Curvature, concentration, and error estimates for Markov chain Monte Carlo

Aldéric Joulin, Yann Ollivier

Abstract

Under a "positive curvature" assumption expressing a kind of metric ergodicity, we provide explicit non-asymptotic estimates for the rate of convergence of empirical means of Markov chains, together with a Gaussian or exponential control on the deviations of empirical means.

The goal of the Markov chain Monte Carlo method is to provide an efficient way to approximate the integral $\pi(f) := \int f(x)\,\pi(\mathrm{d}x)$ of a function f under a finite measure π on some space $\mathcal{X}$. This approach, which has been very successful, consists in constructing a hopefully easy-to-simulate Markov chain $(X_1, X_2, \ldots, X_k, \ldots)$ on $\mathcal{X}$ with stationary distribution π, waiting for a time T_0 (the *burn-in*) so that the chain gets close to its stationary distribution, and then estimating $\pi(f)$ by the empirical mean on the next T steps of the trajectory, with T large enough:

$$\hat{\pi}(f) := \frac{1}{T} \sum_{k=T_0+1}^{T_0+T} f(X_k).$$

We refer e.g. to [RR04] for a review of the topic.

Under suitable assumptions [MT93], it is known that $\hat{\pi}(f)$ almost surely tends to $\pi(f)$ as $T \to \infty$, that the variance of $\hat{\pi}(f)$ decreases asymptotically like $1/T$ and that a central limit theorem holds for the errors $\hat{\pi}(f) - \pi(f)$. Unfortunately, these theorems are asymptotic only, and thus mainly of theoretical interest since they do not allow to give explicit confidence intervals for $\pi(f)$ at a given time T. Some even say that confidence intervals disappeared the day MCMC methods appeared.

In this paper, we aim at establishing rigorous non-asymptotic upper bounds for the error $|\hat{\pi}(f) - \pi(f)|$, which will provide good deviation estimates and confidence intervals for $\pi(f)$. An important point is that we will try to express all results in terms of explicit quantities that are readily computable given a choice of a Markov chain; and, at the same time, recover correct order of magnitudes in a surprising variety of examples.

Our non-asymptotic estimates have the same qualitative behavior as theory predicts in the asymptotic regime: the variance of $\hat{\pi}(f)$ decreases like $1/T$, and the bias decreases exponentially in T_0. Moreover, we provide a Gaussian or exponential control on deviations of $\hat{\pi}(f)$, which allows for good confidence intervals. Finally,

we find that the influence of the choice of the starting point on the variance of $\hat{\pi}(f)$ decreases like $1/T^2$.

Our results hold under an assumption of *positive curvature* [Oll07, Oll09], which can be understood as a kind of "metric ergodicity". Not all Markov chains satisfy this assumption, but important examples include spin systems at high temperature, several of the usual types of waiting queues, processes such as the Ornstein–Uhlenbeck process on $\mathbb{R}^d$ or Brownian motion on positively curved manifolds. We refer to [Oll09] for more examples and discussions on how one can check this assumption, but let us stress out that, at least in principle, this curvature can be computed explicitly given a Markov transition kernel. This property or similar ones can be traced back to Dobrushin [Dob70, DS85], and have been used a few times in the Markov chain literature [CW94, Dob96, BD97, DGW04, Jou07, Oll07, Oll09, Jou, Oli].

Similar concentration inequalities have been recently investigated in [Jou] for time-continuous Markov jump processes. More precisely, the first author obtained Poisson-type tail estimates for Markov processes with positive Wasserstein curvature. Actually, the latter is nothing but a continuous-time version of the Ricci curvature emphasized in the present paper, so that we expect to recover such results by a simple limiting argument (cf. Section 2).

Estimates for the deviations of empirical means have previously been given in [Lez98] using the *spectral gap* of the Markov chain, under different conditions (namely that the chain is reversible, that the law of the initial point has a density w.r.t. π, and that f is bounded). The positive curvature assumption, which is a stronger property than the spectral gap used by Lezaud, allows to lift these restrictions: our results apply to an arbitrary starting point, the function f only has to be Lipschitz, and reversibility plays no particular role. In a series of papers (for instance [Wu00, CG08, GLWY]), the spectral approach has been extended into a general framework for deviations of empirical means using various types of functional inequalities; in particular [GLWY] contains a very nice characterization of asymptotic variance of empirical means of Lipschitz functions in terms of a functional inequality $W_1 I$ satisfied by the invariant distribution.

1 Preliminaries and statement of the results

1.1 Notation

Markov chains. In this paper, we consider a Markov chain $(X_N)_{N\in\mathbb{N}}$ in a Polish (i.e. metric, complete, separable) state space $(\mathcal{X}, d)$. The associated transition kernel is denoted $(P_x)_{x\in\mathcal{X}}$ where each P_x is a probability measure on $\mathcal{X}$, so that $P_x(\mathrm{d}y)$ is the transition probability from x to y. The N-step transition kernel is defined inductively as

$$P_x^N(\mathrm{d}y) := \int_{\mathcal{X}} P_x^{N-1}(\mathrm{d}z)\, P_z(\mathrm{d}y)$$

(with $P_x^1 := P_x$). The distribution at time N of the Markov chain given the initial probability measure μ is the measure μP^N given by

$$\mu P^N(\mathrm{d}y) = \int_{\mathcal{X}} P_x^N(\mathrm{d}y)\, \mu(\mathrm{d}x).$$

Let as usual $\mathbb{E}_x$ denote the expectation of a random variable knowing that the initial point of the Markov chain is x. For any measurable function $f : \mathcal{X} \to \mathbb{R}$, define the iterated averaging operator as

$$P^N f(x) := \mathbb{E}_x f(X_N) = \int_{\mathcal{X}} f(y)\, P_x^N(\mathrm{d}y), \qquad x \in \mathcal{X}.$$

A probability measure π on $\mathcal{X}$ is said to be *invariant* for the chain if $\pi = \pi P$. Under suitable assumptions on the Markov chain $(X_N)_{N\in\mathbb{N}}$, such an invariant measure π exists and is unique, as we will see below.

Denote by $\mathcal{P}_d(\mathcal{X})$ the set of those probability measures μ on $\mathcal{X}$ such that $\int_{\mathcal{X}} d(y, x_0)\, \mu(\mathrm{d}y) < \infty$ for some (or equivalently for all) $x_0 \in \mathcal{X}$. We will always assume that the map $x \mapsto P_x$ is measurable, and that $P_x \in \mathcal{P}_d(\mathcal{X})$ for every $x \in \mathcal{X}$. These assumptions are always satisfied in practice.

Wasserstein distance. The L^1 transportation distance, or Wasserstein distance, between two probability measures $\mu_1, \mu_2 \in \mathcal{P}_d(\mathcal{X})$ represents the "best" way to send μ_1 on μ_2 so that on average, points are moved by the smallest possible distance. It is defined [Vil03] as

$$W_1(\mu_1, \mu_2) := \inf_{\xi \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{X}} \int_{\mathcal{X}} d(x, y)\, \xi(\mathrm{d}x, \mathrm{d}y),$$

where $\Pi(\mu_1, \mu_2)$ is the set of probability measures ξ on $\mathcal{P}_d(\mathcal{X} \times \mathcal{X})$ with marginals μ_1 and μ_2, i.e. such that $\int_y \xi(\mathrm{d}x, \mathrm{d}y) = \mu_1(\mathrm{d}x)$ and $\int_x \xi(\mathrm{d}x, \mathrm{d}y) = \mu_2(\mathrm{d}y)$. (So intuitively $\xi(\mathrm{d}x, \mathrm{d}y)$ represents the amount of mass travelling from x to y.)

Ricci curvature of a Markov chain. Our main assumption in this paper is the following, which can be seen geometrically as a "positive Ricci curvature" [Oll09] property of the Markov chain.

Standing assumption.
There exists $\kappa > 0$ such that

$$W_1(P_x, P_y) \leqslant (1 - \kappa)\, d(x, y)$$

for any $x, y \in \mathcal{X}$.

In practice, it is not necessary to compute the exact value of the Wasserstein distance $W_1(P_x, P_y)$: it is enough to exhibit one choice of $\xi(\mathrm{d}x, \mathrm{d}y)$ providing a good value of $W_1(P_x, P_y)$.

An important remark is that on a "geodesic" space $\mathcal{X}$, it is sufficient to control $W_1(P_x, P_y)$ only for nearby points $x, y \in \mathcal{X}$, and not for all pairs of points (Proposition 19 in [Oll09]). For instance, on a graph it is enough to check the assumption on pairs of neighbors.

These remarks make the assumption possible to check in practice, as we will see from the examples below.

More notation: eccentricity, diffusion constant, local dimension, granularity. Under the assumption above, Corollary 21 in [Oll09] entails the existence of a unique invariant measure $\pi \in \mathcal{P}_d(\mathcal{X})$, with moreover the following geometric ergodicity in W_1-distance (instead of the classical total variation distance, which is obtained by choosing the trivial metric $d(x, y) = 1_{\{x \neq y\}}$):

$$W_1(\mu P^N, \pi) \leqslant (1 - \kappa)^N W_1(\mu, \pi), \tag{1}$$

and in particular

$$W_1(P_x^N, \pi) \leqslant (1 - \kappa)^N E(x), \tag{2}$$

where the *eccentricity* E at point $x \in \mathcal{X}$ is defined as

$$E(x) := \int_{\mathcal{X}} d(x, y)\, \pi(\mathrm{d}y).$$

Note that eccentricity satisfies the bounds [Oll09]:

$$E(x) \leqslant \begin{cases} \operatorname{diam} \mathcal{X}; \\ E(x_0) + d(x, x_0), \quad x_0 \in \mathcal{X}; \\ \frac{1}{\kappa} \int_{\mathcal{X}} d(x, y)\, P_x(\mathrm{d}y). \end{cases}$$

Such *a priori* estimates are useful in various settings. In particular, the last one is "local" in the sense that it is easily computable given the Markov kernel P_x.

