beta}(R)}+\frac{2\bar{\beta}(R)(R-1)}{c(R)(1-\bar{\beta}(R))}-\frac{2\bar{\beta}(R)(R-1)}{1-\bar{\beta}(R)}\\&\le\frac{\bar{V}(x,x')-2\bar{\beta}(R)R}{1-\bar{\beta}(R)}+\frac{2\bar{\beta}(R)(R-1)}{c(R)(1-\bar{\beta}(R))}.\end{aligned}$$

We are now ready to use our new coupling argument, combining the metric $d_V(x,y)=\bar{V}(x,y)=V(x)+V(y)$ if $x\neq y$ and $d_V(x,y)=0$ else of [11] with the Γ-weak dependence notion of [26]. A main difference with [11] is that the coupling argument of [26] does not require any contractivity of the Markov kernel with respect to the metric d_V. Denoting $K=(1+2\bar{\beta}(R)((R-1)/c(R)-R)/(1-\bar{\beta}(R))$, we have

$$\mathbb{E}_{x,x'}\Big[\sum_{k=0}^{\infty}d_V(X_k,X_k')\Big]=\mathbb{E}_{x,x'}\Big[\sum_{k=0}^{\tau_A}d_V(X_k,X_k')\Big]\le Kd_V(x,x')\tag{2.6}$$

as $X_k=X_k'$ for $k>\tau_A$. Recall the following definition from [26]:

Definition 2.1. A Markov chain is $\Gamma_{d_V,d_V}(1)$-weakly dependent if for any $(x,x')\in E^2$ there exist coefficients $\gamma_{k,0}(1)\ge 0$ and a coupling scheme $(X_k,X_k')_{1\le k\le n}$ satisfying

$$\mathbb{E}_{x,x'}[d_V(X_k,Y_k)]\le\gamma_{k,0}(1)d_V(x,x'),\qquad 0\le k\le n.$$

In view of (2.6), we claim that the Markov chain $(X_k)_{1\le k\le n}$ is $\Gamma_{d_V,d_V}(1)$-weakly dependent with dependence coefficients satisfying

$$\sum_{k=0}^{\infty}\gamma_{k,0}(1)\le K.$$

We denote $X=(X_1,\dots,X_n)$ on E^n starting from x with distribution P_x and $d_{V,L}$ the metric on E^n such that

$$d_{V,L}(x,y)=\sum_{k=1}^{n}L_kd_V(x_k,y_k).$$

Recall the definition of the Wasserstein distance between P_x and any measure Q on E^n

$$W_{1,d_{V,L}}(P,Q)=\inf_{\pi}\mathbb{E}_{\pi}[d_{V,L}(X,Y)],$$

where π is any coupling measure such that $(X,Y)\sim\pi$, $X\sim P_x$ and $Y\sim Q$. From Eq. (3.11) of [26], we have the following result: denoting $Q_{Y^{(j-1)}}$ the conditional probability

of Y given $Y^{(j-1)} = (Y_1, \dots, Y_n)$ and noticing that $P_{Y^{(j-1)}} = P_{Y_{j-1}}$ by the strong Markov property, the Wasserstein distance satisfies

$$\begin{aligned} W_{1,d_{V,L}}(P,Q) &\leq \sum_{j=1}^{n}\sum_{k=j}^{n} \gamma_{k-j,0}(1)\mathbb{E}_Q[L_j W_{1,d_V}(P_{Y_{j-1}}, Q_{Y^{(j-1)}})] \\ &\leq K\sum_{j=1}^{n} L_j \mathbb{E}_Q[W_{1,d_V}(P_{Y_{j-1}}, Q_{Y^{(j-1)}})]. \end{aligned}$$

We estimate the right hand side term applying successively Cauchy-Schwarz and Young's inequalities

$$\begin{aligned} W_{1,d_V}(P_{Y_{j-1}}, Q_{Y^{(j-1)}}) =& \inf_{\pi} \mathbb{E}_\pi[(V(X') + V(Y'))\,\mathbb{1}_{X\neq Y}] \\ \leq& \sqrt{\mathbb{E}_\pi[V^2(X')]\mathbb{E}[\pi(X'\neq Y' \mid X')^2]} \\ &+ \sqrt{\mathbb{E}_\pi[V^2(Y')]\mathbb{E}[\mathbb{P}(X'\neq Y' \mid Y')^2]} \\ \leq& \frac{\lambda}{2}(\mathbb{E}_\pi[V^2(X')] + \mathbb{E}_\pi[V^2(Y')]) \\ &+ \frac{\mathbb{E}_\pi[\pi(X'\neq Y' \mid X')^2] + \mathbb{E}_\pi[\mathbb{P}(X'\neq Y' \mid Y')^2]}{2\lambda}. \end{aligned}$$

