Electron. Commun. Probab. **20** (2015), no. 64, 1–12.
DOI: 10.1214/ECP.v20-3966
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

Subgaussian concentration inequalities for geometrically ergodic Markov chains

Jérôme Dedecker[*] Sébastien Gouëzel[†]

Abstract

We prove that an irreducible aperiodic Markov chain is geometrically ergodic if and only if any separately bounded functional of the stationary chain satisfies an appropriate subgaussian deviation inequality from its mean.

Keywords: Concentration inequalities ; Markov chains ; Geometric ergodicity ; Coupling.
AMS MSC 2010: 60J10 ; 60E15.
Submitted to ECP on December 4, 2014, final version accepted on September 9, 2015.

Let $K(x_0, \dots, x_{n-1})$ be a function of n variables, which is separately bounded in the following sense: there exist constants L_i such that for all $x_0, \dots, x_{n-1}, x_i'$,

$$|K(x_0, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{n-1}) - K(x_0, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_{n-1})| \leqslant L_i. \tag{0.1}$$

It is well known that, if the random variables $X_0, X_1, \dots$ are i.i.d., then $K(X_0, \dots, X_{n-1})$ satisfies a subgaussian concentration inequality around its average, of the form

$$\mathbb{P}(|K(X_0, \dots, X_{n-1}) - \mathbb{E}K(X_0, \dots, X_{n-1})| > t) \leqslant 2e^{-2t^2/\sum L_i^2}, \tag{0.2}$$

see for instance [9].

Such concentration inequalities have also attracted a lot of interest for dependent random variables, due to the wealth of possible applications. For instance, Markov chains with good mixing properties have been considered, as well as weakly dependent sequences.

A particular instance of function K is a sum $\sum f(x_i)$ (also referred to as an additive functional). In this case, one can hope for better estimates than (0.2), involving for instance the asymptotic variance instead of only L_i (Bernstein-like inequalities). However, for the case of a general functional K, estimates of the form (0.2) are rather natural.

Under very strong assumptions ensuring that the dependence is uniformly small (say, uniformly ergodic Markov chains, or Φ-mixing dependent sequences), subgaussian concentration inequalities are well known (see [15] for the extension of (0.2) and [16] for other concentration inequalities). For additive functionals, Lezaud [7, p 861] proved a Prohorov-type inequality under a spectral gap condition in $\mathbb{L}^2$, from which a subgaussian bound follows. However, there are very few such results under weaker assumptions (say, geometrically ergodic Markov chains, or α-mixing dependent sequences), where

[*] Laboratoire MAP5 CNRS UMR 8145, Université Paris Descartes, Sorbonne Paris Cité, 75006 Paris, France.
E-mail: jerome.dedecker@parisdescartes.fr

[†] IRMAR, CNRS UMR 6625, Université de Rennes 1, 35042 Rennes, France.
E-mail: sebastien.gouezel@univ-rennes1.fr

other type of exponential bounds are more usual (let us cite [10] for α-mixing sequences and [2] for geometrically ergodic Markov chains; see also the references in those two papers for a quite complete picture of the literature). As an exception, let us mention the result of Adamczak, who proves in [1] subgaussian concentration inequalities for geometrically ergodic Markov chains under the additional assumptions that the chain is strongly aperiodic and that the functional K is invariant under permutation of its variables.

Our goal in this note is to prove subgaussian concentration inequalities for aperiodic geometrically ergodic Markov chains, extending the above result of [1]. Such a setting has a wide range of applications, in particular to MCMC (see for instance Section 3.2 in [2]). Our proof is mainly a reformulation in probabilistic terms of the proof given in [4] for dynamical systems. It is based on a classical coupling estimate (Lemma 0.6 below), but used in an unusual way along an unusual filtration (the relationship between coupling and concentration has already been explored in [5]). Similar results can also be proved for Markov chains that mix more slowly (for instance, if the return times to a small set have a polynomial tail, then polynomial concentration inequalities hold). The interested reader is referred to the articles [4] and [6] where such results are proved for dynamical systems: the proofs given there can be readily adapted to Markov chains using the techniques we describe in the current paper (the only difficulty is to prove an appropriate coupling lemma extending Lemma 0.6). Since the main case of interest for applications is geometrically ergodic Markov chains, and since the proof is more transparent in this case, we only give details for this situation.

Our results are valid for Markov chains on a general state space $\mathcal{S}$, but they are already new and interesting for countable state Markov chains. The reader who is unfamiliar with general state space Markov chains is therefore invited to pretend that $\mathcal{S}$ is countable. We chose to present our results for general state space firstly because of the wealth of applications, and secondly because of a peculiarity of general state space that does not exist for countable state space: there is a distinction between strongly aperiodic and aperiodic chains, and several mixing results only apply in the strongly aperiodic case (i.e., $m = 1$ in Definition 0.1 below) while our argument always applies.

From this point on, we consider an irreducible aperiodic positive Markov chain $(X_n)_{n\geqslant 0}$ on a general state space $\mathcal{S}$, which we assume as usual to be countably generated. We refer to the books [13] or [11] for the classical background on Markov chains on general state spaces. Let us nevertheless recall the meaning of some of the above terms, since it may vary slightly between sources.

First, we are given a measurable transition kernel P of the chain, that is, for any measurable set A in $\mathcal{S}$,

$$P(x, A) = \mathbb{E}\left(\mathbb{1}_{X_1 \in A} \mid X_0 = x\right).$$

Starting from any point x_0, we obtain a chain $X_0 = x_0, X_1, X_2, \dots$, where X_i is distributed according to the measure $P(X_{i-1}, \cdot)$. This chain is irreducible, aperiodic and positive if there exists a (necessarily unique) stationary probability measure π such that, for all x, all sets A with $\pi(A) > 0$ and all large enough n (depending on x and A), one has $P^n(x, A) > 0$ (where P^n denotes the kernel of the Markov chain at time n). Other definitions of irreducibility only require this property to hold for almost every x (in this case, one can restrict to an absorbing set of full π-measure to obtain it for all x there), we follow the definition of [11].

