Electron. J. Probab. **20** (2015), no. 79, 1–32.
ISSN: 1083-6489 DOI: 10.1214/EJP.v20-4039

Concentration inequalities for Markov chains by Marton couplings and spectral methods

Daniel Paulin*

Abstract

We prove a version of McDiarmid's bounded differences inequality for Markov chains, with constants proportional to the mixing time of the chain. We also show variance bounds and Bernstein-type inequalities for empirical averages of Markov chains. In the case of non-reversible chains, we introduce a new quantity called the "pseudo spectral gap", and show that it plays a similar role for non-reversible chains as the spectral gap plays for reversible chains.

Our techniques for proving these results are based on a coupling construction of Katalin Marton, and on spectral techniques due to Pascal Lezaud. The pseudo spectral gap generalises the multiplicative reversiblication approach of Jim Fill.

Keywords: concentration inequalities ; Markov chain; mixing time ; spectral gap; coupling.
AMS MSC 2010: Primary 60E15; 60J05; 60J10; 28A35, Secondary 05C81; 68Q87.
Submitted to EJP on January 6, 2015, final version accepted on June 3, 2015.

1 Introduction

Consider a vector of random variables

$$X := (X_1, X_2, \dots, X_n)$$

taking values in $\Lambda := (\Lambda_1 \times \ldots \times \Lambda_n)$, and having joint distribution $\mathbb{P}$. Let $f : \Lambda \to \mathbb{R}$ be a measurable function. Concentration inequalities are tail bounds of the form

$$\mathbb{P}(|f(X_1, \dots, X_n) - \mathbb{E}f(X_1, \dots, X_n)| \geq t) \leq g(t),$$

with $g(t)$ typically being of the form $2\exp(-t^2/C)$ or $2\exp(-t/C)$ (for some constant C, which might depend on n).

Such inequalities are known to hold under various assumptions on the random variables $X_1, \dots, X_n$ and on the function f. With the help of these bounds able to get information about the tails of $f(X)$ even in cases when the distribution of $f(X)$ is

*Department of Statistics and Applied Probability, National University of Singapore, Singapore.
E-mail: paulindani@gmail.com

complicated. Unlike limit theorems, these bounds hold non-asymptotically, that is for any fixed n. Our references on concentration inequalities are [24], and [3].

Most of the inequalities in the literature are concerned with the case when $X_1, \ldots, X_n$ are independent. In that case, very sophisticated, and often sharp bounds are available for many different types of functions. Such bounds have found many applications in discrete mathematics (via the probabilistic method), computer science (running times of randomized algorithms, pattern recognition, classification, compressed sensing), and statistics (model selection, density estimation).

Various authors have tried to relax the independence condition, and proved concentration inequalities under different dependence assumptions. However, unlike in the independent case, these bounds are often not sharp.

In this paper, we focus on an important type of dependence, that is, Markov chains. Many problems are more suitably modelled by Markov chains than by independent random variables, and MCMC methods are of great practical importance. Our goal in this paper is to generalize some of the most useful concentration inequalities from independent random variables to Markov chains.

We have found that for different types of functions, different methods are needed to obtain sharp bounds. In the case of sums, the sharpest inequalities can be obtained using spectral methods, which were developed by [28]. In this case, we show variance bounds and Bernstein-type concentration inequalities. For reversible chains, the constants in the inequalities depend on the spectral gap of the chain (if we denote it by γ, then the bounds are roughly $1/\gamma$ times weaker than in the independent case). In the non-reversible case, we introduce the "pseudo spectral gap",

$$\gamma_{\mathrm{ps}} := \text{ maximum of (the spectral gap of } (P^*)^k P^k \text{ divided by } k) \text{ for } k \geq 1,$$

and prove similar bounds using it. Moreover, we show that just like $1/\gamma$, $1/\gamma_{\mathrm{ps}}$ can also be bounded above by the mixing time of the chain (in total variation distance). For more complicated functions than sums, we show a version of McDiarmid's bounded differences inequality, with constants proportional to the mixing time of the chain. This inequality is proven by combining the martingale-type method of [4] and a coupling structure introduced by Katalin Marton.