As $X' \sim P_{Y_{j-1}}$ one can identify $\mathbb{E}[V^2(X')] = PV_j^2$. We then use the following improvement of the Marton's inequality [17] (see Lemma 8.3 of [3] combined with Lemma 2 of [24])

$$\mathbb{E}_\pi[\pi(X'\neq Y' \mid X')^2] + \mathbb{E}_\pi[\pi(X'\neq Y' \mid Y')^2] \leq 2\mathcal{K}(Q_{Y^{(j-1)}}, P_{Y_{j-1}}),$$

where $\mathcal{K}(Q,P)$ is the Kulback-Leibler divergence between two probability measures P and Q:

$$\mathcal{K}(Q,P) = \mathbb{E}_Q[\log(dQ/dP)].$$

We obtain

$$W_{1,d_V}(P_{Y_{j-1}}, Q_{Y^{(j-1)}}) \leq \frac{\lambda}{2}(PV_j^2 + \mathbb{E}_\pi[V^2(Y')]) + \lambda^{-1}\mathcal{K}(Q_{Y^{(j-1)}}, P_{Y_{j-1}}).$$

Combining those inequalities, as $Y' \sim Q_{Y^{(j-1)}}$ so that $\mathbb{E}_Q[\mathbb{E}_\pi[V^2(Y')]] = \mathbb{E}_Q[V_j^2]$, we obtain

$$\begin{aligned} W_{1,d_{V,L}}(P,Q) &\leq K\sum_{j=1}^{n} L_j \mathbb{E}_Q[W_{1,d_V}(P_{Y_{j-1}}, Q_{Y^{(j-1)}})] \\ &\leq K\sum_{j=1}^{n}\left(\frac{\lambda L_j^2}{2}(\mathbb{E}_\pi[V^2(X')] + \mathbb{E}_\pi[V^2(Y')])\right. \\ &\quad \left. + \frac{\mathbb{E}_\pi[\pi(X'\neq Y' \mid X')^2] + \mathbb{E}_\pi[\pi(X'\neq Y' \mid Y')^2]}{2\lambda}\right) \\ &\leq K\mathbb{E}_Q\left[\sum_{j=1}^{n}\left(\frac{\lambda L_j^2}{2}(PV_j^2 + V_j^2) + \lambda^{-1}\mathcal{K}(Q_{Y^{(j-1)}}, P_{Y_{j-1}})\right)\right]. \end{aligned}$$

From the identity

$$\mathbb{E}_Q\Big[\sum_{j=1}^{n}\mathcal{K}(Q_{Y^{(j-1)}},P_{Y_{j-1}})\Big]=\mathcal{K}(Q,P_x)$$

we obtain

$$W_{1,d_{V,L}}(P_x,Q)\le K\mathbb{E}_Q\left[\sum_{k=1}^{n}\frac{\lambda L_k^2}{2}(PV_k^2+V_k^2)\right]+\lambda^{-1}\mathcal{K}(Q,P_x).$$

Then we apply the Kantorovich's duality (see for instance [25]):

$$W_{1,d_{V,L}}(P_x,Q)=\sup_g \mathbb{E}_Q[g]-\mathbb{E}_x[g]$$

where g is 1-Lipschitz with respect to the $d_{V,L}$ metric:

$$|g(x)-g(y)|\le\sum_{k=1}^{n}L_j(V(x_k)+V(y_k))\,\mathbb{1}_{x_k\neq y_k}.$$

Thus, as any f satisfying (2.3) also satisfies such Lipschitz condition, we obtain

$$\mathbb{E}_Q\Big[\lambda(f-\mathbb{E}_x[f])-K\sum_{k=1}^{n}\frac{(\lambda L_k)^2}{2}(PV_k^2+V_k^2)\Big]\le\mathcal{K}(Q,P_x).$$

Choosing the probability measure Q as

$$dQ\propto\exp\Big(\lambda(f-\mathbb{E}_x[f])-K\sum_{k=1}^{n}\frac{(\lambda L_k)^2}{2}(PV_k^2+V_k^2)\Big)dP_x$$

we obtain the desired inequality for the trajectory X starting from $x\in E$.