We will be interested in a specific class of such Markov chains, called geometrically ergodic. There are many equivalent definitions of this class, in terms of the properties of the return time to a nice set, or of mixing properties. Essentially, geometrically ergodic

Markov chains are those Markov chains that mix exponentially fast, see [11, Chapters 15 and 16] for several equivalent characterizations. For instance, they can be defined as follows [11, Theorem 15.0.1(ii)].

Definition 0.1. *An irreducible aperiodic positive Markov chain is* geometrically ergodic *if the tails of the return time to some small set are exponential. More precisely, there exist a set C, an integer $m > 0$, a probability measure ν, and $\delta \in (0,1)$, $\kappa > 1$ such that*

- *For all $x \in C$, one has*
$$P^m(x,\cdot) \geqslant \delta\nu. \tag{0.3}$$
- *The return time τ_C to C satisfies*
$$\sup_{x\in C} \mathbb{E}_x(\kappa^{\tau_C}) < \infty. \tag{0.4}$$

A set C satisfying (0.3) is called a *small* set (there is a related notion of *petite* set, these notions coincide in irreducible aperiodic Markov chains, see [11, Theorem 5.5.7]).

In the case of a countable state space, this property is equivalent to the fact that the return time to some (or equivalently any) point has an exponential moment.

From Theorem 15.0.1 of [11], it follows that if a chain is geometrically ergodic in the sense of Definition 0.1, then

$$\|P^n(x,\cdot) - \pi\| \leqslant V(x)\rho^n\,, \tag{0.5}$$

where $\|\cdot\|$ is the total variation norm, $\rho \in (0,1)$ and V is a positive function such that the set $S_V = \{x : V(x) < \infty\}$ is absorbing and of full measure. This property (0.5) is in fact another classical definition for geometric ergodicity: from Theorem 15.4.2 in [11] (or Theorem 6.14 in [13]) it follows that if a chain is irreducible, aperiodic, positively recurrent (so that there exists an unique stationary distribution π) and satisfies (0.5), then there exists a small set C for which (0.4) holds.

We prove the following theorem.

Theorem 0.2. *Let (X_n) be an irreducible aperiodic Markov chain which is geometrically ergodic on a space $\mathcal{S}$. Let π be its stationary distribution. Let C be a small set as in Definition 0.1. There exists a constant M_0 (depending on C) with the following property. Let $n \in \mathbb{N}$. Let $K(x_0,\dots,x_{n-1})$ be a function of n variables on $\mathcal{S}^n$, which is separately bounded with constants L_i, as in* (0.1)*. Then, for all $t > 0$,*

$$\mathbb{P}_\pi(|K(X_0,\dots,X_{n-1}) - \mathbb{E}_\pi K(X_0,\dots,X_{n-1})| > t) \leqslant 2e^{-M_0^{-1}t^2/\sum L_i^2}\,, \tag{0.6}$$

and for all x in the small set C,

$$\mathbb{P}_x(|K(X_0,\dots,X_{n-1}) - \mathbb{E}_x K(X_0,\dots,X_{n-1})| > t) \leqslant 2e^{-M_0^{-1}t^2/\sum L_i^2}\,. \tag{0.7}$$

As will be clear from the proof, the constant M_0 can be written explicitly in terms of simple numerical properties of the Markov chain, more precisely of its coupling time and of the return time to the small set C. We shall in fact prove (0.7), and show how it implies (0.6) (see the first step of the proof of Theorem 0.2).

Note that there is no strong aperiodicity assumption in our theorems (i.e., we are not requiring $m = 1$), contrary to several mixing results for Markov chains. The reason for this is that we will use the splitting method of Nummelin (see Definition 0.5 below) only to control coupling times, but we will not need the independence of the blocks between two successive entrance times to the atom of the split chain as in [1]. Following the classical strategy of McDiarmid, we will rather decompose K as a sum of martingale increments, and estimate each of them. However, if we try to use the natural filtration given by the time, we have no control on what happens away from C. The main unusual idea in our argument is to use another filtration indexed by the next return to C, the rest is mainly routine.

The following remarks show that the above theorem is sharp: it is not possible to weaken the boundedness assumption (0.1), nor the assumption of geometric ergodicity.

Remark 0.3. It is often desirable to have estimates for functions which are unbounded. A typical example in geometrically ergodic Markov chains is the following. Consider an appropriate drift function, i.e., a function $V \geqslant 1$ which is bounded on a small set C and satisfies $PV(x) \leqslant \rho V(x) + A\mathbb{1}_C(x)$ for some numbers $\rho < 1$ and $A \geqslant 0$ (where P is the Markov operator of the chain). One thinks of V as being “large close to infinity”. A natural candidate for stronger concentration inequalities would be functions K satisfying

$$|K(x_0, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{n-1}) - K(x_0, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_{n-1})| \\ \leqslant L_i f(V(x_i) \vee V(x_i')), \quad (0.8)$$

for some positive function f going to infinity at infinity, for instance $f(t) = \log(1+t)$. Unfortunately, subgaussian concentration inequalities do *not* hold for such functionals of geometrically ergodic Markov chains: there exists a geometrically ergodic Markov chain such that, for any M_0, for any function f going to infinity, there exist n and a functional K satisfying (0.8) for which the inequality (0.6) is violated. Even more, concentration inequalities fail for additive functionals.

Consider for instance the chain on $\{1, 2, \dots\}$ given by $\mathbb{P}(1 \to s) = 2^{-s}$ for $s \geqslant 1$ and $\mathbb{P}(s \to s-1) = 1$ for $s > 1$. The function $V(s) = 2^{s/2}$ satisfies the drift condition, for the small set $C = \{1\}$, since $PV(s) = 2^{-1/2}V(s)$ for $s > 1$ and $PV(1) = 2^{-1/2}/(1-2^{-1/2}) < \infty$. The stationary measure π is given by $\pi(s) = 2^{-s}$. In particular, V is integrable.