An important feature of our inequalities is that they only depend on the spectral gap and the mixing time of the chain. These quantities are well studied for many important Markov chain models, making our bounds easily applicable.

Now we describe the organisation of the paper.

In Section 1.1, we state basic definitions about general state space Markov chains. This is followed by two sections presenting our results. In Section 2, we define Marton couplings, a coupling structure introduced in [35], and use them to show a version of McDiarmid's bounded differences inequality for dependent random variables, in particular, Markov chains. Examples include m-depedent random variables, hidden Markov chains, and a concentration inequality for the total variational distance of the empirical distribution from the stationary distribution. In Section 3, we show concentration results for sums of functions of Markov chains using spectral methods, in particular, variance bounds, and Bernstein-type inequalities. Several applications are given, including error bounds for hypothesis testing. In Section 4, we compare our results with the previous inequalities in the literature, and finally Section 5 contains the proofs of the main results.

This work grew out of the author's attempt to solve the "Spectral transportation cost inequality" conjecture stated in Section 6.4 of [22].

1.1 Basic definitions for general state space Markov chains

In this section, we are going to state some definitions from the theory of general state space Markov chains, based on [43]. If two random elements $X \sim P$ and $Y \sim Q$ are defined on the same probability space, then we call (X, Y) a coupling of the distributions P and Q. We define the total variational distance of two distributions P and Q defined on the same state space $(\Omega, \mathcal{F})$ as

$$d_{\mathrm{TV}}(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|, \tag{1.1}$$

or equivalently

$$d_{\mathrm{TV}}(P, Q) := \inf_{(X,Y)} \mathbb{P}(X \neq Y), \tag{1.2}$$

where the infimum is taken over all couplings (X, Y) of P and Q. Couplings where this infimum is achieved are called *maximal couplings* of P and Q (their existence is shown, for example, in [30]).

Note that there is also a different type of coupling of two random vectors called *maximal coupling* by some authors in the concentration inequalities literature, introduced by [16]. We will call this type of coupling as Goldstein's maximal coupling (which we will define precisely in Proposition 2.6). Let Ω be a Polish space. The *transition kernel* of a Markov chain with *state space* Ω is a set of probability distributions $P(x, \mathrm{d}y)$ for every $x \in \Omega$. A time homogenous Markov chain $X_0, X_1, \ldots$ is a sequence of random variables taking values in Ω satisfying that the conditional distribution of X_i given $X_0 = x_0, \ldots, X_{i-1} = x_{i-1}$ equals $P(x_{i-1}, \mathrm{d}y)$. We say that a distribution π on Ω is a stationary distribution for the chain if

$$\int_{x \in \Omega} \pi(\mathrm{d}x) P(x, \mathrm{d}y) = \pi(\mathrm{d}y).$$

A Markov chain with stationary distribution π is called *periodic* if there exist $d \geq 2$, and disjoints subsets $\Omega_1, \ldots, \Omega_d \subset \Omega$ with $\pi(\Omega_1) > 0$, $P(x, \Omega_{i+1}) = 1$ for all $x \in \Omega_i$, $1 \leq i \leq d-1$, and $P(x, \Omega_1) = 1$ for all $x \in \Omega_d$. If this condition is not satisfied, then we call the Markov chain *aperiodic*.

We say that a time homogenous Markov chain is *ϕ-irreducible*, if there exists a non-zero σ-finite measure ϕ on Ω such that for all $A \subset \Omega$ with $\phi(A) > 0$, and for all $x \in \Omega$, there exists a positive integer $n = n(x, A)$ such that $P^n(x, A) > 0$ (here $P^n(x, \cdot)$ denotes the distribution of X_n conditioned on $X_0 = x$).

The properties aperiodicity and ϕ-irreduciblility are sufficient for convergence to a stationary distribution.