In the stationary case, one replaces $P_{Y_0}=P_x(X_1\in\cdot)$ by P_0 the unconditional distribution of X_1. Adding artificially an initial point $X_0=Y_0=x_0$ for a fixed point $x_0\in E$, we check the Γ_{d_V,d_V} weak dependence of the stationary trajectory $(X_1,\dots,X_n)$ even if that notion of dependence is defined conditionally to the past, see [8] and [26] for more details. Thus the same reasoning holds and the result follows similarly in the stationary case. □

3 Application to non asymptotic confidence intervals for MCMC algorithms

In this section we are considering the approximation of $\int g(x)dP_0(x)=\mathbb{E}[g]$ for some unbounded function g and some density P_0, known up to the normalizing constant. The Markov Chain Monte Carlo (MCMC) algorithms generates the approximation $\frac{1}{n}\sum_{k=1}^{n}g(X_k)$ where $(X_k)_{1\le k\le n}$ is a Markov chain admitting P_0 as its unique stationary distribution. We refer to [22] for a survey on MCMC algorithms. Usually, one has to consider a burn-in

period to deal with the bias $|\mathbb{E}[g]-\mathbb{E}_x[g]|$ due to the arbitrary choice of the initial state x of the Markov chain. However, recent algorithms based on regeneration schemes start automatically under the stationary distribution, see [21] and [5] for instance. We will only focus on such algorithms to avoid the issue of the burn-in period and corresponding quantitative bounds on the bias $|\mathbb{E}_x[g] - \mathbb{E}[g]|$.

3.1 Estimation errors for MCMC algorithms

An interesting case is when $|g|$ is proportional to a drift function $L|g| = V$. In the stationary case, we have

$$\mathbb{E}\Big[\exp\Big(\lambda\sum_{k=1}^{n}(g(X_k)-\mathbb{E}[g]) - \frac{1+\bar\beta(R)((R-1)/c(R)-R)}{1-\bar\beta(R)}\sum_{k=1}^{n}\frac{\lambda^2}{2}(Pg_k^2+g_k^2)\Big)\Big] \le 1.$$

Notice that the square integrability of g is satisfied if g^2 is also proportional to a Lyapunov's function. Then the mean ergodic theorem applies and we obtain the a.s. convergence

$$\frac{1+\bar\beta(R)((R-1)/c(R)-R)}{1-\bar\beta(R)}\frac{1}{2n}\sum_{k=1}^{n}Pg_k^2+g_k^2 \to_{k\to\infty} \frac{1+\bar\beta(R)((R-1)/c(R)-R)}{1-\bar\beta(R)}\mathbb{E}_\pi[g^2].$$

Moreover, the CLT applies and $(\sum_{k=1}^n g(X_k)-\mathbb{E}_x[g])/\sqrt{n} \to^d \sigma^2(g)N$ where $N\sim\mathcal{N}(0,1)$ and the asymptotic variance $\sigma^2(g)$ can be expressed as

$$\sigma^2(g) = \mathrm{Var}_\pi[g^2] + 2\sum_{k=1}^{\infty}\mathrm{Cov}_\pi[g(X_0),g(X_k)].$$

Thus, if one could consider the exponential inequality asymptotically, one would obtain

$$\mathbb{E}[\exp(\lambda\sigma(g)N)] \le \exp\left(\lambda^2\frac{1+\bar\beta(R)((R-1)/c(R)-R)}{1-\bar\beta(R)}\mathbb{E}_\pi[g^2]\right), \qquad \lambda>0.$$

The quantity $(1+\bar\beta(R)((R-1)/c(R)-R))\mathbb{E}_\pi[g^2]/(1-\bar\beta(R))$ appears as a natural over estimator of $\sigma^2(g)/2$. Similar upper bounds have been derived under the spectral gap condition in [23] and under the Ricci curvature condition in [14]. The spectral gap assumption relies on the control of the correlations for any square integrable functions of the Markov chain. The Ricci curvature condition relies on the contraction properties of any Lipschitz functions of the Markov chain. The advantage of the drift condition's approach is that the constants b and β are related only with the Lyapunov's function V. So the estimate can be much sharper if the Lyapunov's function can be well chosen, i.e. close to g. A careful look at the proof of Theorem 2.1 shows that, using our coupling argument, one can improve the over estimator of the asymptotic variance to

$$\sigma^2(g) \le \mathbb{E}_\pi[g^2]\left(1+\frac{2\bar\beta(R)}{1-\bar\beta(R)}\right)+\frac{\bar\beta(R)((R-1)/c-R)}{1-\bar\beta(R)}.$$