Let us also introduce the *coarse diffusion constant* $\sigma(x)$ of the Markov chain at a point $x \in \mathcal{X}$, which controls the size of the steps, defined by

$$\sigma(x)^2 := \frac{1}{2} \iint d(y, z)^2\, P_x(\mathrm{d}y) P_x(\mathrm{d}z).$$

Let the *local dimension* n_x at point $x \in \mathcal{X}$ be given by

$$\begin{aligned} n_x &:= \inf_{\substack{f : \mathcal{X} \to \mathbb{R} \\ f \text{ 1-Lipschitz}}} \frac{\iint d(y, z)^2\, P_x(\mathrm{d}y)\, P_x(\mathrm{d}z)}{\iint d(f(y), f(z))^2\, P_x(\mathrm{d}y)\, P_x(\mathrm{d}z)} \\ &\geqslant 1. \end{aligned}$$

Let the *granularity* of the Markov chain be

$$\sigma_\infty := \frac{1}{2} \sup_{x \in \mathcal{X}} \operatorname{diam} \operatorname{Supp} P_x$$

which we will often assume to be finite.

For example, for the simple random walk in a graph we have $\sigma_\infty \leqslant 1$ and $\sigma(x)^2 \leqslant 2$.

Finally, we will denote by $\|\cdot\|_{\mathrm{Lip}}$ the usual Lipschitz seminorm of a function f on $\mathcal{X}$:

$$\|f\|_{\mathrm{Lip}} := \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)}.$$

1.2 Results

Back to the introduction, choosing integers $T \geqslant 1$ and $T_0 \geqslant 0$ and setting

$$\hat{\pi}(f) := \frac{1}{T} \sum_{k=T_0+1}^{T_0+T} f(X_k),$$

the purpose of this paper is to understand how fast the difference $|\hat{\pi}(f) - \pi(f)|$ goes to 0 as T goes to infinity, for a large class of functions f. Namely, we will consider Lipschitz functions (recall that Corollary 21 in [Oll09] implies that all Lipschitz functions are π-integrable).

Bias and non-asymptotic variance. Our first interest is in the non-asymptotic mean quadratic error

$$\mathbb{E}_x \left[|\hat{\pi}(f) - \pi(f)|^2 \right]$$

given any starting point $x \in \mathcal{X}$ for the Markov chain.

There are two contributions to this error: a *variance* part, controlling how $\hat{\pi}(f)$ differs between two independent runs both starting at x, and a *bias* part, which is the difference between $\pi(f)$ and the average value of $\hat{\pi}(f)$ starting at x. Namely, the mean quadratic error decomposes as the sum of the squared bias plus the variance:

$$\mathbb{E}_x \left[|\hat{\pi}(f) - \pi(f)|^2 \right] = |\mathbb{E}_x \hat{\pi}(f) - \pi(f)|^2 + \mathrm{Var}_x\, \hat{\pi}(f) \tag{3}$$

where $\mathrm{Var}_x\, \hat{\pi}(f) := \mathbb{E}_x \left[|\hat{\pi}(f) - \mathbb{E}_x \hat{\pi}(f)|^2 \right]$.

As we will see, these two terms have different behaviors depending on T_0 and T. For instance, the bias is expected to decrease exponentially fast as the burn-in period T_0 is large, whereas if T is fixed, the variance term does not vanish as $T_0 \to \infty$.

Let us start with control of the bias term, which depends, of course, on the starting point of the Markov chain. All proofs are postponed to Section 3.

Proposition 1 (Bias of empirical means).
For any Lipschitz function $f : \mathcal{X} \to \mathbb{R}$, we have the upper bound on the bias:

$$|\mathbb{E}_x \hat{\pi}(f) - \pi(f)| \leqslant \frac{(1-\kappa)^{T_0+1}}{\kappa T} E(x)\, \|f\|_{\mathrm{Lip}}\,. \tag{4}$$

The variance term is more delicate to control. For comparison, let us first mention that under the invariant measure π, the variance of a Lipschitz function f is bounded as follows (and this estimate is often sharp):

$$\operatorname{Var}_\pi f \leqslant \|f\|_{\mathrm{Lip}}^2 \sup_{x\in\mathcal{X}} \frac{\sigma(x)^2}{n_x\kappa} \tag{5}$$

(cf. Lemma 9 below or Proposition 32 in [Oll09]). This implies that, were one able to sample from the invariant distribution π, the ordinary Monte Carlo method of estimating $\pi(f)$ by the average over T independent samples would yield a variance bounded by $\frac{\|f\|_{\mathrm{Lip}}^2}{T} \sup_{x\in\mathcal{X}} \frac{\sigma(x)^2}{n_x\kappa}$. Because of correlations, this does not hold for the MCMC method. Nevertheless we get the following.

Theorem 2 (Variance of empirical means, 1).
Provided the inequalities make sense, we have

$$\operatorname{Var}_x \hat{\pi}(f) \leqslant \begin{cases} \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa T} \sup_{x\in\mathcal{X}} \frac{\sigma(x)^2}{n_x\kappa} & \text{if } T_0 = 0; \\ \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa T} \left(1 + \frac{1}{\kappa T}\right) \sup_{x\in\mathcal{X}} \frac{\sigma(x)^2}{n_x\kappa} & \text{otherwise.} \end{cases} \tag{6}$$

The most important feature of this formula is the $1/\kappa T$ factor, which means there is an additional $1/\kappa$ factor with respect to the ordinary Monte Carlo case. Intuitively, the idea is that correlations disappear after roughly $1/\kappa$ steps, and so T steps of the MCMC method are "worth" only κT independent samples. This $1/\kappa T$ factor will appear repeatedly in our text.

To get convinced that this $1/\kappa T$ factor in our estimate (6) is natural, observe that if the burn-in T_0 is large enough, then the law of X_{T_0} will be close to the invariant distribution π so that $\operatorname{Var}_x \hat{\pi}(f)$ will behave like $\operatorname{Var}_{X_0\sim\pi} \hat{\pi}(f)$. Then we have

$$\operatorname{Var}_{X_0\sim\pi} \hat{\pi}(f) = \frac{1}{T^2} \left(\sum_{i=1}^{T} \operatorname{Var}_{X_0\sim\pi}(f(X_i)) + 2 \sum_{1\leqslant i<j\leqslant T} \operatorname{Cov}_{X_0\sim\pi}(f(X_i), f(X_j)) \right)$$

but our assumption on κ easily implies that correlations decrease exponentially fast with rate $1-\kappa$ so that (at least in the reversible case) we have $\operatorname{Cov}_{X_0\sim\pi}(f(X_0), f(X_t)) \leqslant (1-\kappa)^t \operatorname{Var}_{X_0\sim\pi} f(X_0)$. In particular, for any fixed i we have $\operatorname{Var}_{X_0\sim\pi}(f(X_i)) + 2\sum_{j>i} \operatorname{Cov}_{X_0\sim\pi}(f(X_i), f(X_j)) \leqslant \frac{2}{\kappa} \operatorname{Var}_{X_0\sim\pi} f(X_i)$. Plugging this into the above yields $\operatorname{Var}_{X_0\sim\pi} \hat{\pi}(f) \leqslant \frac{2}{\kappa T} \operatorname{Var}_\pi f$, which explains the $1/\kappa T$ factor.

Unbounded diffusion constant. In the formulas above, a supremum of $\sigma(x)^2/n_x$ appeared. This is fine when considering e.g. the simple random walk on a graph, because then $\sigma(x)^2/n_x \approx 1$ for all $x \in \mathcal{X}$. However, in some situations (for instance

binomial distributions on the cube), this supremum is much larger than a typical value, and, in some continuous-time limits on infinite spaces, the supremum may even be infinite (as in the example of the $M/M/\infty$ queueing process below). For such situations, one expects the variance to depend, asymptotically, on the average of $\sigma(x)^2/n_x$ under the invariant measure π, rather than its supremum.

The next result generalizes Theorem 2 to Markov chains with unbounded diffusion constant $\sigma(x)^2/n_x$. We will assume that $\sigma(x)^2/n_x$ has at most linear growth (this is consistent with the usual theorems on Markov processes, in which linear growth on the coefficients of the diffusion is usually assumed).

Of course, if one starts the Markov chain at a point x with large $\sigma(x)^2$, meaning that the chain has a large diffusion constant at the beginning, then the variance of the empirical means started at x will be accordingly large, at least for small T. This gives rise, in the estimates below, to a variance term depending on x; it so happens that this terms decreases like $1/T^2$ with time.