Theorem (Theorem 4 of [43])**.** *If a Markov chain on a state space with countably generated σ-algebra is ϕ-irreducible and aperiodic, and has a stationary distribution π, then for π-almost every $x \in \Omega$,*

$$\lim_{n \to \infty} d_{\mathrm{TV}}(P^n(x, \cdot), \pi) = 0.$$

We define uniform and geometric ergodicity.

Definition 1.1. *A Markov chain with stationary distribution π, state space Ω, and transition kernel $P(x, dy)$ is* uniformly ergodic *if*

$$\sup_{x \in \Omega} d_{\mathrm{TV}}\left(P^n(x, \cdot), \pi\right) \leq M \rho^n, \qquad n = 1, 2, 3, \ldots$$

for some $\rho < 1$ and $M < \infty$, and we say that it is geometrically ergodic *if*

$$d_{\mathrm{TV}}\left(P^n(x, \cdot), \pi\right) \leq M(x) \rho^n, \qquad n = 1, 2, 3, \ldots$$

for some $\rho < 1$, where $M(x) < \infty$ for π-almost every $x \in \Omega$.

Remark 1.2. Aperiodic and irreducible Markov chains on finite state spaces are uniformly ergodic. Uniform ergodicity implies ϕ-irreducibility (with $\phi = \pi$), and aperiodicity.

The following definitions of the mixing time for Markov chains with general state space are based on Sections 4.5 and 4.6 of [26].

Definition 1.3 (Mixing time for time homogeneous chains)**.** *Let X_1, X_2, $X_3, \dots$ be a time homogeneous Markov chain with transition kernel $P(x, dy)$, Polish state space Ω, and stationary distribution π. Then t_{mix}, the mixing time of the chain, is defined by*

$$d(t) := \sup_{x \in \Omega} d_{\text{TV}}\left(P^t(x, \cdot), \pi\right), \; t_{\text{mix}}(\epsilon) := \min\{t : d(t) \le \epsilon\}, \text{ and}$$
$$t_{\text{mix}} := t_{\text{mix}}(1/4).$$

The fact that $t_{\text{mix}}(\epsilon)$ is finite for some $\epsilon < 1/2$ (or equivalently, t_{mix} is finite) is equivalent to the *uniform ergodicity* of the chain, see [43], Section 3.3. We will also use the following alternative definition, which also works for time inhomogeneous Markov chains.

Definition 1.4 (Mixing time for Markov chains without assuming time homogeneity)**.** *Let $X_1, \dots, X_N$ be a Markov chain with Polish state space $\Omega_1 \times \ldots \times \Omega_N$ (that is $X_i \in \Omega_i$). Let $\mathcal{L}(X_{i+t}|X_i = x)$ be the conditional distribution of X_{i+t} given $X_i = x$. Let us denote the minimal t such that $\mathcal{L}(X_{i+t}|X_i = x)$ and $\mathcal{L}(X_{i+t}|X_i = y)$ are less than ϵ away in total variational distance for every $1 \le i \le N - t$ and $x, y \in \Omega_i$ by $\tau(\epsilon)$, that is, for $0 < \epsilon < 1$, let*

$$\overline{d}(t) := \max_{1 \le i \le N-t} \sup_{x,y \in \Omega_i} d_{\text{TV}}\left(\mathcal{L}(X_{i+t}|X_i = x), \mathcal{L}(X_{i+t}|X_i = y)\right),$$
$$\tau(\epsilon) := \min\left\{t \in \mathbb{N} : \overline{d}(t) \le \epsilon\right\}.$$

Remark 1.5. One can easily see that in the case of time homogeneous Markov chains, by triangle inequality, we have

$$\tau(2\epsilon) \le t_{\text{mix}}(\epsilon) \le \tau(\epsilon). \tag{1.3}$$

Similarly to Lemma 4.12 of [26] (see also proposition 3.(e) of [43]), one can show that $\overline{d}(t)$ is subadditive

$$\overline{d}(t + s) \le \overline{d}(t) + \overline{d}(s), \tag{1.4}$$

and this implies that for every $k \in \mathbb{N}$, $0 \le \epsilon \le 1$,

$$\tau(\epsilon^k) \le k\tau(\epsilon), \text{ and thus } t_{\text{mix}}\left((2\epsilon)^k\right) \le k t_{\text{mix}}(\epsilon). \tag{1.5}$$