Better upper bounds for the asymptotic variance have already been obtained in [15] by a direct application of the Nummelin's scheme on $(X_1, \ldots, X_n)$ (and not on the coupling scheme). It is an open question if such sharper over estimators of the asymptotic variance satisfy an empirical Bernstein's inequality similar than (2.4). It seems that our large over estimation (see Section 4 for numerical values) is partly due to the fact that the approximation of $\mathbb{E}_\pi[g^2]$ can be quite unstable (eventually $\mathbb{E}_\pi[|g|^{2+\delta}] = \infty$ for all $\delta > 0$) but also because the coupling technique used in the proof seems very conservative, see discussions in Section 4.

3.2 Confidence interval for the regenerative Metropolis algorithm

We consider the Random Walk Metropolis algorithm to simulate a Markov chain $(X_k)_{1 \le k \le n}$ on $E = \mathbb{R}^d$, $d \ge 1$, with stationary distribution P proportional to some positive continuous function π given. For some continuous symmetric positive density q one simulates Z_k iid and U_k iid uniform on $[0, 1]$ and independent of the Z_k. Then one computes recursively the Markov chain X_k from the relation

$$X_k = X_{k-1} + Z_k \mathbb{1}_{U_k \le \min(1, \pi(X_{k-1}+Z_k)/\pi(X_k))}, \qquad k \ge 1, \qquad X_0 = x.$$

Mengersen and Tweedie provide in [20] sufficient conditions (that are almost necessary) on π for the geometric ergodicity of the Random Walk Metropolis algorithm, the α log-concavity in the tails assumption ($\alpha > 0$): there exists $x_1 > 0$ such that

$$\frac{\pi(y)}{\pi(x)} \le \exp(-\alpha(|y| - |x|)), \qquad |y| > |x| > x_1, \tag{3.1}$$

where $|\cdot|$ is some norm on E. Let us recall the result in Theorem 3.2 of [20]:

Theorem 3.1. *If $d = 1$, π satisfies (3.1) and $q(x) \le be^{-\alpha|x|}$ for some $\alpha > 0$ then the Random Walk Metropolis algorithm is geometrically ergodic with the drift function $V(x) = e^{s|x|}$, $s < \alpha$.*

To overcome the bias issue we simulate under the stationary measure using the regenerative Metropolis algorithm of Brockwell and Kadane [5] in a simple version (the algorithm 1 in [5] with q as the re-entry proposal distribution). It creates an artificial atom that has to be removed to obtain the Markov chain $(X_k)_{1 \le k \le n}$. The visits to the atom corresponds to the state $A = 1$. The chain X_k is only updated outside the atom when $A = 0$. The only drawback of the approach is that it requires more than n steps to obtain $(X_k)_{1 \le k \le n}$ because of the rejection steps. To overcome this issue, one can use a parrallelized version of the algorithm, see [4]. Let (V_k) be iid uniform over $[0, 1]$ independent of the Z_ks and the U_ks. The pseudo code of the algorithm is given in Figure 1. The idea is to mix the rejection sampling and the Metropolis Random Walk algorithm. Doing so, the rejection step is very robust to the choice of the constant $k > 0$ in the threshold $\pi/(kq)$. Here we assume that $k = 1$ for simplicity. The algorithm simulates automatically the Markov chain under the stationary measure. It also appears that the rejection step increases the

Initialization $A = 1$.
Compute recursively X_{i_k}, $k \geq 1$,

- if $A = 1$
  - if $V_k < \pi(Z_k)/q(Z_k)$ then $X_{i_k} = Z_k$ and $A = 0$,
  - else $A = 1$.
- if $A = 0$ then $Y_k = X_{i_{k-1}} + Z_k \mathbb{I}_{U_k \leq \pi(X_{i_{k-1}} + Z_k)/\pi(X_{i_{k-1}}))}$.
  - if $V_k > q(Y_k)/\pi(Y_k)$ then $X_{i_k} = Y_k$ and $A = 0$
  - else $A = 1$.