Assume by contradiction that a concentration inequality (0.6) holds for all functionals satisfying the bound (0.8), for some function f going to infinity and some $M_0 > 0$. Let $\tilde{f}$ be a nondecreasing function with $\tilde{f}(x) \leqslant \min(f(x), x)$, tending to infinity at infinity. Define a function $g(s) = \tilde{f}(V(s))$, except for $s = 1$ where $g(1)$ is chosen so that $\int g \,\mathrm{d}\pi = 0$. Let $K(x_0, \dots, x_{n-1}) = \sum g(x_i)$, it satisfies (0.8) with $L_i = L$ constant and $\mathbb{E}_\pi K = 0$.

For any $N > 0$ and $n > 0$, the Markov chain has a probability 2^{-n-N} to start from $X_0 = n + N$, and then the next n iterates are $n + N - i \geqslant N$. In this case, $g(X_0) + \dots + g(X_{n-1}) \geqslant ng(N)$. Applying (0.6), we get

$$2^{-n-N} = \pi(n+N) \leqslant \mathbb{P}_\pi(|K - \mathbb{E}_\pi K| \geqslant ng(N)) \leqslant 2e^{-M_0^{-1}(ng(N))^2/(nL^2)} = 2e^{-M_0^{-1}L^{-2}g(N)^2 n}.$$

Letting n tend to infinity, we deduce that $M_0^{-1}L^{-2}g(N)^2 \leqslant \log 2$. This is a contradiction if N is large enough, since g tends to infinity.

For instance, if one takes $f(t) = \sqrt{\ln(t \vee e)}$, then g satisfies the subgaussian condition $\mathbb{E}_\pi(\exp(g(X_0)^2)) < \infty$, but nevertheless the subgaussian inequality for the additive functional $g(X_0) + \dots + g(X_{n-1})$ fails.

Remark 0.4. One may wonder if the subgaussian concentration inequality (0.6) can be proved in larger classes of Markov chains. This is not the case: (0.6) *characterizes* geometrically ergodic Markov chains, as we now explain.

Consider an irreducible aperiodic Markov chain such that (0.6) holds for any separately bounded functional. We want to prove that it is geometrically ergodic. By [11, Theorem 5.2.2], there exists a small set, i.e., a set C satisfying (0.3), for some $m \geqslant 1$. If the original chain satisfies subgaussian concentration inequalities, then the chain at times which are multiples of m (called its m-skeleton) also does. Moreover, an irreducible aperiodic Markov chain is geometrically ergodic if and only if its m-skeleton is, by [11, Theorem 15.3.6]. It follows that is suffices to prove the characterization when $m = 1$, which we assume from now on.

The proof uses the *split chain* of Nummelin (see [12] and [13]), which we describe now.

Definition 0.5. *Let P be a transition kernel satisfying* (0.3) *for $\delta \in (0,1)$ and ν a probability measure. The split chain is a Markov chain Y_n on $\bar{\mathcal{S}} = \mathcal{S} \times [0,1]$, whose transition kernel $\bar{P}$ is as follows: if $x \notin C$, then $\bar{P}((x,t),\cdot) = P(x,\cdot) \otimes \lambda$, where λ is the uniform measure on $[0,1]$. If $x \in C$, then if $t \in [0,\delta]$ one sets $\bar{P}((x,t),\cdot) = \nu \otimes \lambda$, and if $t \in (\delta,1]$ then $\bar{P}((x,t),\cdot) = (1-\delta)^{-1}(\bar{P}(x,\cdot) - \delta\nu) \otimes \lambda$.*

Essentially, the corresponding chain behaves as the chain on $\mathcal{S}$, except when it enters C where the part of the transition kernel corresponding to $\delta\nu$ is explicitly separated from the rest.

For $x \in \mathcal{S}$, let $\mathbb{P}_{\bar{x}}$ denote the distribution of the Markov chain Y_n started from $\delta_x \otimes \lambda$. The first component of Y_n, living on $\mathcal{S}$, is then distributed as the original Markov chain started from x. In the same way, the chain Y_n started from $\bar{\pi} = \pi \otimes \lambda$ has a first projection which is distributed as the original Markov chain started from π. For obvious reasons, we still denote by X_n the first component of Y_n.

Let $\bar{C} = C \times [0,\delta]$. This is an atom of the chain Y_n, i.e., $\bar{P}(y,\cdot)$ does not depend on $y \in C$. We will show that the return time $\tau_{\bar{C}}$ to $\bar{C}$ has an exponential moment. Let $C' = C \times [0,1]$, and let U_n be the second component of Y_n. Each time the chain X_n enters C, i.e., Y_n enters C', then Y_n enters $\bar{C}$ if and only if $U_n \leqslant \delta$. Denote by t_k the k-th visit to C' of the chain Y_n, and note that (t_k) is an increasing sequence of stopping times. By the strong Markov property, it follows that (U_{t_k}) is an i.i.d. sequence of random variables with common distribution λ. Let $K(X_1,\dots,X_n) = \sum_{i=1}^n \mathbb{1}_C(X_i)$ denote the number of visits of X_i to C. For any $k \leqslant n$, $\{K(X_1,\dots,X_n) \geqslant k\} = \{t_k \leqslant n\}$. It follows that, for any $k \leqslant n$,

$$\begin{aligned}\mathbb{P}_{\bar{\pi}}(\tau_{\bar{C}} > n) &\leqslant \mathbb{P}_{\pi}(K(X_1,\dots,X_n) < k) + \mathbb{P}_{\bar{\pi}}(t_k \leqslant n, \tau_{\bar{C}} > n)\\ &\leqslant \mathbb{P}_{\pi}(K(X_1,\dots,X_n) < k) + \mathbb{P}_{\bar{\pi}}(t_k \leqslant n, U_{t_1} > \delta, \dots, U_{t_k} > \delta)\\ &\leqslant \mathbb{P}_{\pi}(K(X_1,\dots,X_n) < k) + (1-\delta)^k .\end{aligned}$$