Theorem 3 (Variance of empirical means, 2).
Assume that there exists a Lipschitz function S with $\|S\|_{\mathrm{Lip}} \leqslant C$ such that

$$\frac{\sigma(x)^2}{n_x \kappa} \leqslant S(x), \quad x \in \mathcal{X}.$$

Then the variance of the empirical mean is bounded as follows:

$$\mathrm{Var}_x\, \hat{\pi}(f) \leqslant \begin{cases} \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa T} \left(\mathbb{E}_\pi S + \frac{C}{\kappa T} E(x) \right) & \text{if } T_0 = 0; \\ \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa T} \left(\left(1 + \frac{1}{\kappa T}\right) \mathbb{E}_\pi S + \frac{2C(1-\kappa)^{T_0}}{\kappa T} E(x) \right) & \text{otherwise.} \end{cases} \tag{7}$$

In particular, the upper bound behaves asymptotically like $\|f\|_{\mathrm{Lip}}^2\, \mathbb{E}_\pi S/\kappa T$, with a correction of order $1/(\kappa T)^2$ depending on the initial point x.

Note that in some situations, $\mathbb{E}_\pi S$ is known in advance from theoretical reasons. In general, it is always possible to chose any origin $x_0 \in \mathcal{X}$ (which may or may not be the initial point x) and apply the estimate $\mathbb{E}_\pi S \leqslant S(x_0) + CE(x_0)$.

Concentration results. To get good confidence intervals for $\hat{\pi}(f)$ it is necessary to investigate deviations, i.e. the behavior of the probabilities

$$\mathbb{P}_x \left(|\hat{\pi}(f) - \pi(f)| > r \right),$$

which reduce to deviations of the centered empirical mean

$$\mathbb{P}_x \left(|\hat{\pi}(f) - \mathbb{E}_x \hat{\pi}(f)| > r \right)$$

if the bias is known. Of course the Bienaymé–Chebyshev inequality states that $\mathbb{P}_x \left(|\hat{\pi}(f) - \mathbb{E}_x \hat{\pi}(f)| > r \right) \leqslant \frac{\mathrm{Var}_x\, \hat{\pi}(f)}{r^2}$, but this does not decrease very fast with r, and Gaussian-type deviation estimates are often necessary to get good confidence

intervals. Our next results show that the probability of a deviation of size r is bounded by an explicit Gaussian or exponential term. (The same Gaussian-exponential transition also appears in [Lez98] and other works.)

Of course, deviations for the function $10f$ are 10 times bigger than deviations for f, so we will use the rescaled deviation $\frac{\hat{\pi}(f)-\mathbb{E}_x\hat{\pi}(f)}{\|f\|_{\mathrm{Lip}}}$.

In the sequel, we assume that $\sigma_\infty < \infty$. Once more the proofs of the following results are established in Section 3.

Theorem 4 (Concentration of empirical means, 1).
Denote by V^2 the quantity:

$$V^2 := \frac{1}{\kappa T}\left(1+\frac{T_0}{T}\right)\sup_{x\in\mathcal{X}}\frac{\sigma(x)^2}{n_x\kappa}.$$

Then empirical means satisfy the following concentration result:

$$\mathbb{P}_x\left(\frac{|\hat{\pi}(f)-\mathbb{E}_x\hat{\pi}(f)|}{\|f\|_{\mathrm{Lip}}}\geqslant r\right)\leqslant\begin{cases}2\,e^{-\frac{r^2}{16V^2}} & \text{if}\quad r\in(0,r_{max})\\ 2\,e^{-\frac{\kappa T r}{12\sigma_\infty}} & \text{if}\quad r\geqslant r_{max}\end{cases}\tag{8}$$

where the boundary of the Gaussian window is given by $r_{max} := 4V^2\kappa T/3\sigma_\infty$.

Note that $r_{\max}\gg 1/\sqrt{T}$ for large T, so that the Gaussian window gets better and better when normalizing by the standard deviation. This is in accordance with the central limit theorem for Markov chains [MT93].

As for the case of variance above, we also provide an estimate using the average of $\sigma(x)^2/n_x$ rather than its supremum.

Theorem 5 (Concentration of empirical means, 2).
Assume that there exists a Lipschitz function S with $\|S\|_{\mathrm{Lip}}\leqslant C$ such that

$$\frac{\sigma(x)^2}{n_x\kappa}\leqslant S(x),\quad x\in\mathcal{X}.$$

Denote by V_x^2 the following term depending on the initial condition x:

$$V_x^2 := \frac{1}{\kappa T}\left(1+\frac{T_0}{T}\right)\mathbb{E}_\pi S+\frac{CE(x)}{\kappa^2T^2}.$$

Then the following concentration inequality holds:

$$\mathbb{P}_x\left(\frac{|\hat{\pi}(f)-\mathbb{E}_x\hat{\pi}(f)|}{\|f\|_{\mathrm{Lip}}}\geqslant r\right)\leqslant\begin{cases}2\,e^{-\frac{r^2}{16V_x^2}} & \text{if}\quad r\in(0,r_{max})\\ 2\,e^{-\frac{\kappa T r}{4\max\{2C,3\sigma_\infty\}}} & \text{if}\quad r\geqslant_{max}\end{cases}\tag{9}$$

where $r_{max} := 4V_x^2\kappa T/\max\{2C,3\sigma_\infty\}$.

The two quantities V^2 and V_x^2 in these theorems are essentially similar to the estimates of the empirical variance $\mathrm{Var}_x\,\hat{\pi}(f)$ given in Theorems 2 and 3, so that the same comments apply.

Randomizing the starting point. As can be seen from the above, we are mainly concerned with Markov chains starting from a deterministic point. The random case might be treated as follows. Assume that the starting point X_0 of the chain is taken at random according to some probability measure μ. Then an additional variance term appears in the variance/bias decomposition (3), namely:

$$\begin{aligned}\mathbb{E}_\mu\left[|\hat{\pi}(f)-\pi(f)|^2\right] &= |\mathbb{E}_\mu\hat{\pi}(f)-\pi(f)|^2+\int_{\mathcal{X}}\operatorname{Var}_x\hat{\pi}(f)\,\mu(\mathrm{d}x)\\ &\quad+\operatorname{Var}\left[\mathbb{E}\left(\hat{\pi}(f)|X_0\right)\right].\end{aligned}$$

The new variance term depends on how "spread" the initial distribution μ is and can be easily bounded. Indeed we have

$$\mathbb{E}\left(\hat{\pi}(f)\mid X_0=x\right)=\frac{1}{T}\sum_{k=T_0+1}^{T_0+T}P^kf(x)$$

so that if f is, say, 1-Lipschitz,

$$\begin{aligned}\operatorname{Var}\left[\mathbb{E}\left(\hat{\pi}(f)|X_0\right)\right] &= \frac{1}{2T^2}\int_{\mathcal{X}}\int_{\mathcal{X}}\left|\sum_{k=T_0+1}^{T_0+T}\left(P^kf(x)-P^kf(y)\right)\right|^2\mu(\mathrm{d}x)\mu(\mathrm{d}y)\\ &\leqslant \frac{(1-\kappa)^{2(T_0+1)}}{2\kappa^2T^2}\int_{\mathcal{X}}\int_{\mathcal{X}}d(x,y)^2\,\mu(\mathrm{d}x)\mu(\mathrm{d}y),\end{aligned}$$

since P^kf is $(1-\kappa)^k$-Lipschitz. This is fast-decreasing both in T_0 and T.

Note also that the bias can be significantly reduced if μ is known, for some reason, to be close to the invariant measure π. More precisely, the eccentricity $E(x)$ in the bias formula (4) above is replaced with the L^1 transportation distance $W_1(\mu,\pi)$.

Convergence of $\hat{\pi}$ to π. The fact that $\hat{\pi}$ yields estimates close to π when integrating Lipschitz functions does not mean that $\hat{\pi}$ itself is close to π. To see this, consider the simple case when $\mathcal{X}$ is a set of N elements equipped with any metric. Consider the trivial Markov chain on $\mathcal{X}$ which sends every point $x\in\mathcal{X}$ to the uniform probability measure on $\mathcal{X}$ (so that $\kappa=1$ and the MCMC method reduces to the ordinary Monte Carlo method). Then it is clear that for any function f, the bias vanishes and the empirical variance is

$$\operatorname{Var}_x\hat{\pi}(f)=\frac{1}{T}\operatorname{Var}_\pi f$$

which in particular does not depend directly on N and allows to estimate $\pi(f)$ with a sample size independent of N, as is well-known to any statistician. But the empirical measure $\hat{\pi}$ is a sum of Dirac masses at T points, so that its Wasserstein distance to the uniform measure cannot be small unless T is comparable to N.

This may seem to contradict the Kantorovich–Rubinstein duality theorem [Vil03], which states that

$$W_1(\hat{\pi}, \pi) = \sup_{f \text{ 1-Lipschitz}} \hat{\pi}(f) - \pi(f).$$

Indeed, we know that for a function f fixed in advance, very probably $\hat{\pi}(f)$ is close to $\pi(f)$. But for every realization of the random measure $\hat{\pi}$ there may be a particular function f yielding a large error. What is true, is that the *averaged* empirical measure $\mathbb{E}_x\hat{\pi}$ starting at x tends to π fast enough, namely

$$W_1(\mathbb{E}_x\hat{\pi}, \pi) \leqslant \frac{(1-\kappa)^{T_0+1}}{\kappa T}\, E(x)$$

which is just a restatement of our bias estimate above (Proposition 1). But as we have just seen, $\mathbb{E}_x W_1(\hat{\pi}, \pi)$ is generally much larger.

2 Examples and applications

We now show how these results can be applied to various settings where the positive curvature assumption is satisfied, ranging from discrete product spaces to waiting queues, diffusions on $\mathbb{R}^d$ or manifolds, and spin systems. In several examples, our results improve on the literature.