Figure 1: the hybrid Metropolis Random Walk algorithm

irreducible property of the chain. In particular, the Markov chain satisfies condition (2.2) on $\{V \leq R\}$ for any $R > 0$ with $\nu = q$ and

$$c(R) = \mathbb{E}[\pi(Z)/q(Z) \wedge 1] \min \left((1 - \mathbb{E}[\pi(Z)/q(Z) \wedge 1]), \min_{\{V(x) \leq R\}} \mathbb{E}[q(x + Z)]/\pi_\infty \right), \quad (3.2)$$

for π_∞ satisfying $\pi(x) \leq \pi_\infty$, $x \in E$.

Define as above the Lapounov's function $V(x) = e^{s|x|}$, $x \in \mathbb{R}$, and denote $\|g\|_V = \sup_{x \in E} |g(x)|/V(x)$. We have the following result, for $d \geq 1$,

Theorem 3.2. *Assume that* π *satisfies* (3.1) *and* $q(x) \leq Ce^{-\alpha|x|}$ *for some* $C > 0$ *and* $\alpha > 2s$. *Assume that* $R \geq V(x_1)$ *is sufficiently large such that*

$$\bar{\beta}(R) := \mathbb{E}[\exp((s - \alpha)|Z|)] + \frac{2V(x_1)\mathbb{E}[V(Z)]}{1 + R} < 1.$$

Then for any function g *such that* $\|g\|_V < \infty$, *we have, for any* $y > 0$ *and* $n \geq 1$,

$$\left| \frac{1}{n} \sum_{k=1}^{n} g(X_k) - \mathbb{E}[g] \right| \leq \frac{x\|g\|_V}{\sqrt{n}} \sqrt{(\hat{\sigma}_n^2(V) + y) \left(1 + \frac{1}{2} \log(\hat{\sigma}_n^2(V)/y + 1) \right)}, \quad (3.3)$$

with probability $1 - \exp(-x^2/2)$, $x > \sqrt{2}$ *and with the over estimator of the asymptotic variance* $\sigma^2(V)$*:*

$$\hat{\sigma}_n^2(V) := \frac{1 + \bar{\beta}(R)((R-1)/c(R) - R)}{1 - \bar{\beta}(R)} \left(\frac{1 + \mathbb{E}[V^2(Z)]}{n} \sum_{k=1}^{n} V^2(X_k) + \varepsilon_n \right)$$

and $\varepsilon_n = (\mathbb{E}[V^2(X)] - V_n^2 \mathbb{E}[V^2(Z)])/n$ *is considered as a non observable negligible term.*

Proof. We already show in (3.2) that the minorization condition (2.2) is satisfied on the small set $\{V(x) \leq R\}$ with the constant $c(R) \to 0$ as $R \to \infty$.

Let us check that the Markov chain satisfies the drift condition (2.1) with the Lyapunov's function $V(x) = \exp(s|x|)$. First consider the case $A = 1$, then $\mathbb{E}_x[V(X_1)] \leq \mathbb{E}[V(Z)]$, $x \in E$ and $\mathbb{E}[V(Z)] = \mathbb{E}[\exp(s|Z|)]$ is finite because $q(x) \leq be^{-\alpha|x|}$. Second, consider the case $A = 0$ and $|x| > x_1$, then under (3.1) we have

$$\begin{aligned}
\mathbb{E}_x[V(X_1)] &= \mathbb{E}_x[V(X_1)\,\mathbb{1}_{|X_1|\leq|x|}] + \mathbb{E}_x[V(X_1)\,\mathbb{1}_{|X_1|>|x|}] \\
&\leq V(x)P_x(|X_1| \leq |x|) + \mathbb{E}_x\Big[V(x+Z_1)\pi(x+Z_1)/\pi(x)\,\mathbb{1}_{|x+Z_1|>|x|}\Big] \\
&\leq \exp(s|x|)\Big(1 + \mathbb{E}\Big[(\exp((s-\alpha)(|x+Z_1| - |x|)) - 1)\,\mathbb{1}_{|x+Z_1|>|x|}\Big]\Big).
\end{aligned}$$