Take $k = \varepsilon n$ for $\varepsilon = \pi(C)/2 < \pi(C)$. The subgaussian concentration inequality (0.6) applied to K gives, for some $c > 0$, the inequality $\mathbb{P}_{\pi}(K(X_1,\dots,X_n) \leqslant \varepsilon n) \leqslant 2e^{-cn}$. We deduce that $\tau_{\bar{C}}$ has an exponential moment, as desired, first for $\bar{\pi}$, then for its restriction to $\bar{C}$ since $\bar{\pi}(\bar{C}) > 0$, and then for any point in $\bar{C}$ since it is an atom (i.e., all starting points in $\bar{C}$ give rise to a chain with the same distribution after time 1). Hence, for some $\kappa > 1$,

$$\sup_{y \in \bar{C}} \mathbb{E}_y(\kappa^{\tau_{\bar{C}}}) < \infty.$$

By definition, this shows that the extended chain Y_n is geometrically ergodic in the sense of Definition 1. It is then easy to deduce that X_n also is, as follows. By (0.5), there exists a measurable function $\bar{V}$ which is finite π-almost everywhere such that

$$\|\bar{P}^n(y,\cdot) - \bar{\pi}\| \leqslant \bar{V}(y)\rho^n$$

for $\rho \in (0,1)$ and all y. We may take $\bar{V}(y) = \sup_{n \geqslant 1} \rho^{-n} \|\bar{P}^n(y,\cdot) - \bar{\pi}\|$. For $x \notin C$, this function $\bar{V}$ is constant on $\{x\} \times [0,1]$ since the chains Y_n starting from (x,t) or (x,t') have the same distribution after time 1. In the same way, for $x \in C$, the function $\bar{V}$ is constant on $\{x\} \times [0,\delta]$ and on $\{x\} \times (\delta,1]$. In particular, $\bar{V}$ is bounded, hence integrable, on π-almost every fiber $\{x\} \times [0,1]$. Letting $V(x) = \int \bar{V}(x,t)\,\mathrm{d}t$, we get $\|(\delta_x \otimes \lambda)\bar{P}^n - \bar{\pi}\| \leqslant V(x)\rho^n$ (we use the standard notation: for any measure ν on $\bar{\mathcal{S}}$, the measure $\nu\bar{P}^n$ on $\bar{\mathcal{S}}$ is defined by $\nu\bar{P}^n(A) = \int \bar{P}^n(y, A)\nu(dy)$). Since the first marginal of the chain Y_n started from $\delta_x \otimes \lambda$ is X_n started from x, this yields $\|P^n(x,\cdot) - \pi\| \leqslant V(x)\rho^n$, where V is finite π-almost everywhere. As we already mentioned, this implies that the chain is geometrically ergodic in the sense of Definition 1, by Theorem 15.4.2 in [11].

For the proof of Theorem 0.2, we will use the following coupling lemma. It says that the chains starting from any point in C or from the stationary distribution can be coupled in such a way that the coupling time has an exponential moment.

Let us first be more precise about what we call a coupling time. In general, a *coupling* between two random variables U and V is a way to realize these two random variables on a common probability space, usually to assert some closeness property between them. Formally, it is a probability space Ω^* together with two random variables U^* and V^* on Ω^*, distributed respectively like U and V. Abusing notations, we will usually implicitly identify U and U^*, and V and V^*.

Let μ and $\tilde{\mu}$ be two initial distributions on $\mathcal{S}$. They give rise to two chains X_n and $\tilde{X}_n$. We will construct couplings (X_n^*) and $(\tilde{X}_n^*)$ between these two chains with the following additional property: there exists a random variable $\tau : \Omega^* \to \mathbb{N}$, the *coupling time*, such that $X_n^* = \tilde{X}_n^*$ for all $n \geqslant \tau$.

Lemma 0.6. *Consider an irreducible aperiodic geometrically ergodic Markov chain and a small set C as in Definition 0.1. There exist constants $M_1 > 0$ and $\kappa > 1$ with the following property. Fix $x \in C$. Consider the Markov chains X_n and X_n' starting respectively from x, and from the stationary measure π. Then there exists a coupling between them with a coupling time τ such that*

$$\mathbb{E}(\kappa^\tau) \leqslant M_1.$$

While this lemma has a very classical flavor, we have not been able to locate a precise reference in the literature. We stress that the constants κ and M_1 are uniform, i.e., they do not depend on $x \in C$.

Proof. We will first give the proof when the chain is strongly aperiodic, i.e., m in Definition 0.1 is equal to 1. Then, we will deduce the general case from the strongly aperiodic one.

Assume $m = 1$. We use the split chain Y_n on $\bar{\mathcal{S}} = \mathcal{S} \times [0,1]$ introduced in Definition 0.5. We will use the notations of Remark 0.4, in particular $\bar{C} = C \times [0,\delta]$ and $\bar{\pi} = \pi \otimes \lambda$ is the stationary distribution for Y_n. Every time the Markov chain X_n on $\mathcal{S}$ returns to C, there is by definition a probability δ that the lifted chain Y_n enters $\bar{C}$. Hence, it follows from (0.4) that, for some $\kappa_1 > 1$,

$$\sup_{(x,s)\in C\times[0,1]} \mathbb{E}_{(x,s)}(\kappa_1^{\tau_{\bar{C}}}) < \infty. \tag{0.9}$$

In the same way, the entrance time to C starting from π has an exponential moment, by Theorem 2.5 (i) in [14]. It follows that, for some $\kappa_2 > 1$,

$$\mathbb{E}_{\bar{\pi}}(\kappa_2^{\tau_{\bar{C}}}) < \infty. \tag{0.10}$$

Define $T_0 = \inf\{n > 0 : Y_n \in \bar{C}\}$ and the return times

$$T_0 + \cdots + T_{i+1} = \inf\{n > T_0 + \cdots + T_i \ : \ Y_n \in \bar{C}\}.$$