A simple example: discrete product spaces. Let us first consider a very simple example. This is mainly illustrative, as in this case the invariant measure π is very easy to simulate. Let $\mathcal{X} = \{0,1\}^N$ be the space of N-bit sequences equipped with the uniform probability measure. We shall use the Hamming distance on $\mathcal{X}$, namely, the distance between two sequences of 0's and 1's is the number of positions at which they differ. The Markov chain we shall consider consists, at each step, in choosing a position $1 \leqslant i \leqslant N$ at random, and replacing the i-th bit of the sequence with either a 0 or a 1 with probability 1/2. Namely, starting at $x = (x_1, \dots, x_N)$ we have $P_x(x) = 1/2$ and $P_x(x_1, \dots, 1 - x_i, \dots, x_N) = 1/2N$.

A typical Lipschitz function for this example is the function f_0 equal to the proportion of "0" bits in the sequence, for which $\|f_0\|_{\text{Lip}} = 1/N$.

Then an elementary computation (Example 8 in [Oll09]) shows that $\kappa = 1/N$, so that our theorems apply. The various quantities of interest are estimated as $\sigma_\infty = 1$, $\sigma(x)^2 \leqslant 2$ and $n_x \geqslant 1$; using Remark 40 in [Oll09] yields a slightly better estimate $\frac{\sigma(x)^2}{n_x} \leqslant 1/2$. Moreover $E(x) = N/2$ for any $x \in \mathcal{X}$.

Then our bias estimate (4) for a Lipschitz function f is

$$|\mathbb{E}_x\hat{\pi}(f) - \pi(f)| \leqslant \frac{N^2}{2T}\,(1-1/N)^{T_0+1}\ \|f\|_{\text{Lip}} \leqslant \frac{N^2}{2T}\, e^{-T_0/N}\ \|f\|_{\text{Lip}}$$

So taking $T_0 \approx 2N \log N$ is enough to ensure small bias. This estimate of the mixing time is known to be the correct order of magnitude: indeed, if each bit has

been updated at least once (which occurs after a time $\approx N \log N$) then the measure is exactly the invariant measure and so, under this event, the bias exactly vanishes. In contrast, the classical estimate using the spectral gap yields only $O(N^2)$ for the mixing time [DS96].

The variance estimate (6) reads

$$\operatorname{Var} \hat{\pi}(f) \leqslant \frac{N^2}{2T}(1 + N/T) \ \|f\|_{\mathrm{Lip}}^2$$

so that, for example for the function f_0 above, taking $T \approx N$ will yield a variance $\approx 1/N$, the same order of magnitude as the variance of f_0 under the uniform measure. (With a little work, one can convince oneself that this order of magnitude is correct for large T.)

The concentration result (8) reads, say with $T_0 = 0$ and for the Gaussian part:

$$\mathbb{P}_x \left(|\hat{\pi}(f) - \mathbb{E}_x \hat{\pi}(f)| \geqslant r \right) \leqslant 2e^{-Tr^2/8N^2 \|f\|_{\mathrm{Lip}}^2}$$

so that e.g. for f_0 we simply get get $2e^{-Tr^2/8}$. For comparison, the spectral estimate from [Lez98] behaves like $2^{N/2} e^{-Tr^2/4}$ for small r, so that we roughly improve the estimate by a factor $2^{N/2}$, due to the fact that the density of the law of the starting point (a Dirac mass) plays no role in our setting.

Heat bath for the Ising model. Let G be a finite graph. Consider the classical Ising model from statistical mechanics [Mar04], namely the configuration space $\mathcal{X} := \{-1, 1\}^G$ together with the energy function $U(s) := -\sum_{x \sim y \in G} s(x)s(y) - h \sum_{x \in G} s(x)$ for $s \in \mathcal{X}$, where $h \in \mathbb{R}$. For some $\beta \geqslant 0$, equip $\mathcal{X}$ with the Gibbs distribution $\pi := e^{-\beta U}/Z$ where as usual $Z := \sum_s e^{-\beta U(s)}$. The distance between two states is defined as the number of vertices of G at which their values differ, namely $d(s, s') := \frac{1}{2} \sum_{x \in G} |s(x) - s'(x)|$.

For $s \in \mathcal{X}$ and $x \in G$, denote by s_{x+} and s_{x-} the states obtained from s by setting $s_{x+}(x) = +1$ and $s_{x-}(x) = -1$, respectively. Consider the following random walk on $\mathcal{X}$, known as the *heat bath* or *Glauber dynamics* [Mar04]: at each step, a vertex $x \in G$ is chosen at random, and a new value for $s(x)$ is picked according to local equilibrium, i.e. $s(x)$ is set to 1 or -1 with probabilities proportional to $e^{-\beta U(s_{x+})}$ and $e^{-\beta U(s_{x-})}$ respectively (note that only the neighbors of x influence the ratio of these probabilities). The Gibbs distribution π is invariant (and reversible) for this Markov chain.

When $\beta = 0$, this Markov chain is identical to the Markov chain on $\{0, 1\}^N$ described above, with $N = |G|$. Therefore, it comes as no surprise that for β small enough, curvature is positive. More precisely one finds [Oll09]

$$\kappa \geqslant \frac{1}{|G|} \left(1 - v_{\max} \frac{e^{\beta} - e^{-\beta}}{e^{\beta} + e^{-\beta}} \right)$$

where $v_{\max}$ is the maximal valency of a vertex of G. In particular, if $\beta < \frac{1}{2} \ln \left(\frac{v_{\max}+1}{v_{\max}-1} \right)$ then κ is positive. This is not surprising, as the current research

interest in transportation distances can be traced back to [Dob70] (where the name *Vasershtein distance* is introduced), in which a criterion for convergence of spin systems is introduced. Dobrushin's criterion was a contraction property of the Markov chain in Wasserstein distance, and thus, in this context, precisely coincides with our notion of $\kappa > 0$. (See also [Per].)

Let us see how our theorems apply, for example, to the magnetization $f_0(s) := \frac{1}{|G|}\sum_{x\in G} s(x)$. With the metric we use we have $\|f_0\|_{\mathrm{Lip}} = \frac{2}{|G|}$.

Let $\gamma := 1 - v_{\max}\frac{e^{\beta}-e^{-\beta}}{e^{\beta}+e^{-\beta}}$, so that $\kappa = \frac{\gamma}{|G|}$, and assume that $\gamma > 0$. Using the gross inequalities $\sigma_\infty = 1$, $\sigma(s)^2 \leqslant 2$ and $n_s \geqslant 1$, the variance estimate of Theorem 2 reads, with $T_0 = 0$:

$$\mathrm{Var}_s\, \hat{\pi}(f_0) \leqslant \frac{8}{\gamma^2 T}$$

where s is any initial configuration. For example, taking $T \approx |G|$ (i.e. each site of G is updated a few times by the heat bath) ensures that $\mathrm{Var}_s\, \hat{\pi}(f_0)$ is of the same order of magnitude as the variance of f_0 under the invariant measure.

Theorem 4 provides a Gaussian estimate for deviations, with similar variance up to numerical constants. The transition for Gaussian to non-Gaussian regime becomes very relevant when the external magnetic field h is large, because then the number of spins opposing the magnetic field has a Poisson-like rather than Gaussian-like behavior (compare Section 3.3.3 in [Oll09]).

The bias is controlled as follows: using $E(s) \leqslant \operatorname{diam} \mathcal{X} = |G|$ in Proposition 1 one finds $|\mathbb{E}_s\hat{\pi}(f_0) - \pi(f_0)| \leqslant 2\,|G|\,(1-\gamma/\,|G|)^{T_0}/\gamma T$ so that taking $T_0 \approx |G|\log|G|$ is a good choice.

These results are not easily compared with the literature, which often focusses on getting non-explicit constants for systems of infinite size [Mar04]. However, we have seen that even in the case $\beta = 0$ our estimates improve on the spectral estimate, and our results provide very explicit bounds on the time necessary to run a heat bath simulation, at least for β not too large.

The $M/M/\infty$ queueing process. We now focus on a continuous-time example, namely the $M/M/\infty$ queueing process. This is a continuous-time Markov chain $(X_t)_{t\geqslant 0}$ on $\mathbb{N}$ with transition kernel given for small t by

$$P_t(x,y) = \begin{cases} \lambda t + o(t) & \text{if } y = x+1; \\ xt + o(t) & \text{if } y = x-1; \\ 1-(\lambda+x)t + o(t) & \text{if } y = x, \end{cases}$$

where λ is a positive parameter. The (reversible) invariant measure is the Poisson distribution π on $\mathbb{N}$ with parameter λ. Although this process is very simple in appearance, the unboundedness of the associated transition rates makes the determination of concentration inequalities technically challenging. Here we will get a convenient concentration inequality for Lipschitz functions f with respect to the

classical metric on $\mathbb{N}$, in contrast with the situation of [Jou] where Poisson-like concentration estimates are provided for Lipschitz functions with respect to an *ad hoc* metric. The techniques used here allow us to overcome this difficulty.

First, let us consider, given $d \in \mathbb{N}^\star$, $d > \lambda$, the so-called binomial Markov chain $(X_N^{(d)})_{N\in\mathbb{N}}$ on $\{0, 1, \ldots, d\}$, with transition probabilities given by

$$P_x^{(d)}(y) = \begin{cases} \frac{\lambda}{d}\left(1 - \frac{x}{d}\right) & \text{if} \quad y = x + 1; \\ \left(1 - \frac{\lambda}{d}\right)\frac{x}{d} & \text{if} \quad y = x - 1; \\ \frac{\lambda x}{d^2} + \left(1 - \frac{\lambda}{d}\right)\left(1 - \frac{x}{d}\right) & \text{if} \quad y = x. \end{cases}$$

The invariant measure is the binomial distribution $\pi^{(d)}$ on $\{0, 1, \ldots, d\}$ with parameters d and λ/d. It is not difficult to show that the Ricci curvature is $\kappa = 1/d$ and that $\sigma(x)^2 \leqslant (\lambda + x)/d$ for $x \in \{0, 1, \ldots, d\}$.