2 Marton couplings

In this section, we are going to prove concentration inequalities using Marton couplings. First, in Section 2.1, we introduce Marton couplings (which were originally defined in [35]), which is a coupling structure between dependent random variables. We are going to define a coupling matrix, measuring the strength of dependence between the random variables. We then apply this coupling structure to Markov chains by breaking the chain into blocks, whose length is proportional to the mixing time of the chain.

2.1 Preliminaries

In the following, we will consider dependent random variables $X = (X_1, \dots, X_N)$ taking values in a Polish space

$$\Lambda := \Lambda_1 \times \ldots \times \Lambda_N.$$

Let P denote the distribution of X, that is, $X \sim P$. Suppose that $Y = (Y_1, \dots, Y_N)$ is another random vector taking values in Λ, with distribution Q. We will refer to distribution of a vector $(X_1, \dots, X_k)$ as $\mathcal{L}(X_1, \dots, X_k)$, and

$$\mathcal{L}(X_{k+1}, \dots, X_N | X_1 = x_1, \dots, X_k = x_k)$$

will denote the conditional distribution of $X_{k+1}, \dots, X_N$ under the condition $X_1 = x_1, \dots, X_k = x_k$. Let $[N] := \{1, \dots, N\}$. We will denote the operator norm of a square matrix Γ by $\|\Gamma\|$. The following is one of the most important definitions of this paper. It has appeared in [35].

Definition 2.1 (Marton coupling)**.** *Let $X := (X_1, \dots, X_N)$ be a vector of random variables taking values in $\Lambda = \Lambda_1 \times \dots \times \Lambda_N$. We define a* Marton coupling *for X as a set of couplings*

$$\left(X^{(x_1, \dots, x_i, x_i')}, X'^{(x_1, \dots, x_i, x_i')} \right) \in \Omega \times \Omega,$$

for every $i \in [N]$, every $x_1 \in \Omega_1, \dots, x_i \in \Omega_i, x_i' \in \Omega_i$, satisfying the following conditions.

(i) $X_1^{(x_1, \dots, x_i, x_i')} = x_1, \quad \dots, \quad X_i^{(x_1, \dots, x_i, x_i')} = x_i,$
$X_1'^{(x_1, \dots, x_i, x_i')} = x_1, \quad \dots, \quad X_{i-1}'^{(x_1, \dots, x_i, x_i')} = x_{i-1}, \quad X_i'^{(x_1, \dots, x_i, x_i')} = x_i'.$

(ii) $\left(X_{i+1}^{(x_1, \dots, x_i, x_i')}, \dots, X_N^{(x_1, \dots, x_i, x_i')} \right)$
$\sim \mathcal{L}(X_{i+1}, \dots, X_N | X_1 = x_1, \dots, X_i = x_i),$
$\left(X'^{(x_1, \dots, x_i, x_i')}_{i+1}, \dots, X'^{(x_1, \dots, x_i, x_i')}_N \right)$
$\sim \mathcal{L}(X_{i+1}, \dots, X_N | X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_i = x_i').$

(iii) *If $x_i = x_i'$, then $X^{(x_1, \dots, x_i, x_i')} = X'^{(x_1, \dots, x_i, x_i')}$.*

For a Marton coupling, we define the mixing matrix $\Gamma := (\Gamma_{i,j})_{i,j \le N}$ *as an upper diagonal matrix with $\Gamma_{i,i} := 1$ for $i \le N$, and*

$$\Gamma_{j,i} := 0, \Gamma_{i,j} := \sup_{x_1, \dots, x_i, x_i'} \mathbb{P}\left[X_j^{(x_1, \dots, x_i, x_i')} \neq X'^{(x_1, \dots, x_i, x_i')}_j \right] \text{ for } 1 \le i < j \le N.$$