Then T_0 is independent of $(T_i)_{i>0}$ and $T_1, T_2, \ldots$ are i.i.d. Denote by $\mathbb{P}_{\bar{\pi}}$ the probability measure on the underlying space starting from the invariant distribution $\bar{\pi}$, and by $\mathbb{P}_{\bar{x}}$

the probability measure starting from $\delta_x \otimes \lambda$ for $x \in \mathcal{S}$: the corresponding Markov chains lift the Markov chains on $\mathcal{S}$ starting from π and x respectively. We infer from (0.9) and (0.10) that there exist $\kappa_3 > 1$ and $M < \infty$ such that

$$\sup_{x \in C} \mathbb{E}_{\bar{x}}(\kappa_3^{T_0}) \leqslant M, \quad \mathbb{E}_{\bar{\pi}}(\kappa_3^{T_0}) \leqslant M \quad \text{and} \quad \mathbb{E}(\kappa_3^{T_1}) \leqslant M. \tag{0.11}$$

Let now (Y_n) and (Y'_n) be independant Markov chains on $\bar{\mathcal{S}}$ with common transition kernel $\bar{P}$, starting from $Y_0 \sim \delta_x \otimes \lambda$ with $x \in C$, and $Y'_0 \sim \bar{\pi}$. It follows from (0.11) that their respective return times $T_0 + \dots + T_i$ and $T'_0 + \dots + T'_i$ to $\bar{C}$ are such that:

- Both T_0 and T'_0 have a uniformly bounded exponential moment, i.e., $\mathbb{E}(\kappa_3^{T_0}) \leqslant M$ and $\mathbb{E}(\kappa_3^{T'_0}) \leqslant M$.
- The times T_i and T'_i for $i \geqslant 1$ are all independent, identically distributed, and their common distribution p is aperiodic with an exponential moment.

Define τ as

$$\tau = \inf\{n \geqslant 0 \,:\, \exists i \text{ with } n = T_0 + \dots + T_i \text{ and } \exists j \text{ with } n = T'_0 + \dots + T'_j\} + 1.$$

Lindvall [8, Page 66] proves that, under the above two assumptions, τ has an exponential moment: there exist $\kappa < 1$, $M_1 < \infty$, depending only on κ_3, M and p, such that $\mathbb{E}(\kappa^\tau) \leqslant M$.

Let $Y^*_n = Y_n$ if $n < \tau$ and $Y^*_n = Y'_n$ if $n \geqslant \tau$. As both $Y_{\tau-1}$ and $Y'_{\tau-1}$ belong to the atom $\bar{C}$ by definition of τ, the strong Markov property shows that $(Y^*_n)_{n\in\mathbb{N}}$ is distributed as $(Y_n)_{n\in\mathbb{N}}$. Hence, we have constructed a coupling between Y_n and Y'_n, with a coupling time τ which has an exponential moment, uniformly in x. Considering their first marginals, this yields the desired coupling between X_n (the Markov chain on $\mathcal{S}$ started from x) and X'_n (the Markov chain on $\mathcal{S}$ started from π). This concludes the proof when $m = 1$.

Assume now $m > 1$. In this case, one uses the m-skeleton of the original Markov chain, i.e., the Markov chain at times in $m\mathbb{N}$. By [11, Theorem 15.3.6], this m-skeleton is still geometrically ergodic, and the return times to C have a uniformly bounded exponential moment. Hence, the result with $m = 1$ yields a coupling between the chains $(X_{mn})_{n\in\mathbb{N}}$ and $(X'_{mn})_{n\in\mathbb{N}}$ started respectively from $x \in C$ and from π, with a coupling time τ having a uniformly bounded exponential moment. Thus, we deduce a coupling between $(X_n)_{n\in\mathbb{N}}$ and $(X'_n)_{n\in\mathbb{N}}$ together with a stopping time τ taking values in $m\mathbb{N}$, such that $X_{nm} = X'_{nm}$ for all $nm \in [\tau, +\infty) \cap m\mathbb{N}$ (from the technical point of view, this follows by seeing the fact that (X_{nm}) is a subsequence of (X_n) as a coupling between these two sequences, and then using the transitivity of couplings given by Lemma A.1 of [3]). This is not yet the desired coupling since there is no guarantee that $X_i = X'_i$ for $i \geqslant \tau$, $i \notin m\mathbb{N}$. Let $X^*_i = X_i$ for $i \leqslant \tau$, and $X^*_i = X'_i$ for $i > \tau$. It is distributed as (X_n) by the strong Markov property since $X_\tau = X'_\tau$, and satisfies $X^*_i = X'_i$ for all $i \geqslant \tau$ as desired. □

The following lemma readily follows.

Lemma 0.7. *Under the assumptions of Lemma 0.6, let $K(x_0, \dots)$ be a function of finitely or infinitely many variables, satisfying the boundedness condition* (0.1) *for some constants L_i. Then, for all $x \in C$,*

$$|\mathbb{E}_x(K(X_0, X_1, \dots)) - \mathbb{E}_\pi(K(X_0, X_1, \dots))| \leqslant M_1 \sum_{i \geqslant 0} L_i \rho^i,$$

where $M_1 > 0$ and $\rho < 1$ do not depend on x or K.

Proof. Consider the coupling given by the previous lemma, between the Markov chain X_n started from x and the Markov chain X'_n started from the stationary distribution π. Replacing successively X'_i with X_i for $i < \tau$, we get

$$|K(X_0, X_1, \dots) - K(X'_0, X'_1, \dots)| \leqslant \sum_{i<\tau} L_i.$$

Taking the expectation, we obtain

$$\begin{aligned}|\mathbb{E}(K(X_0, X_1, \dots)) - \mathbb{E}(K(X'_0, X'_1, \dots))| &\leqslant \mathbb{E}\left(\sum_{i<\tau} L_i\right) = \sum_i L_i \mathbb{P}(\tau > i) \\ &\leqslant \sum_i L_i \kappa^{-i} \mathbb{E}(\kappa^\tau) \leqslant M_1 \sum L_i \kappa^{-i}. \qquad \square\end{aligned}$$

We start the proof of Theorem 0.2. To simplify the notations, consider K as a function of infinitely many variables, with $L_i = 0$ for $i \geqslant n$. We start with several simple reductions in the first steps, before giving the real argument in Step 5.