But now, take instead the continuous-time version of the above, namely the Markov process $(X_t^{(d)})_{t\geqslant 0}$ whose transition kernel is defined for any $t \geqslant 0$ as

$$P_t^{(d)}(x, y) = e^{-t}\sum_{k=0}^{+\infty} \frac{t^k}{k!}\,(P_x^{(d)})^k(y), \quad x, y \in \{0, 1, \ldots, d\}.$$

As $d \to \infty$, the invariant measure $\pi^{(d)}$ converges weakly to the Poisson measure π, which is nothing but the invariant measure of the $M/M/\infty$ queueing process. One can check (using e.g. Theorem 4.8 in [JS03]) that the process $(X_t^{(d)})_{t\geqslant 0}$ sped up by a factor d converges to the $M/M/\infty$ queueing process $(X_t)_{t\geqslant 0}$ in a suitable sense (in the *Skorokhod space* of càdlàg functions equipped with the Skorokhod topology).

To derive a concentration inequality for the empirical mean $\hat{\pi}(f) := t^{-1}\int_0^t f(X_s)\,\mathrm{d}s$, where f is 1-Lipschitz on $\mathbb{N}$ and time t is fixed, we proceed as follows. First, we will obtain a concentration estimate for the continuous-time binomial Markov chain $(X_t^{(d)})_{t\geqslant 0}$ by using Theorem 5 for the chain $(X_{\varepsilon N}^{(d)})_{N\in\mathbb{N}}$ with $\varepsilon \to 0$, and then we will approximate the $M/M/\infty$ queueing process $(X_t)_{t\geqslant 0}$ by the sped-up process $(X_{td}^{(d)})_{t\geqslant 0}$ with $d \to \infty$.

For small ε, the Markov chain $(X_{\varepsilon N}^{(d)})_{N\in\mathbb{N}}$ has Ricci curvature bounded below by ε/d, eccentricity $E(x) \leqslant x + E(0) = x + \lambda$, square diffusion constant $\sigma(x)^2$ of order $\varepsilon(\lambda + x)/d$, and $n_x \geqslant 1$, so that we may take $S(x) := \lambda + x$ in Theorems 3 and 5 above (with $T_0 = 0$ for simplicity). Let f be a 1-Lipschitz function. For a given $t > 0$ we have $\mathbb{P}_x$-almost surely the Riemann approximation:

$$\hat{\pi}^{(d)}(f) := \frac{1}{t}\int_0^t f(X_s^{(d)})\,\mathrm{d}s = \lim_{T\to+\infty} \hat{\pi}^{(d),T}(f)$$

where $\hat{\pi}^{(d),T}(f) := \frac{1}{T}\sum_{k=1}^{T} f\left(X_{kt/T}^{(d)}\right)$. So applying Theorem 5 to the Markov

chain $(X_{\varepsilon N}^{(d)})_{N\in\mathbb{N}}$ with $\varepsilon = t/T$, we get by Fatou's lemma:

$$\begin{aligned}\mathbb{P}_x\left(\left|\hat{\pi}^{(d)}(f) - \mathbb{E}_x\hat{\pi}^{(d)}(f)\right| > r\right) &\leqslant \liminf_{T\to+\infty} \mathbb{P}_x\left(\left|\hat{\pi}^{(d),T}(f) - \mathbb{E}_x\hat{\pi}^{(d),T}(f)\right| > r\right)\\ &\leqslant \begin{cases} 2\,e^{-\frac{t^2r^2}{16d(2\lambda t+(\lambda+x)d)}} & \text{if} \quad r\in(0, r_{\max}^{(d)})\\ 2\,e^{-\frac{tr}{12d}} & \text{if} \quad r\geqslant r_{\max}^{(d)}\end{cases}\end{aligned}$$

where $r_{\max}^{(d)} := (8\lambda t + 4(\lambda+x)d)/3t$. Finally, we approximate $(X_t)_{t\geqslant 0}$ by the sped-up process $(X_{td}^{(d)})_{t\geqslant 0}$ with $d\to\infty$ and apply Fatou's lemma again to obtain the following.

Corollary 6.
Let $(X_s)_{s\geqslant 0}$ be the $M/M/\infty$ queueing process with parameter λ. Let $f:\mathbb{N}\to\mathbb{R}$ be a 1-Lipschitz function. Then for any $t>0$, the empirical mean $\hat{\pi}(f) := t^{-1}\int_{s=0}^t f(X_s)\,\mathrm{d}s$ under the process starting at $x\in\mathbb{N}$ satisfies the concentration inequality

$$\mathbb{P}_x\left(|\hat{\pi}(f) - \mathbb{E}_x\hat{\pi}(f)| > r\right) \leqslant \begin{cases} 2\,e^{-\frac{tr^2}{16(2\lambda+(\lambda+x)/t)}} & \textit{if} \quad r\in(0, r_{max})\\ 2\,e^{-\frac{tr}{12}} & \textit{if} \quad r\geqslant_{max}\end{cases}$$

where $r_{max} := (8\lambda t + 4(\lambda+x))/3t$.

Let us mention that a somewhat similar, albeit much less explicit, concentration inequality has been derived in [GLWY] via transportation-information inequalities and a drift condition of Lyapunov-type.

Our results generalize to other kinds of waiting queues, such as queues with a finite number of servers and positive abandon rate.

Euler scheme for diffusions. Let $(X_t)_{t\geqslant 0}$ be the solution of the following stochastic differential equation on the Euclidean space $\mathbb{R}^d$:

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sqrt{2}\,\rho(X_t)\,\mathrm{d}W_t$$

where $(W_t)_{t\geqslant 0}$ is a standard Brownian motion in $\mathbb{R}^d$, the function $b:\mathbb{R}^d\to\mathbb{R}^d$ is measurable, as is the $d\times d$ matrix-valued function ρ. For a given matrix A, we define the Hilbert–Schmidt norm $\|A\|_{\mathrm{HS}} := \sqrt{\mathrm{tr}AA^*}$ and the operator norm $\|A\|_{\mathbb{R}^d} := \sup_{v\neq 0}\frac{\|Av\|}{\|v\|}$.

We assume that the following stability condition [BHW97, DGW04] is satisfied:

(C) the functions b and ρ are Lipschitz, and there exists $\alpha>0$ such that

$$\|\rho(x)-\rho(y)\|_{\mathrm{HS}}^2 + \langle x-y, b(x)-b(y)\rangle \leqslant -\alpha\,\|x-y\|^2, \quad x,y\in\mathbb{R}^d.$$

A typical example is the Ornstein–Uhlenbeck process, defined by $\rho = \mathrm{Id}$ and $b(x) = -x$. As we will see, this assumption implies that $\kappa > 0$.

The application of Theorems 3, 4 and 5 on this example requires careful approximation arguments (see below). The result is the following.

Corollary 7.
Let $\hat{\pi}(f) := t^{-1} \int_{s=0}^{t} f(X_s)\,\mathrm{d}s$ be the empirical mean of the 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ under the diffusion process $(X_t)_{t \geqslant 0}$ above, starting at point x. Let $S : \mathcal{X} \to \mathbb{R}$ be a C-Lipschitz function with $S(x) \geqslant \frac{2}{\alpha} \|\rho(x)\|^2_{\mathbb{R}^d}$. Set

$$V_x^2 := \frac{1}{\alpha t} \mathbb{E}_\pi S + \frac{CE(x)}{\alpha^2 t^2}$$

Then one has $\mathrm{Var}_x\, \hat{\pi}(f) \leqslant V_x^2$ and

$$\mathbb{P}_x \left(|\hat{\pi}(f) - \mathbb{E}_x \hat{\pi}(f)| > r \right) \leqslant \begin{cases} 2\, e^{-\frac{r^2}{16 V_x^2}} & \text{if} \quad r \in (0, r_{max}) \\ 2\, e^{-\frac{\alpha t r}{8C}} & \text{if} \quad r \geqslant_{max} \end{cases}$$

where $r_{max} := 2V_x^2 \alpha t / C$.

An interesting case is when ρ is constant or bounded, in which case one can take $S(x) := \sup_x \frac{2\,\|\rho(x)\|^2_{\mathbb{R}^d}}{\alpha}$ so that $C = 0$. Then $r_{\max} = \infty$ and the exponential regime disappears. For this particular case our result is comparable to [GLWY], but note however that their result requires some regularity on the distribution of the starting point of the process, in contrast with ours.

Note that the final result features the average of S under the invariant distribution. Sometimes this value is known from theoretical reasons, but in any case the assumption (C) implies very explicit bounds on the expectation of $d(x_0, x)^2$ under the invariant measure π [BHW97], which can be used to bound $\mathbb{E}_\pi S$ knowing $S(x_0)$, as well as to bound $E(x)$.

The Lipschitz growth of $\|\rho\|^2_{\mathbb{R}^d}$ allows to treat stochastic differential equations where the diffusion constant ρ grows like $\sqrt{x}$, such as naturally appear in population dynamics or superprocesses.