Remark 2.2. The definition says that a Marton coupling is a set of couplings the $\mathcal{L}(X_{i+1}, \dots, X_N | X_1 = x_1, \dots, X_i = x_i)$ and $\mathcal{L}(X_{i+1}, \dots, X_N | X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_i = x_i')$ for every $x_1, \dots, x_i, x_i'$, and every $i \in [N]$. The mixing matrix quantifies how close is the coupling. For independent random variables, we can define a Marton coupling whose mixing matrix equals the identity matrix. Although it is true that

$$\Gamma_{i,j} \ge \sup_{x_1, \dots, x_i, x_i'} d_{\mathrm{TV}} \left[\mathcal{L}(X_j | X_1 = x_1, \dots, X_i = x_i), \right.$$
$$\left. \mathcal{L}(X_j | X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_i = x_i') \right],$$

the equality does not hold in general (so we cannot replace the coefficients $\Gamma_{i,j}$ by the right hand side of the inequality). At first look, it might seem to be more natural to make a coupling between $\mathcal{L}(X_{i+1}, \dots, X_N | X_1 = x_1, \dots, X_i = x_i)$ and $\mathcal{L}(X_{i+1}, \dots, X_N | X_1 = x_1', \dots, X_i = x_i')$. For Markov chains, this is equivalent to our definition. The requirement in this definition is less strict, and allows us to get sharp inequalities for more dependence structures (for example, random permutations) than the stricter definition would allow.

We define the partition of a set of random variables.

Definition 2.3 (Partition)**.** *A* partition *of a set S is the division of S into disjoint non-empty subsets that together cover S. Analogously, we say that $\hat{X} := (\hat{X}_1, \dots, \hat{X}_n)$ is a*

partition of a vector of random variables $X = (X_1, \dots, X_N)$ *if* $(\hat{X}_i)_{1 \le i \le n}$ *is a partition of the set* $\{X_1, \dots, X_N\}$*. For a partition* $\hat{X}$ *of* X*, we denote the number of elements of* $\hat{X}_i$ *by* $s(\hat{X}_i)$ *(*size *of* $\hat{X}_i$*), and call* $s(\hat{X}) := \max_{1 \le i \le n} s(\hat{X}_i)$ *the* size of the partition*.*

Furthermore, we denote the set of indices of the elements of $\hat{X}_i$ *by* $\mathcal{I}(\hat{X}_i)$*, that is,* $X_j \in \hat{X}_i$ *if and only if* $j \in \mathcal{I}(X_i)$*. For a set of indices* $S \subset [N]$*, let* $X_S := \{X_j : j \in S\}$*. In particular,* $\hat{X}_i = X_{\mathcal{I}(\hat{X}_i)}$*. Similarly, if* X *takes values in the set* $\Lambda := \Lambda_1 \times \ldots \times \Lambda_N$*, then* $\hat{X}$ *will take values in the set* $\hat{\Lambda} := \hat{\Lambda}_1 \times \ldots \times \hat{\Lambda}_n$*, with* $\hat{\Lambda}_i := \Lambda_{\mathcal{I}(\hat{X}_i)}$*.*

Our main result of this section will be a McDiarmid-type inequality for dependent random variables, where the constant in the exponent will depend on the size of a particular partition, and the operator norm of the mixing matrix of a Marton coupling for this partition. The following proposition shows that for uniformly ergodic Markov chains, there exists a partition and a Marton coupling (for this partition) such that the size of the partition is comparable to the mixing time, and the operator norm of the coupling matrix is an absolute constant.