If $x > 0$, as the integrand is negative we have:

$$\mathbb{E}\Big[(\exp((s-\alpha)(|x+Z_1| - |x|)) - 1)\,\mathbb{1}_{|x+Z_1|>|x|}\Big] \leq \mathbb{E}\Big[(\exp((s-\alpha)Z_1) - 1)\,\mathbb{1}_{Z_1>0}\Big].$$

The same reasoning applies if $x < 0$ and as q is symmetric we obtain

$$\mathbb{E}\Big[(\exp((s-\alpha)(|x+Z_1| - |x|)) - 1)\,\mathbb{1}_{|x+Z_1|>|x|}\Big] \leq \frac{1}{2}\mathbb{E}[\exp((s-\alpha)|Z_1|) - 1].$$

Finally, when $A = 0$ and $|x| \leq x_1$ we use the upper bound $\mathbb{E}_x[V(X_1)] \leq V(x_1)\mathbb{E}[V(Z)]$. Thus, the drift condition (2.1) is satisfied by $V(x) = e^{s|x|}$ with $b_1 = V(x_1)\mathbb{E}[V(Z)]$ and $\beta_1 = (1+\mathbb{E}[\exp(s-\alpha)|Z|)])/2$. Notice that by similar arguments we also have the drift condition (2.1) satisfied by V^2 with $b_2 = V^2(x_1)\mathbb{E}[V^2(Z)]$ and $\beta_2 = (1 + \mathbb{E}[\exp(2s-\alpha)|Z|)])/2$. So PV_k^2 are well defined as the second moments are finite. We apply the stationary version of Theorem 2.1 to obtain

$$\mathbb{E}\Big[\exp\Big(\lambda\sum_{k=1}^{n}(g(X_k) - \mathbb{E}[g]) - \frac{1 + \bar{\beta}(R)((R-1)/c(R) - R)}{1 - \bar{\beta}(R)}\sum_{k=1}^{n}\frac{1}{2}(PV_k^2 + V_k^2)\Big)\Big] \leq 1.$$

As PV_k^2 is not observed, we over estimate it by $V_{k-1}^2\mathbb{E}[V^2(Z)]$ for $2 \leq k \leq n$. The negligible term ε_n correspond to the fact that $PV_1^2 = \mathbb{E}[V^2(X)]$ is replaced by $V_n^2\mathbb{E}[V^2(Z)]$ in the expression of $\hat{\sigma}_n^2(V)$. Finally we apply Corollary 2.2 of [7] to obtain the desired result. □

4 Discussion and simulations study

We provide in this section some discussions accompanied by some simulations study.

Discussion about the Luyapunov's function V: Compared with [6], the approach is very dependent on the choice of the Lyapunov's function V. It is good because then the constants involved can be reasonable if V is well chosen. Moreover, for the MCMC application when $f(X_1, \ldots, X_n) = \sum_{k=1}^{n} g(X_k)$, it seems more efficient to take V as close to $|g|$ as possible, i.e. as small as possible. Indeed, the larger is V and the larger b is in (2.1).

So the larger is R in $\tilde{\beta}$ and the smaller is $c(R)$ in (2.2). By a convex argument, one can actually shows that the drift condition (2.1) holds for all Lyapunov's functions V^p with $0 < p < 1$. So the range of admissible Lyapunov's function is quite large. For instance, in the Metropolis algorithm, any $V(x) = \exp(s|x|)$ for $s < 2\alpha$ is admissible. However, we are not aware of any other Lypunov's functions for this algorithm and the Metropolis algorithm will have good properties for functions g with exponential shape. An interested issue is to know wether, given an unbounded g, one can always find an algorithm such that (2.1) is almost satisfied for $|g|$.

Discussion about the quantitative bounds: The explicit constant in Theorem 2.1 is very large. For instance, the contracting normals toy-example considered in [2] satisfies our conditions; it corresponds to the case of an AR(1) model $X_k = 0.5X_{k-1} + \sqrt{3/4}N_k$ where N_k are iid standard gaussian random variables. The stationary solution is the standard gaussian distribution, $g(x) = x$, $\mathbb{E}[g] = 0$ and $V(x) = 1 + x^2$, see [2] and [15] for more details. Then the constant $K = (1+2\bar{\beta}(R)((R-1)/c(R)-R)/(1-\bar{\beta}(R)) \approx 7,000,000,000$, is larger by 3 orders of magnitude than the constants in [16]. Note that [15] improved the constants of [16] by 5 orders of magnitude. Our bounds are much larger because of the use of our coupling argument. It would be interesting to obtain an empirical Bernstein's inequality by applying the Nummelin's scheme directly on the Markov chain trajectory $(X_1, \ldots, X_n)$.