Proof.
Consider the underlying Euler scheme with (small) constant step δt for the stochastic differential equation above, i.e. the Markov chain $(X_N^{(\delta t)})_{N \in \mathbb{N}}$ defined by

$$X_{N+1}^{(\delta t)} = X_N^{(\delta t)} + b(X_N^{(\delta t)})\delta t + \sqrt{2\delta t}\, \rho(X_N^{(\delta t)})\, Y_N$$

where (Y_N) is any sequence of i.i.d. standard Gaussian random vectors. When $\delta t \to 0$, this process tends to a weak solution of the stochastic differential equation [BHW97].

Let us see how Theorems 3, 4 and 5 may be applied. The measure P_x is a Gaussian with expectation $x + b(x)\delta t$ and covariance matrix $2\delta t \rho\rho^*(x)$. Let

$(X_N^{(\delta t)}(x))_{N\in\mathbb{N}}$ be the chain starting at x. Under (C), we have

$$\begin{aligned}\mathbb{E}\left[\|X_1^{(\delta t)}(x) - X_1^{(\delta t)}(y)\|^2\right] &= \|x-y\|^2 + 2\delta t\langle x-y, b(x)-b(y)\rangle \\ &\quad +2\delta t\|\rho(x)-\rho(y)\|_{\mathrm{HS}}^2 + \delta t^2\|b(x)-b(y)\|^2 \\ &\leqslant \|x-y\|^2\left(1-\alpha\,\delta t + O(\delta t^2)\right)^2,\end{aligned}$$

so that we obtain $\kappa \geqslant \alpha\,\delta t + O(\delta t^2)$. Moreover, the diffusion constant $\sigma(x)$ is given by

$$\begin{aligned}\sigma(x)^2 &= \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|y-z\|^2\,P_x(\mathrm{d}y)P_x(\mathrm{d}z) \\ &= 2\delta t\|\rho(x)\|_{\mathrm{HS}}^2\end{aligned}$$

by a direct computation.

Next, using the Poincaré inequality for Gaussian measures in $\mathbb{R}^d$, with a little work one gets that the local dimension is

$$n_x = \frac{\|\rho(x)\|_{\mathrm{HS}}^2}{\|\rho(x)\|_{\mathbb{R}^d}^2}.$$

For example, if ρ is the $d\times d$ identity matrix we have $n_x = d$, whereas $n_x = 1$ if ρ is of rank 1.

So we get that $\frac{\sigma(x)^2}{\kappa n_x}$ is bounded by the function

$$S(x) := \frac{2}{\alpha}\,\|\rho(x)\|_{\mathbb{R}^d}^2 + O(\delta t).$$

However, here we have $\sigma_\infty = \infty$. This can be circumvented either by directly plugging into Lemma 10 the well-known Laplace transform estimate for Lipschitz functions of Gaussian variables, or slightly changing the approximation scheme as follows. Let us assume that $\sup_x \|\rho(x)\|_{\mathbb{R}^d} < \infty$. Now, replace the Gaussian random vectors Y_N with random vectors whose law is supported in a large ball of radius R and approximates a Gaussian (the convergence theorems of [BHW97] cover this situation as well). Then we have $\sigma_\infty = R\sqrt{2\delta t}\,\sup_x\|\rho(x)\|_{\mathbb{R}^d}$. This modifies the quantities $\sigma(x)^2$ and $\rho(x)$ by a factor at most $1+o(1)$ as $R\to\infty$.

Therefore, provided S is Lipschitz, we can apply Theorem 5 to the empirical mean $\hat{\pi}(f) := t^{-1}\int_{s=0}^t f(X_s)\,\mathrm{d}s$ by using the Euler scheme at time $T = t/\delta t$ with $\delta t\to 0$, and using Fatou's lemma as we did above for the case of the $M/M/\infty$ process. Note in particular that $\sigma_\infty\to 0$ as $\delta t\to 0$, so that σ_∞ will disappear from $r_{\max}$ in the final result.

Finally, the constraint $\sup_x\|\rho(x)\|_{\mathbb{R}^d}<\infty$ can be lifted by considering that, under our Lipschitz growth assumptions on b and $\|\rho(x)\|_{\mathbb{R}^d}^2$, with arbitrary high probability the process does not leave a compact set and so, up to an arbitrarily small error, the deviation probabilities considered depend only on the behavior of ρ and b in a compact set. □

Diffusions on positively curved manifolds. Consider a diffusion process $(X_t)_{t\geqslant 0}$ on a smooth, compact N-dimensional Riemannian manifold M, given by the stochastic differential equation

$$\mathrm{d}X_t = b\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

with infinitesimal generator

$$L := \Delta + b\cdot\nabla$$

where b is a vector field on M, Δ is the Laplace-Beltrami operator and B_t is the standard Brownian motion in the Riemannian manifold M. The Ricci curvature of this operator in the Bakry–Émery sense [BE85], applied to a tangent vector v, is $\mathrm{Ric}(v,v) - v\cdot\nabla_v b$ where Ric is the usual Ricci tensor. Assume that this quantity is at least K for any unit tangent vector v.

Consider as above the Euler approximation scheme at time δt for this stochastic differential equation: starting at a point x, follow the flow of b for a time δt, to obtain a point x'; now take a random tangent vector w at x' whose law is a Gaussian in the tangent plane at x' with covariance matrix equal to the metric, and follow the geodesic generated by w for a time $\sqrt{2\delta t}$. Define P_x to be the law of the point so obtained. When $\delta t \to 0$, this Markov chain approximates the process $(X_t)_{t\geqslant 0}$ (see e.g. Section I.4 in [Bis81]). Just as above, actually the Gaussian law has to be truncated to a large ball so that $\sigma_\infty < \infty$.

For this Euler approximation, we have $\kappa \geqslant K\delta t + O(\delta t^{3/2})$ where K is a lower bound for Ricci–Bakry–Émery curvature [Oll09]. We have $\sigma(x)^2 = 2N\delta t + O(\delta t^{3/2})$ and $n_x = N + O(\sqrt{\delta t})$. The details are omitted, as they are very similar to the case of $\mathbb{R}^d$ above, except that in a neighborhood of size $\sqrt{\delta t}$ of a given point, distances are distorted by a factor $1 \pm O(\sqrt{\delta t})$ w.r.t. the Euclidean case. We restrict the statement to compact manifolds so that the constants hidden in the $O()$ notation are uniform in x.

So applying Theorem 4 to the Euler scheme at time $T = t/\delta t$ we get:

Corollary 8.
Let $(X_t)_{t\geqslant 0}$ be a process as above on a smooth, compact N-dimensional Riemannian manifold $\mathcal{X}$, with Bakry–Émery curvature at least $K > 0$. Let $\hat{\pi}(f) := t^{-1}\int_{s=0}^{t} f(X_s)\,\mathrm{d}s$ be the empirical mean of the 1-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$ under the diffusion process (X_t) starting at some point $x \in \mathcal{X}$. Then

$$\mathbb{P}_x(|\hat{\pi}(f) - \mathbb{E}_x\hat{\pi}(f)| > r) \leqslant 2e^{-\frac{K^2\,t\,r^2}{32}}.$$

Once more, a related estimate appears in [GLWY], except that their result features an additional factor $\|\mathrm{d}\beta/\mathrm{d}\pi\|_2$ where β is the law of the initial point of the Markov chain and π is the invariant distribution, thus preventing it from being applied with β a Dirac measure at x.

Non-linear state space models. Given a Polish state space $(\mathcal{X}, d)$, we consider the Markov chain $(X_N)_{N\in\mathbb{N}}$ solution of the following equation

$$X_{N+1} = F(X_N, W_{N+1}), \quad X_0 \in \mathcal{X},$$

which models a noisy dynamical system. Here $(W_N)_{N\in\mathbb{N}}$ is a sequence of i.i.d. random variables with values in some parameter space, with common distribution μ. We assume that there exists some $r < 1$ such that

$$\mathbb{E}\, d\left(F(x, W_1), F(y, W_1)\right) \leqslant r d(x, y), \quad x, y \in \mathcal{X}, \tag{10}$$

and that moreover the following function is L^2-Lipschitz on $\mathcal{X}$:

$$x \mapsto \mathbb{E}\left[d(F(x, W_1), F(x, W_2))^2\right].$$

Note that the assumption (10) already appears in [DGW04] (Condition (3.3)) to study the propagation of Gaussian concentration to path-dependent functionals of $(X_N)_{N\in\mathbb{N}}$.

Since the transition probability P_x is the image measure of μ by the function $F(x, \cdot)$, it is straightforward that the Ricci curvature κ is at least $1 - r$, which is positive. Hence we may apply Theorem 3 with the $L^2/(2(1-r))$-Lipschitz function

$$S(x) := \frac{1}{2(1-r)}\, \mathbb{E}\left[d(F(x, W_1), F(x, W_2))^2\right],$$

to obtain the variance inequality:

$$\sup_{\|f\|_{\mathrm{Lip}}\leqslant 1} \mathrm{Var}_x\, \hat{\pi}(f) \leqslant \begin{cases} \frac{1}{(1-r)T}\left\{\frac{L^2}{2(1-r)^2 T} E(x) + \mathbb{E}_\pi S\right\} & \text{if} \quad T_0 = 0; \\ \frac{1}{\kappa T}\left\{\left(1 + \frac{1}{(1-r)T}\right)\mathbb{E}_\pi S + \frac{L^2 r^{T_0}}{(1-r)^2 T} E(x)\right\} & \text{otherwise.} \end{cases}$$

Note that to obtain a qualitative concentration estimate via Theorem 5, we need the additional assumption $\sigma_\infty < \infty$, which depends on the properties of μ and of the function $F(x, \cdot)$ and states that at each step the noise has a bounded influence.