First step: It suffices to prove (0.7)*, i.e., the concentration estimate starting from a point $x_0 \in C$.*

Indeed, fix some large $N > 0$, and consider the function

$$K_N(x_0, \dots, x_{n+N-1}) = K(x_N, \dots, x_{N+n-1}).$$

It satisfies $L_i(K_N) = 0$ for $i < N$ and $L_i(K_N) = L_{i-N}(K)$ for $N \leqslant i < n+N$. In particular, $\sum L_i(K_N)^2 = \sum L_i(K)^2$. Applying the inequality (0.7) to K_N, we get

$$\mathbb{P}_{x_0}(|K(X_N, \dots, X_{N+n-1}) - \mathbb{E}_{x_0} K(X_N, \dots, X_{N+n-1})| > t) \leqslant 2e^{-M_0^{-1} t^2 / \sum_{i \geqslant 0} L_i^2}. \quad (0.12)$$

Let

$$g_n(x) = \mathbb{E}(K(X_0, \dots, X_{n-1}) | X_0 = x) = \mathbb{E}(K(X_N, \dots, X_{N+n-1}) | X_N = x).$$

When $N \to \infty$, the distribution of X_N converges towards π in total variation, by (0.5). Since g_n is bounded, it follows that

$$\mathbb{E}_{x_0} K(X_N, \dots, X_{N+n-1}) = \mathbb{E}_{x_0} g_n(X_N) \to \mathbb{E}_\pi g_n(X_0) = \mathbb{E}_\pi K(X_0, \dots, X_{n-1}) \quad \text{as } N \to \infty.$$

Hence, for any $\varepsilon > 0$, their difference is bounded by ε if N is large enough. We obtain

$$\begin{aligned}\mathbb{P}_\pi(|K(X_0, \dots, &X_{n-1}) - \mathbb{E}_\pi K(X_0, \dots, X_{n-1})| > t) \\ &\leqslant \mathbb{P}_\pi(|K(X_0, \dots, X_{n-1}) - \mathbb{E}_{x_0} K(X_N, \dots, X_{N+n-1})| > t - \varepsilon) \\ &\leqslant \varepsilon + \mathbb{P}_{x_0}(|K(X_N, \dots, X_{N+n-1}) - \mathbb{E}_{x_0} K(X_N, \dots, X_{N+n-1})| > t - \varepsilon),\end{aligned}$$

using again the fact that the total variation between π and the distribution of X_N starting from x_0 is bounded by ε. Using (0.12) and letting then ε tend to 0, we obtain the desired concentration estimate (0.6) starting from π, i.e.,

$$\mathbb{P}_\pi(|K(X_0, \dots, X_{n-1}) - \mathbb{E}_\pi K(X_0, \dots, X_{n-1})| \geqslant t) \leqslant 2e^{-M_0^{-1} t^2 / \sum_{i \geqslant 0} L_i^2}.$$

Second step: It suffices to prove that, for $x_0 \in C$,

$$\mathbb{E}_{x_0}(e^{K - \mathbb{E}_{x_0} K}) \leqslant e^{M_2 \sum_{i \geqslant 0} L_i^2}, \quad (0.13)$$

for some constant M_2 independent of K.

Indeed, assume that this holds. Then, for any $\lambda > 0$,

$$\mathbb{P}_{x_0}(K - \mathbb{E}_{x_0} K > t) \leqslant \mathbb{E}_{x_0}(e^{\lambda K - \lambda \mathbb{E}_{x_0} K - \lambda t}) \leqslant e^{-\lambda t} e^{\lambda^2 M_2 \sum_{i \geqslant 0} L_i^2},$$

by (0.13). Taking $\lambda = t/(2M_2 \sum L_i^2)$, we get a bound $e^{-t^2/(4M_2 \sum L_i^2)}$. Applying also the same bound to $-K$, we obtain

$$\mathbb{P}_{x_0}(|K - \mathbb{E}_{x_0} K| > t) \leqslant 2e^{-\frac{t^2}{4M_2 \sum L_i^2}},$$

as desired.

Third step: Fix some $\varepsilon_0 > 0$*. It suffices to prove* (0.13) *assuming moreover that each* L_i *satisfies* $L_i \leqslant \varepsilon_0$*.*

Indeed, assume that (0.13) is proved whenever $L_i(K) \leqslant \varepsilon_0$ for all i. Consider now a general function K. Take an arbitrary point $x_* \in \mathcal{S}$. Define a new function $\tilde{K}$ by

$$\tilde{K}(x_0, \dots, x_{n-1}) = K(y_0, \dots, y_{n-1}),$$

where $y_i = x_i$ if $L_i(K) \leqslant \varepsilon_0$, and $y_i = x_*$ if $L_i(K) > \varepsilon_0$. This new function $\tilde{K}$ satisfies $L_i(\tilde{K}) = L_i(K) \mathbb{1}(L_i(K) \leqslant \varepsilon_0) \leqslant \varepsilon_0$. Therefore, it satisfies (0.13). Moreover, $|K - \tilde{K}| \leqslant \sum_{L_i(K) > \varepsilon_0} L_i(K) \leqslant \sum L_i(K)^2/\varepsilon_0$. Hence,

$$\mathbb{E}_{x_0}(e^{K - \mathbb{E}_{x_0} K}) \leqslant e^{2 \sum L_i(K)^2/\varepsilon_0} \mathbb{E}_{x_0}(e^{\tilde{K} - \mathbb{E}_{x_0} \tilde{K}}) \leqslant e^{2 \sum L_i(K)^2/\varepsilon_0} e^{M_2 \sum L_i(\tilde{K})^2}.$$

This is the desired inequality.