Proposition 2.4 (Marton coupling for Markov chains)**.** *Suppose that* $X_1, \dots, X_N$ *is a uniformly ergodic Markov chain, with mixing time* $\tau(\epsilon)$ *for any* $\epsilon \in [0, 1)$*. Then there is a partition* $\hat{X}$ *of* X *such that* $s(\hat{X}) \le \tau(\epsilon)$*, and a Marton coupling for for this partition* $\hat{X}$ *whose mixing matrix* Γ *satisfies*

$$\Gamma = (\Gamma_{i,j})_{i,j \le n} \le \begin{pmatrix} 1 & 1 & \epsilon & \epsilon^2 & \epsilon^3 & \dots \\ 0 & 1 & 1 & \epsilon & \epsilon^2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \tag{2.1}$$

with the inequality meant in each element of the matrices.

Remark 2.5. Note that the norm of Γ now satisfies that $\|\Gamma\| \le 1 + \frac{1}{1-\epsilon} = \frac{2-\epsilon}{1-\epsilon}$.

This result is a simple consequence of Goldstein's maximal coupling. The following proposition states this result in a form that is convenient for us (see [16], equation (2.1) on page 482 of [12], and Proposition 2 on page 442 of [45]).

Proposition 2.6 (Goldstein's maximal coupling)**.** *Suppose that* P *and* Q *are probability distributions on some common Polish space* $\Lambda_1 \times \ldots \times \Lambda_n$*, having densities with respect to some underlying distribution* ν *on their common state space. Then there is a coupling of random vectors* $X = (X_1, \dots, X_n), Y = (Y_1, \dots, Y_n)$ *such that* $\mathcal{L}(X) = P$*,* $\mathcal{L}(Y) = Q$*, and*

$$\mathbb{P}(X_i \ne Y_i) \le d_{\mathrm{TV}}(\mathcal{L}(X_i, \dots, X_n), \mathcal{L}(Y_i, \dots, Y_n)).$$

Remark 2.7. [32] assumes maximal coupling in each step, corresponding to

$$\Gamma = (\Gamma_{i,j})_{i,j \le n} \le \begin{pmatrix} 1 & a & a^2 & a^3 & \dots \\ 0 & 1 & a & a^2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \text{ with}$$

$$a := \sup_{x,y \in \Omega} d_{\mathrm{TV}}(P(x, \cdot), P(y, \cdot)). \tag{2.2}$$

The papers [45], [4], [5], [22] use the Marton coupling generated by Proposition 2.6. [35] shows that Marton couplings different from those generated by Proposition 2.6 can be also useful, especially when there is no natural sequential relation between the random variables (such as when they satisfy some Dobrushin-type condition). [42], and [8] generalise this coupling structure to bounded metric spaces. Our contribution is the introduction of the technique of partitioning.

Remark 2.8. In the case of time homogeneous Markov chains, Marton couplings (Definition 2.1) are in fact equivalent to couplings (X, X') between the distributions $\mathcal{L}(X_1, \dots, X_N | X_0 = x_0)$ and $\mathcal{L}(X_1, \dots, X_N | X_0 = x_0')$. Since the seminal paper [9], such couplings have been widely used to bound the convergence of Markov chains to their stationary distribution in total variation distance. If T is a random time such that for every $i \geq T$, $X_i = X_i'$ in the above coupling, then

$$d_{\mathrm{TV}}\left(P^t(x_0, \cdot), P^t(x_0', \cdot)\right) \leq \mathbb{P}(T > t).$$

In fact, even less suffices. Under the so called faithfulness condition of [44], the same bound holds if $X_T = X_T'$ (that is, the two chains are equal at a single time).

2.2 Results

Our main result in this section is a version of McDiarmid's bounded difference inequality for dependent random variables. The constants will depend on the size of the partition, and the norm of the coupling matrix of the Marton coupling.