3 Proofs

3.1 Proof of Proposition 1

Let f be a 1-Lipschitz function. Let us recall from [Oll09] that for $k \in \mathbb{N}$, the function $P^k f$ is $(1-\kappa)^k$-Lipschitz. Then we have by the invariance of π:

$$\begin{aligned} |\mathbb{E}_x \hat{\pi}(f) - \pi(f)| &= \frac{1}{T}\left|\sum_{k=T_0+1}^{T_0+T} \int_{\mathcal{X}} \left(P^k f(x) - P^k f(y)\right) \pi(\mathrm{d}y)\right| \\ &\leqslant \frac{1}{T}\sum_{k=T_0+1}^{T_0+T} (1-\kappa)^k \int_{\mathcal{X}} d(x, y)\, \pi(\mathrm{d}y) \\ &\leqslant \frac{(1-\kappa)^{T_0+1}}{\kappa T} \int_{\mathcal{X}} d(x, y)\, \pi(\mathrm{d}y), \end{aligned}$$

so that we obtain the result.

3.2 Proof of Theorems 2 and 3

Let us start with a variance-type result under the measure after N steps. The proof relies on a simple induction argument and is left to the reader.

Lemma 9.
For any $N \in \mathbb{N}^\star$ and any Lipschitz function f on $\mathcal{X}$, we have:

$$P^N(f^2) - (P^N f)^2 \leqslant \|f\|_{\mathrm{Lip}}^2 \sum_{k=0}^{N-1} (1-\kappa)^{2(N-1-k)}\, P^k\left(\frac{\sigma^2}{n}\right). \tag{11}$$

In particular if the rate $x \mapsto \sigma(x)^2/n_x$ is bounded then letting N tend to infinity above entails a variance estimate under the invariant measure π:

$$\mathrm{Var}_\pi\, f \leqslant \|f\|_{\mathrm{Lip}}^2 \sup_{x\in\mathcal{X}} \frac{\sigma(x)^2}{n_x \kappa}. \tag{12}$$

Now we are able to prove the variance bounds of Theorems 2 and 3. Given a 1-Lipschitz function f, consider the functional

$$f_{x_1,\dots,x_{T-1}}(x_T) := \frac{1}{T}\sum_{k=1}^{T} f(x_k),$$

the others coordinates $x_1, \dots, x_{T-1}$ being fixed. The function $f_{x_1,\dots,x_{T-1}}$ is $1/T$-Lipschitz, hence $1/\kappa T$-Lipschitz since $\kappa \leqslant 1$. Moreover, for each $k \in \{T-1, T-2, \dots, 2\}$, define by a downward induction the conditional expectation of $\hat{\pi}(f)$ knowing $X_1 = x_1, \dots, X_k = x_k$:

$$f_{x_1,\dots,x_{k-1}}(x_k) := \int_{\mathcal{X}} f_{x_1,\dots,x_k}(x_{k+1})\, P_{x_k}(\mathrm{d}x_{k+1})$$

and

$$f_\varnothing(x_1) := \int_{\mathcal{X}} f_{x_1}(x_2)\, P_{x_1}(\mathrm{d}x_2).$$

By Lemma 3.2 (step 1) in [Jou], we know that $f_{x_1,\dots,x_{k-1}}$ is Lipschitz with constant s_k, where

$$s_k := \frac{1}{T}\sum_{j=0}^{T-k} (1-\kappa)^j \leqslant \frac{1}{\kappa T}.$$

Hence we can use the variance bound (11) with $N = 1$ for the function $f_{x_1,\dots,x_{k-1}}$,

successively for $k = T, T-1, \ldots, 2$, to obtain:

$$\begin{aligned}
\mathbb{E}_x[\hat{\pi}(f)^2] &= \int_{\mathcal{X}^T} f_{x_1,\ldots,x_{T-1}}(x_T)^2 \, P_{x_{T-1}}(\mathrm{d}x_T) \cdots P_{x_1}(\mathrm{d}x_2) \, P_x^{T_0+1}(\mathrm{d}x_1) \\
&\leqslant \int_{\mathcal{X}^{T-1}} f_{x_1,\ldots,x_{T-2}}(x_{T-1})^2 \, P_{x_{T-2}}(\mathrm{d}x_{T-1}) \cdots P_{x_1}(\mathrm{d}x_2) \, P_x^{T_0+1}(\mathrm{d}x_1) \\
&\quad + \frac{1}{\kappa^2 T^2} \, P^{T_0+T-1}\left(\frac{\sigma^2}{n}\right)(x) \\
&\leqslant \int_{\mathcal{X}^{T-2}} f_{x_1,\ldots,x_{T-3}}(x_{T-2})^2 \, P_{x_{T-3}}(\mathrm{d}x_{T-2}) \cdots P_{x_1}(\mathrm{d}x_2) \, P_x^{T_0+1}(\mathrm{d}x_1) \\
&\quad + \frac{1}{\kappa^2 T^2} \left(P^{T_0+T-2}\left(\frac{\sigma^2}{n}\right)(x) + P^{T_0+T-1}\left(\frac{\sigma^2}{n}\right)(x) \right) \\
&\leqslant \cdots \\
&\leqslant \int_{\mathcal{X}} f_{\varnothing}(x_1)^2 \, P_x^{T_0+1}(\mathrm{d}x_1) + \frac{1}{\kappa^2 T^2} \sum_{k=T_0+1}^{T_0+T-1} P^k\left(\frac{\sigma^2}{n}\right)(x) \\
&\leqslant (\mathbb{E}_x \hat{\pi}(f))^2 + \frac{1}{\kappa^2 T^2} \sum_{k=0}^{T_0} (1-\kappa)^{2(T_0-k)} P^k\left(\frac{\sigma^2}{n}\right)(x) \\
&\quad + \frac{1}{\kappa^2 T^2} \sum_{k=T_0+1}^{T_0+T-1} P^k\left(\frac{\sigma^2}{n}\right)(x),
\end{aligned}$$

where in the last step we applied the variance inequality (11) to the Lipschitz function $f_{\varnothing}$, with $N = T_0 + 1$. Therefore we get

$$\operatorname{Var}_x \hat{\pi}(f) \leqslant \frac{1}{\kappa^2 T^2} \left(\sum_{k=0}^{T_0} (1-\kappa)^{2(T_0-k)} P^k\left(\frac{\sigma^2}{n}\right)(x) + \sum_{k=T_0+1}^{T_0+T-1} P^k\left(\frac{\sigma^2}{n}\right)(x) \right).$$

Theorem 2 is a straightforward consequence of the latter inequality. To establish

Theorem 3, for instance (7) in the case $T_0 \neq 0$, we rewrite the above as:

$$\begin{aligned}
\mathrm{Var}_x\,\hat{\pi}(f) &\leqslant \frac{1}{\kappa T^2}\left\{\sum_{k=0}^{T_0}(1-\kappa)^{2(T_0-k)}P^kS(x)+\sum_{k=T_0+1}^{T_0+T-1}P^kS(x)\right\}\\
&\leqslant \frac{1}{\kappa T^2}\left\{\sum_{k=0}^{T_0}(1-\kappa)^{2(T_0-k)}\Big(C\,W_1(P_x^k,\pi)+\mathbb{E}_\pi S\Big)\right.\\
&\quad\left.+\sum_{k=T_0+1}^{T_0+T-1}\Big(C\,W_1(P_x^k,\pi)+\mathbb{E}_\pi S\Big)\right\}\\
&\leqslant \frac{1}{\kappa T^2}\left\{\left(1+\frac{1}{\kappa T}\right)T\,\mathbb{E}_\pi S+\sum_{k=0}^{T_0}C(1-\kappa)^{2T_0-k}E(x)\right.\\
&\quad\left.+\sum_{k=T_0+1}^{T_0+T-1}C(1-\kappa)^kE(x)\right\}\\
&\leqslant \frac{1}{\kappa T^2}\left\{\left(1+\frac{1}{\kappa T}\right)T\,\mathbb{E}_\pi S+\frac{2C(1-\kappa)^{T_0}}{\kappa}E(x)\right\}.
\end{aligned}$$

Finally, the proof in the case $T_0 = 0$ is very similar and is omitted.

3.3 Proof of Theorems 4 and 5

The proof of the concentration theorems 4 and 5 follows the same lines as that for variance above, except that Laplace transform estimates $\mathbb{E}e^{\lambda f-\lambda\mathbb{E}f}$ now play the role of the variance $\mathbb{E}[f^2]-(\mathbb{E}f)^2$.

Assume that there exists a Lipschitz function $S : \mathcal{X} \to \mathbb{R}$ with $\|S\|_{\mathrm{Lip}} \leqslant C$ such that

$$\frac{\sigma(x)^2}{n_x\kappa} \leqslant S(x), \quad x \in \mathcal{X}.$$

Let us give first a result on the Laplace transform of Lipschitz funtions under the measure at time N.

Lemma 10.
Let $\lambda \in \left(0, \frac{\kappa T}{\max\{4C, 6\sigma_\infty\}}\right)$. Then for any $N \in \mathbb{N}^\star$ and any $\frac{2}{\kappa T}$-Lipschitz function f on $\mathcal{X}$, we have:

$$P^N(e^{\lambda f}) \leqslant \exp\left\{\lambda P^N f + \frac{4\lambda^2}{\kappa T^2}\sum_{k=0}^{N-1}P^kS\right\}. \tag{13}$$

In the case $C = 0$, the same formula holds for any $\lambda \in (0, \frac{\kappa T}{6\sigma_\infty})$.