Those large constants are due to the poor irreducible properties of the toy-example,

$$c(R) = 2(\Phi(\sqrt{3}d) - \Phi(\sqrt{3}/d) \qquad \text{with} \qquad R = \sqrt{2 + (d^2 - 1)/4},$$

see [2] and [16] for details on those elementary computations. As small values of $c(R)$ are the main issue to control the constant, it is worth to improve the irreducibility properties of the Markov chain. The hybrid algorithms as the one of Brockwell and Kadane [5] offer a simple way of increasing $c(R)$. The only drawback is that it also increases the necessary runs in the algorithms to generate a trajectory of fixed length. In figure 2 we compare the MSE computed on 100 Monte Carlo simulations of $n = 10000$ runs of the Hybrid (1), Rejection (2) and Metropolis (3) algorithms. The proposal distribution is the standard gaussian ($d = 1$) and $\pi(x) = e^{-(x-1)^2}$, $x \in \mathbb{R}$. The initial value for the Metropolis algorithm is 0. The bias issue could explain why the Metropolis algorithm is slightly over performed by the hybrid algorithm. The large number of rejects, even if the ratio $\pi/(kq)$ has been optimized, should explain why the Rejection algorithm is over performed by the hybrid algorithm. When $\pi(x) = e^{-x^2}$ then $c(R)$ is reasonable for the hybrid algorithm and $K \approx 2,650,000$. It still requires more than $100 * \log(10) * K * \log(K)/2 \approx 4,500,000,000$ runs for obtaining a confident interval of level 0.1 and of reasonable length $\approx \sigma(V)/10$.

Discussion about the median trick: We based our comparaison with previous quantitative bounds of [16] and [15] above on confident intervals of level 0.1. As the previous

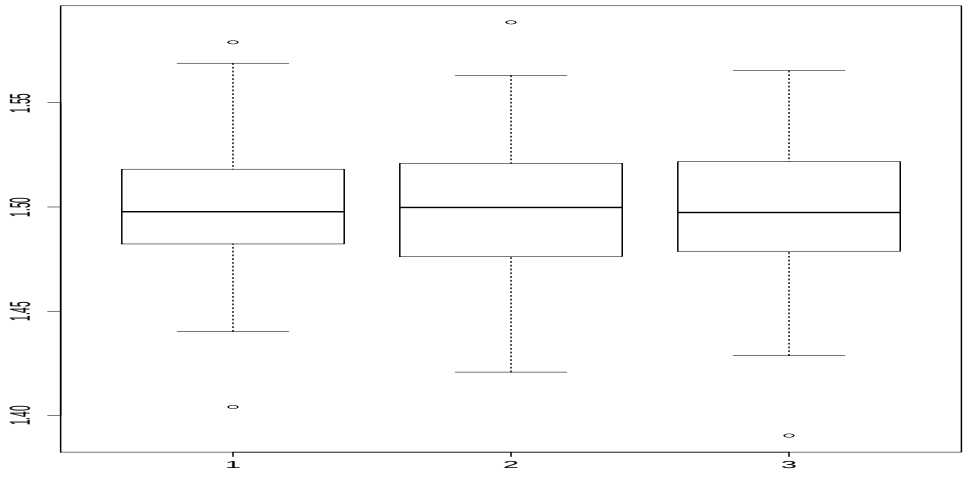

Figure 2: Comparaison of the MSE of the Hybrid (1), Rejection (2) and Metropolis (3) algorithms.

bounds [16] and [15] are based on the Chebychev's inequality

$$\mathbb{P}\left(\left|\sum_{k=1}^{n} g(X_k) - \mathbb{E}[g]\right| > \varepsilon\right) \leq \frac{\|g\|_V^2 \hat{\sigma}_n^2(V)}{n\varepsilon},$$

they are not efficient to produce confidence intervals with small levels. To bypass the problem, the median trick of [13] is used. The trick is to approximate $\mathbb{E}[g]$ thanks to the median of m independent approximations $\frac{1}{n}\sum_{k=1}^{n} g(X_{i,k})$, $1 \leq i \leq m$ of MCMC algorithms with the same confidence interval length of level $a < 1/2$. Then if $m \geq 2\log(\alpha)/\log(4a(1-a))$ the confidence level of the interval around the median is reduced to $\alpha < a$, see Lemma 4.4 in [16]. However, exponential Bernstein's inequalities as (2.4) shows that the interval around the mean of the m independent approximations (based on mn runs) has level $\alpha < a$ when $m \geq \log(\alpha)/\log(a)$. So, when Theorem 2.1 applies, the mean $\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{k=1}^{n} g(X_{i,k})$ seems to have better concentration properties than the median.