Theorem 2.1 (McDiarmid's inequality for dependent random variables)**.** *Let $X = (X_1, \dots, X_N)$ be a sequence of random variables, $X \in \Lambda, X \sim P$. Let $\hat{X} = (\hat{X}_1, \dots, \hat{X}_n)$ be a partition of this sequence, $\hat{X} \in \hat{\Lambda}$, $\hat{X} \sim \hat{P}$. Suppose that we have a Marton coupling for $\hat{X}$ with mixing matrix Γ. Let $c \in \mathbb{R}_+^N$, and define $C(c) \in \mathbb{R}_+^n$ as*

$$C_i(c) := \sum_{j \in \mathcal{I}(\hat{X}_i)} c_j \text{ for } i \leq n. \tag{2.3}$$

If $f : \Lambda \to \mathbb{R}$ is such that

$$f(x) - f(y) \leq \sum_{i=1}^{n} c_i \mathbb{1}[x_i \neq y_i] \tag{2.4}$$

for every $x, y \in \Lambda$, then for any $\lambda \in \mathbb{R}$,

$$\log \mathbb{E}\left(e^{\lambda(f(X) - \mathbb{E}f(X))}\right) \leq \frac{\lambda^2 \cdot \|\Gamma \cdot C(c)\|^2}{8} \leq \frac{\lambda^2 \cdot \|\Gamma\|^2 \|c\|^2 s(\hat{X})}{8}. \tag{2.5}$$

In particular, this means that for any $t \geq 0$,

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X)| \geq t\right) \leq 2 \exp\left(\frac{-2t^2}{\|\Gamma \cdot C(c)\|^2}\right), \tag{2.6}$$

Remark 2.9. Most of the results presented in this paper are similar to (2.6), bounding the absolute value of the deviation of the estimate from the mean. Because of the absolute value, a constant 2 appears in the bounds. However, if one is interested in the bound on the lower or upper tail only, then this constant can be discarded.

A special case of this is the following result.

Corollary 2.10 (McDiarmid's inequality for Markov chains)**.** *Let $X := (X_1, \dots, X_N)$ be a (not necessarily time homogeneous) Markov chain, taking values in a Polish state space $\Lambda = \Lambda_1 \times \ldots \times \Lambda_N$, with mixing time $\tau(\epsilon)$ (for $0 \leq \epsilon \leq 1$). Let*

$$\tau_{\min} := \inf_{0 \leq \epsilon < 1} \tau(\epsilon) \cdot \left(\frac{2 - \epsilon}{1 - \epsilon}\right)^2. \tag{2.7}$$

Suppose that $f : \Lambda \to \mathbb{R}$ satisfies (2.4) for some $c \in \mathbb{R}_+^N$. Then for any $t \geq 0$,

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X)| \geq t\right) \leq 2 \exp\left(\frac{-2t^2}{\|c\|^2 \tau_{\min}}\right). \tag{2.8}$$

Remark 2.11. It is easy to show that for time homogeneous chains,

$$\tau_{\min} \leq \inf_{0\leq\epsilon<1} t_{\mathrm{mix}}(\epsilon/2)\cdot\left(\frac{2-\epsilon}{1-\epsilon}\right)^2 \leq 9t_{\mathrm{mix}}. \tag{2.9}$$

In many situations in practice, the Markov chain exhibits a cutoff, that is, the total variation distance decreases very rapidly in a small interval (see Figure 1 of [31]). If this happens, then $\tau_{\min}\approx 4t_{\mathrm{mix}}$.

Remark 2.12. This corollary could be also obtained as a consequence of theorems in previous papers ([45], [4], [42], [22]) applied to blocks of random variables. Note that by directly applying these theorems on $X_1, X_2, \ldots, X_N$, we would only obtain bounds of the form $2\exp\left(-\mathcal{O}\left(\frac{t^2}{\|c\|^2 t_{\mathrm{mix}}^2}\right)\right)$.

Remark 2.13. In Example 2.17, we are going to use this result to obtain a concentration inequality for the total variational distance between the empirical measure and the stationary distribution. Another application is given in [17], Section 3, where this inequality is used to bound the error of an estimate of the asymptotic variance of MCMC empirical averages.

In addition to McDiarmid's inequality, it is also possible to use Marton couplings to generalise the results of [45] and [35], based on transportation cost inequalities. In the case of Markov chains, this approach can be used to show Talagrand's convex distance inequality, Bernstein's inequality, and self-bounding-type inequalities, with constants proportional to the mixing time of the chain. We have decided not to include them here because of space considerations.