Proof.
Let f be $\frac{2}{\kappa T}$-Lipschitz. By Lemma 38 in [Oll09], we know that if g is an α-Lipschitz

function with $\alpha \leqslant 1$ and if $\lambda \in (0, \frac{1}{3\sigma_\infty})$ then we have the estimate

$$P(\mathrm{e}^{\lambda g}) \leqslant \exp\left\{\lambda P g + \kappa \lambda^2 \alpha^2 S\right\},$$

and by rescaling, the same holds for the function f with $\alpha = \frac{2}{\kappa T}$ whenever $\lambda \in (0, \frac{\kappa T}{6\sigma_\infty})$. Moreover the function $P^N f + \frac{4\lambda}{\kappa T^2} \sum_{k=0}^{N-1} P^k S$ is also $\frac{2}{\kappa T}$-Lipschitz for any $N \in \mathbb{N}^\star$, since $\lambda \in (0, \frac{\kappa T}{4C})$. Hence the result follows by a simple induction argument. □

Now let us prove Theorem 5, using again the notation of Section 3.2 above. Theorem 4 easily follows from Theorem 5 by taking $S := \sup_{x \in \mathcal{X}} \frac{\sigma(x)^2}{n_x \kappa}$ and letting $C \to 0$ in the formula (9).

Let f be a 1-Lipschitz function on $\mathcal{X}$ and let $\lambda \in \left(0, \frac{\kappa T}{\max\{4C, 6\sigma_\infty\}}\right)$. Using the Laplace transform estimate (13) with $N = 1$ for the $\frac{2}{\kappa T}$-Lipschitz functions

$$x_k \mapsto f_{x_1,\dots,x_{k-1}}(x_k) + \frac{4\lambda}{\kappa T^2} \sum_{l=0}^{T-k-1} P^l S(x_k),$$

successively for $k = T-1, T-2, \dots, 2$, we have:

$$\begin{aligned}
&\mathbb{E}_x e^{\lambda \hat{\pi}(f)} \\
&\quad = \int_{\mathcal{X}^T} e^{\lambda f_{x_1,\dots,x_{T-1}}(x_T)} P_{x_{T-1}}(\mathrm{d}x_T) \cdots P_{x_1}(\mathrm{d}x_2)\, P_x^{T_0+1}(\mathrm{d}x_1) \\
&\quad \leqslant \int_{\mathcal{X}^{T-1}} e^{\lambda f_{x_1,\dots,x_{T-2}}(x_{T-1}) + \frac{4\lambda^2}{\kappa T^2} S(x_{T-1})} P_{x_{T-2}}(\mathrm{d}x_{T-1}) \cdots P_{x_1}(\mathrm{d}x_2)\, P_x^{T_0+1}(\mathrm{d}x_1) \\
&\quad \leqslant \int_{\mathcal{X}^{T-2}} e^{\lambda f_{x_1,\dots,x_{T-3}}(x_{T-2}) + \frac{4\lambda^2}{\kappa T^2} \sum_{l=0}^{1} P^l S(x_{T-2})} P_{x_{T-2}}(\mathrm{d}x_{T-1}) \cdots P_{x_1}(\mathrm{d}x_2)\, P_x^{T_0+1}(\mathrm{d}x_1) \\
&\quad \leqslant \cdots \\
&\quad \leqslant \int_{\mathcal{X}} e^{\lambda f_\varnothing(x_1) + \frac{4\lambda^2}{\kappa T^2} \sum_{l=0}^{T-2} P^l S(x_1)} P_x^{T_0+1}(\mathrm{d}x_1) \\
&\quad \leqslant e^{\lambda \mathbb{E}_x \hat{\pi}(f) + \frac{4\lambda^2}{\kappa T^2} \left(\sum_{l=0}^{T-2} P^{T_0+1+l} S(x) + \sum_{l=0}^{T_0} P^l S(x) \right)},
\end{aligned}$$

where in the last line we applied the Laplace transform estimate (13) to the $\frac{2}{\kappa T}$-Lipschitz function

$$x_1 \mapsto f_\varnothing(x_1) + \frac{4\lambda}{\kappa T^2} \sum_{l=0}^{T-2} P^l S(x_1),$$

with $N = T_0 + 1$. Therefore we get:

$$\mathbb{E}_x e^{\lambda(\hat{\pi}(f) - \mathbb{E}_x \hat{\pi}(f))} \leqslant e^{4\lambda^2 V_x^2}.$$

Finally, using Chebychev's inequality and optimizing in $\lambda \in \left(0, \frac{\kappa T}{\max\{4C, 6\sigma_\infty\}}\right)$ entails the result. This ends the proof.

References

[BD97] R. Bubley, M. E. Dyer, *Path coupling: a technique for proving rapid mixing in Markov chains*, FOCS 1997, 223–231.

[BE85] D. Bakry, M. Émery, *Diffusions hypercontractives*, Séminaire de probabilités, XIX, 1983/84. Lecture Notes in Math. **1123**, Springer, Berlin (1985), 177–206.

[Bis81] J.-M. Bismut, *Mécanique aléatoire*, Lecture Notes in Mathematics **866**, Springer-Verlag, Berlin-New York, 1981.

[BHW97] G. K. Basak and I. Hu and C. Z. Wei, Weak convergence of recursions, *Stochastic Process. Appl.*, 68(1):65-82, 1997.

[CG08] P. Cattiaux, A. Guillin, *Deviation bounds for additive functionals of Markov processes*, ESAIM Probab. Stat. **12** (2008), 12–29 (electronic).

[CW94] M.-F. Chen, F.-Y. Wang, *Application of coupling method to the first eigenvalue on manifold*, Sci. China Ser. A **37** (1994), n° 1, 1–14.

[DGW04] H. Djellout, A. Guillin, L. Wu, *Transportation cost-information inequalities and applications to random dynamical systems and diffusions*, Ann. Prob. **32** (2004), n° 3B, 2702–2732.

[Dob70] R. L. Dobrušin, *Definition of a system of random variables by means of conditional distributions* (Russian) Teor. Verojatnost. i Primenen. **15** (1970), 469–497.

[Dob96] R. Dobrushin, *Perturbation methods of the theory of Gibbsian fields*, in R. Dobrushin, P. Groeneboom, M. Ledoux, *Lectures on probability theory and statistics*, Lectures from the 24th Saint-Flour Summer School held July 7–23, 1994, edited by P. Bernard, Lecture Notes in Mathematics **1648**, Springer, Berlin (1996), 1–66.

[DS85] R. L. Dobrushin, S. B. Shlosman, *Constructive criterion for the uniqueness of Gibbs field*, in J. Fritz, A. Jaffe and D. Szász (eds), *Statistical physics and dynamical systems*, papers from the second colloquium and workshop on random fields: rigorous results in statistical mechanics, held in Köszeg, August 26–September 1, 1984, Progress in Physics **10**, Birkhäuser, Boston (1985), 347–370.

[DS96] P. Diaconis, L. Saloff-Coste, *Logarithmic Sobolev inequalities for finite Markov chains*, Ann. Appl. Probab. **6** (1996), n° 3, 695–750.

[GLWY] A. Guillin, C. Léonard, L. Wu and N. Yao, Transportation-information inequalities for Markov processes, To appear in *Probab. Theory Relat. Fields.*

[JS03] J. Jacod and A. N. Shiryaev, *Limit theorems for stochastic processes*, Grundlehren der Mathematischen Wissenschaften, 288, Springer-Verlag, Berlin, 2003.

[Jou07] A. Joulin, *Poisson-type deviation inequalities for curved continuous time Markov chains*, Bernoulli **13** (2007), n° 3, 782–798.

[Jou] A. Joulin, *A new Poisson-type deviation inequality for Markov jump processes with positive Wasserstein curvature*, To appear in *Bernoulli*.

[Lez98] P. Lezaud, *Chernoff-type bound for finite Markov chains*, Ann. Appl. Probab. **8** (1998), n° 3, 849–867.

[Mar04] F. Martinelli, *Relaxation times of Markov chains in statistical mechanics and combinatorial structures*, in H. Kesten (ed), *Probability on discrete structures*, Encyclopaedia of Mathematical Sciences **110**, Springer, Berlin (2004), 175–262.

[MT93] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Communications and Control Engineering Series, Springer-Verlag London, 1993.

[Oli] R. Imbuzeiro Oliveira, *On the convergence to equilibrium of Kac's random walk on matrices*, preprint, `arXiv:0705.2253`

[Oll07] Y. Ollivier, *Ricci curvature of metric spaces*, C. R. Math. Acad. Sci. Paris **345** (2007), n° 11, 643–646.

[Oll09] Y. Ollivier, *Ricci curvature of Markov chains on metric spaces*, J. Funct. Anal. **256** (2009), n° 3, 810–864.

[Per] Y. Peres, *Mixing for Markov chains and spin systems*, lecture notes (2005).

[RR04] G. O. Roberts and J. S. Rosenthal, General state space Markov chains and MCMC algorithms, *Probab. Surv.*, 1:20-71, 2004.

[Vil03] C. Villani, *Topics in optimal transportation*, Graduate Studies in Mathematics **58**, American Mathematical Society, Providence (2003).

[Wu00] L. Wu, *A deviation inequality for non-reversible Markov processes*, Ann. Inst. H. Poincaré Probab. Statist. **36** (2000), n° 4, 435–445.