Let us now start the proof of (0.13) for a function K with $L_i \leqslant \varepsilon_0$ for all i. We consider the Markov chain $X_0, X_1, \dots$ starting from a fixed point $x_0 \in C$. We define a stopping time $\tau_i = \inf\{n \geqslant i \, : \, X_n \in C\}$. Let $\mathcal{F}_i$ be the σ-field corresponding to this stopping time: an event A is $\mathcal{F}_i$-measurable if, for all n, $A \cap \{\tau_i = n\}$ is measurable with respect to $\sigma(X_0, \dots, X_n)$. Let

$$D_i = \mathbb{E}(K \mid \mathcal{F}_i) - \mathbb{E}(K \mid \mathcal{F}_{i-1}).$$

It is $\mathcal{F}_i$-measurable. By the definition of D_i,

$$K(X_0, \dots, X_n) - \mathbb{E}_{x_0}(K(X_0, \dots, X_n))) = \sum_{i=1}^{n} D_i \, .$$

Fourth step: It suffices to prove that

$$\mathbb{E}(e^{D_i} \mid \mathcal{F}_{i-1}) \leqslant e^{M_3 \sum_{k \geqslant i} L_k^2 \rho^{k-i}}, \tag{0.14}$$

for some $M_3 > 0$ *and some* $\rho < 1$*, both independent of* K*.*

Indeed, assume that this inequality holds. Conditioning successively with respect to $\mathcal{F}_n$, then $\mathcal{F}_{n-1}$, and so on, we get

$$\mathbb{E}(e^{K - \mathbb{E}K}) = \mathbb{E}(e^{\sum D_i}) \leqslant e^{M_3 \sum_{i=0}^{n} \sum_{k \geqslant i} L_k^2 \rho^{k-i}} \leqslant e^{M_3/(1-\rho) \cdot \sum_i L_i^2}.$$

This is the desired inequality.

Fifth step: Proof of (0.14)*.*

Note first that on the set $\{\tau_{i-1} > i-1\}$ one has $\tau_{i-1} = \tau_i$, and consequently $D_i = 0$. Hence, the following decomposition holds:

$$\begin{aligned} D_i &= \sum_{j=i}^{\infty} (\mathbb{E}(K \mid \mathcal{F}_i) - \mathbb{E}(K \mid \mathcal{F}_{i-1})) \mathbb{1}_{\tau_i = j, \tau_{i-1} = i-1} \\ &= \sum_{j=i}^{\infty} (g_j(X_0, \dots, X_j) - g_{i-1}(X_0, \dots, X_{i-1})) \mathbb{1}_{\tau_i = j, \tau_{i-1} = i-1}, \end{aligned} \tag{0.15}$$

where

$$g_j(x_0,\dots,x_j) = \mathbb{E}_{X_0=x_j} K(x_0,\dots,x_j,X_1,\dots,X_{n-j-1}).$$

Here, we have used the fact that

$$\mathbb{E}(K \mid \mathcal{F}_i)\mathbb{1}_{\tau_i=j} = \mathbb{E}(K\mathbb{1}_{\tau_i=j} \mid \mathcal{F}_i) = \mathbb{E}(K\mathbb{1}_{\tau_i=j} \mid X_0,\dots,X_j) = \mathbb{E}(K \mid X_0,\dots,X_j)\mathbb{1}_{\tau_i=j},$$

which is commonly used in the proof of the strong Markov property for stopping times.

Let now

$$g_{j,\pi}(x_0,\dots,x_j) = \mathbb{E}_{X_0\sim\pi} K(x_0,\dots,x_j,X_1,\dots,X_{n-j-1}).$$

By Lemma 0.7, for any $x_j \in C$,

$$|g_j(x_0,\dots,x_j) - g_{j,\pi}(x_0,\dots,x_j)| \leqslant M_1 \sum_{k\geqslant j+1} L_k \rho^{k-j}. \tag{0.16}$$

From (0.15) and (0.16), we infer that

$$\begin{aligned} D_i = {}&\sum_{j=i}^{\infty} (g_{j,\pi}(X_0,\dots,X_j) - g_{i-1,\pi}(X_0,\dots,X_{i-1}))\mathbb{1}_{\tau_i=j,\tau_{i-1}=i-1} \\ &+ O\left(\sum_{k\geqslant \tau_i+1} L_k\rho^{k-\tau_i}\right) + O\left(\sum_{k\geqslant i} L_k\rho^{k-i}\right). \end{aligned} \tag{0.17}$$

Since π is the stationary measure, $g_{j,\pi}$ can also be written as

$$g_{j,\pi}(x_0,\dots,x_j) = \mathbb{E}_{X_0\sim\pi} K(x_0,\dots,x_j,X_{j-i+2},\dots,X_{n-i}).$$

It follows that

$$|g_{j,\pi}(x_0,\dots,x_j) - g_{i-1,\pi}(x_0,\dots,x_{i-1})| \leqslant \sum_{k=i}^{j} L_k. \tag{0.18}$$

Write $\tau = \tau_i - (i-1)$ for the return time to C of X_{i-1}. From (0.18), we get that

$$\sum_{j=i}^{\infty} (g_{j,\pi}(X_0,\dots,X_j) - g_{i-1,\pi}(X_0,\dots,X_{i-1}))\mathbb{1}_{\tau_i=j,\tau_{i-1}=i-1} \leqslant \left(\sum_{k=i}^{i+\tau-1} L_k\right)\mathbb{1}_{\tau_{i-1}=i-1}. \tag{0.19}$$

Since $\sum_{k\geqslant i} L_k\rho^{k-i} \leqslant \sum_{k=i}^{i+\tau-1} L_k + \sum_{k\geqslant i+\tau} L_k\rho^{k-i-\tau}$, it follows from (0.17) and (0.19) that

$$|D_i| \leqslant M_4 \left(\sum_{k=i}^{i+\tau-1} L_k + \sum_{k\geqslant i+\tau} L_k\rho^{k-i-\tau}\right)\mathbb{1}_{\tau_{i-1}=i-1}. \tag{0.20}$$