References

[1] Adamczak, R. (2008) A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13 (34)**, 1000–1034.

[2] Baxendale, P.H. (2005) Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.* **15** 700–738.

[3] Boucheron, S., Lugosi, G. and Massart, P. (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press.

[4] Brockwell, A. E. (2006). Parallel Markov chain Monte Carlo simulation by prefetching. *Journal of Computational and Graphical Statistics*, **15**(1), 246–261.

[5] Brockwell, A. E. and Kadane, J. B. (2005) Identification of regeneration times in MCMC simulation, with application to adaptive schemes. *Journal of Computational and Graphical Statistics*, **14** (2).

[6] Dedecker J. and Gouëzel S. (2015) Subgaussian concentration inequalities for geometrically ergodic Markov chains. *Electronic Communications in Probability* **20**, 1–12.

[7] de la Pena, V. H., Klass, M. J. and Lai, T. L. (2004) Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Ann. Probab*, **32**, 1902-1933.

[8] Djellout, H., Guillin, A. and Wu, L. (2004) Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *Ann. Probab.* **32 (3B)**, 2702–2732.

[9] Feigin, P.D. and Tweedie, R.L. (1985) Random coefficient autoregressive processes: a Markov chain analysis of stationarity and finiteness of moments. *J. Time Series Anal.* **6**, 1–14.

[10] Gyori, B. M. and Paulin, D. (2012). Non-asymptotic confidence intervals for MCMC in practice. arXiv preprint arXiv:1212.2016.

[11] Hairer, M. and Mattingly, J. C. (2011) Yet another look at HarrisâĂŹ ergodic theorem for Markov chains. In Seminar on Stochastic Analysis, *Random Fields and Applications*, Springer Basel, **VI** 109–117.

[12] Ibragimov, I. A. (1962) Some limit theorems for stationary processes. *Theory Probab. Appl.* **7**, 349–382.

[13] Jerrum M.R., Valiant L.G. and Vizirani V.V. (1986) Random generation of combinatorial structures from a uniform distribution, *Theoret. Comput. Sci.* **43** 169–188.

[14] Joulin, A. and Ollivier, Y. (2010) Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.* **38 (6)**, 2418–2442.

[15] Latuszynski, K., Miasojedow, B. and Niemiro, W. (2013). Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli*, **19 (5A)**, 2033–2066.

[16] Latuszynski, K. and Niemiro, W. (2011). Rigorous confidence bounds for MCMC under a geometric drift condition.*J. Complexity* 27 23–38.

[17] Marton, K. (1996) A measure concentration inequality for contracting Markov chains. *Geom. Funct. Anal.* **6 (3)**, 556–571.

[18] McDiarmid C. (1989) On the method of bounded differences, in: Surveys of Combinatorics, *Siemons J. (Ed.)*, **Lect. Notes Series 141**, London Math. Soc.

[19] Meyn, S.P. and Tweedie, R.L. (1993) *Markov Chains and Stochastic Stability.* Springer, London.

[20] Mengersen, K. L. and Tweedie, R. L. (1996) Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, **24**, 101–121.

[21] Mykland, P., Tierney, L. and Yu, B. (1995) Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, **90**, 233–241.

[22] Roberts, G. O. and Rosenthal, J. S. (2004) General state space Markov chains and MCMC algorithms. *Probability Surveys*, **1**, 20-71.

[23] Rosenthal, J. S. (2003) Asymptotic variance and convergence rates of nearly-periodic Markov chain Monte Carlo algorithms. *JASA*, **98 (461)**, 169-177.

[24] Samson, P.-M. (2000) Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.* **28 (1)**, 416–461.

[25] Villani, C. (2009) *Optimal transport, old and new.* Springer-Verlag, Berlin, 2009.

[26] Wintenberger, O. (2015) Weak transport inequalities and applications to exponential inequalities and oracle inequalities. *Accepted in EJP* arXiv:1207.4951.