As all the L_k are bounded by ε_0, we obtain

$$|D_i| \leqslant M_4\varepsilon_0(\tau + 1/(1-\rho))\mathbb{1}_{\tau_{i-1}=i-1} \leqslant M_5\varepsilon_0\tau\mathbb{1}_{\tau_{i-1}=i-1}. \tag{0.21}$$

Choose $\sigma \in [\rho, 1)$. The equation (0.20) also gives

$$|D_i| \leqslant M_4\left(\sum_{k\geqslant i} L_k\sigma^{k-i}\sigma^{-\tau}\right)\mathbb{1}_{\tau_{i-1}=i-1}.$$

By the Cauchy-Schwarz inequality, this yields

$$\begin{aligned} |D_i|^2 &\leqslant M_4^2\sigma^{-2\tau}\left(\sum_{k\geqslant i} L_k^2\sigma^{k-i}\right)\left(\sum_{k\geqslant i}\sigma^{k-i}\right)\mathbb{1}_{\tau_{i-1}=i-1} \\ &\leqslant M_6\sigma^{-2\tau}\left(\sum_{k\geqslant i} L_k^2\sigma^{k-i}\right)\mathbb{1}_{\tau_{i-1}=i-1}. \end{aligned} \tag{0.22}$$

We have $e^t \leqslant 1 + t + t^2 e^{|t|}$ for all real t. Applying this inequality to D_i, taking the conditional expectation with respect to $\mathcal{F}_{i-1}$ and using that $\mathbb{E}(D_i \mid \mathcal{F}_{i-1}) = 0$, this gives

$$\mathbb{E}(e^{D_i} \mid \mathcal{F}_{i-1}) \leqslant 1 + \mathbb{E}(D_i^2 e^{|D_i|} \mid \mathcal{F}_{i-1}).$$

Combining this estimate with (0.21) and (0.22), we get

$$\begin{aligned}\mathbb{E}(e^{D_i} \mid \mathcal{F}_{i-1}) &\leqslant 1 + \mathbb{E}\left(M_6 e^{M_5 \varepsilon_0 \tau} \sigma^{-2\tau} \sum_{k \geqslant i} L_k^2 \sigma^{k-i} \mid \mathcal{F}_{i-1} \right) 1\!\!1_{\tau_{i-1} = i-1} \\ &\leqslant 1 + M_6 \left(\sum_{k \geqslant i} L_k^2 \sigma^{k-i} \right) \mathbb{E}\big(e^{M_5 \varepsilon_0 \tau} \sigma^{-2\tau} \mid X_{i-1}\big) 1\!\!1_{X_{i-1} \in C}.\end{aligned}$$

By the definition of geometric ergodicity (see Definition 0.1) one can choose ε_0 small enough and σ close enough to 1 in such a way that

$$\sup_{x \in C} \mathbb{E}\big(e^{M_5 \varepsilon_0 \tau} \sigma^{-2\tau} \mid X_{i-1} = x\big) < \infty\,.$$

It follows that

$$\mathbb{E}(e^{D_i} \mid \mathcal{F}_{i-1}) \leqslant 1 + M_7 \sum_{k \geqslant i} L_k^2 \sigma^{k-i} \leqslant e^{M_7 \sum_{k \geqslant i} L_k^2 \sigma^{k-i}}.$$

This concludes the proof of (0.14), and of Theorem 0.2. □

References

[1] Adamczak, R.: A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13**, (2008), 1000–1034. MR-2424985

[2] Adamczak, R. and Bednorz, W.: Exponential Concentration Inequalities for Additive Functionals of Markov Chains. (2013). arXiv:1201.3569v2

[3] Berkes, I. and Philipp, W.: Approximation theorems for independent and weakly dependent random vectors. *Ann. Probab.* **7**, (1979), 29–54. MR-515811

[4] Chazottes, J.-R. and Gouëzel, S.: Optimal concentration inequalities for dynamical systems. *Comm. Math. Phys.* **316**, (2012), 843–889. MR-2993935

[5] Chazottes, J.-R. and Redig, F.: Concentration inequalities for Markov processes via coupling. *Electron. J. Probab.* **14**, (2009), 1162–1180. MR-2511280

[6] Gouëzel, S. and Melbourne, I.: Moment bounds and concentration inequalities for slowly mixing dynamical systems. *Electron. J. Probab.* **19**, (2014), 30pp. MR-3272326

[7] Lezaud, P.: Chernoff-type bound for finite Markov chains. *Ann. Appl. Probab.* **8**, (1998), 849–867. MR-1627795

[8] Lindvall, T.: On coupling of discrete renewal processes. *Z. Wahrsch. Verw. Gebiete* **48**, (1979), 57–70. MR-533006

[9] McDiarmid, C.: On the method of bounded differences. Surveys in combinatorics, London Math. Soc. Lecture Note Ser. **141**, 148–188. *Cambridge Univ. Press,* Cambridge, 1989. MR-1036755

[10] Merlevède, F., Peligrad, M. and Rio, E.: A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probab. Theory Related Fields* **151**, (2011), 435–474. MR-2851689

[11] Meyn, S. P. and Tweedie, R. L.: Markov chains and stochastic stability. Communications and Control Engineering Series. *Springer-Verlag London Ltd.,* London, 1993. xvi+548 pp. MR-1287609

[12] Nummelin, E.: A splitting technique for Harris recurrent Markov chains. *Z. Wahrsch. Verw. Gebiete* **43**, (1978), 309–318. MR-0501353

[13] Nummelin, E.: General irreducible Markov chains and nonnegative operators. Cambridge Tracts in Mathematics. *Cambridge University Press*, Cambridge, 1984. xi+156 pp. MR-776608

[14] Nummelin, E. and Tuominen, P.: Geometric ergodicity of Harris recurrent Markov chains with applications to renewal theory. *Stochastic Process. Appl.* **12**, (1982), 187–202. MR-651903

[15] Rio, E.: Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math.* **330**, (2000), 905–908. MR-1771956

[16] Samson, P.-M.: Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.* **28**, (2000), 416–461. MR-1756011