2.3 Applications

Example 2.14 (m-dependence)**.** We say that $X_1,\ldots,X_N$ are m-dependent random variables if for each $1\leq i\leq N-m$, $(X_1,\ldots,X_i)$ and $(X_{i+m},\ldots,X_N)$ are independent. Let $n:=\lceil\frac{N}{m}\rceil$, and

$$\hat{X}_1 := (X_1,\ldots,X_m),\ldots,\hat{X}_N := (X_{(n-1)m+1},\ldots,X_N).$$

We define a Marton coupling for $\hat{X}$ as follows.

$$\left(\hat{X}^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')}, \hat{X}'^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')}\right)$$

is constructed by first defining

$$\left(\hat{X}_1^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')},\ldots,\hat{X}_i^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')}\right) := (\hat{x}_1,\ldots,\hat{x}_i),$$
$$\left(\hat{X}'^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')}_1,\ldots,\hat{X}'^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')}_i\right) := (\hat{x}_1,\ldots,\hat{x}_{i-1},\hat{x}_i'),$$

and then defining

$$\left(\hat{X}_{i+1}^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')},\ldots,\hat{X}_n^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')}\right) \sim \mathcal{L}(\hat{X}_{i+1},\ldots,\hat{X}_n|\hat{X}_1=\hat{x}_1,\ldots,\hat{X}_i=\hat{x}_i).$$

After this, we set

$$\left(\hat{X}'^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')}_{i+2},\ldots,\hat{X}'^{(\hat{x}_1,\ldots,\hat{x}_n,\hat{x}_i')}_n\right) := \left(\hat{X}_{i+2}^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')},\ldots,\hat{X}_n^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')}\right),$$

and then define $\hat{X}'^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')}_{i+1}$ such that for any $(\hat{x}_{i+2},\ldots,\hat{x}_n)$,

$$\mathcal{L}(\hat{X}'^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')}_{i+1}|\hat{X}'^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')}_{i+2}=\hat{x}_{i+2},\ldots,\hat{X}_n^{(\hat{x}_1,\ldots,\hat{x}_i,\hat{x}_i')}=\hat{x}_n) =$$
$$\mathcal{L}(\hat{X}_{i+1}|\hat{X}_1=\hat{x}_1,\ldots,\hat{X}_i=\hat{x}_i,\hat{X}_{i+2}=\hat{x}_{i+2},\ldots,\hat{X}_n=\hat{x}_n).$$

Because of the m-dependence condition, this coupling is a Marton coupling, whose mixing matrix satisfies

$$\Gamma = (\Gamma_{i,j})_{i,j\leq n} \leq \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 1 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

We can see that $||\Gamma|| \leq 2$, and $s(\hat{X}) = m$, thus the constants in the exponent in McDiarmid's inequality are about $4m$ times worse than in the independent case.

Example 2.15 (Hidden Markov chains)**.** Let $\tilde{X}_1, \dots, \tilde{X}_N$ be a Markov chain (not necessarily homogeneous) taking values in $\tilde{\Lambda} = \tilde{\Lambda}_1 \times \dots \times \tilde{\Lambda}_N$, with distribution $\tilde{P}$. Let $X_1, \dots, X_N$ be random variables taking values in $\Lambda = \Lambda_1 \times \dots \times \Lambda_N$ such that the joint distribution of $(\tilde{X}, X)$ is given by

$$H(\mathrm{d}\tilde{x}, \mathrm{d}x) := \tilde{P}(\mathrm{d}\tilde{x}) \cdot \prod_{i=1}^{n} P_i(\mathrm{d}x_i|\tilde{x}_i),$$

that is, X_i are conditionally independent given $\tilde{X}$. Then we call $X_1, \dots, X_N$ a *hidden Markov chain*.

Concentration inequalities for hidden Markov chains have been investigated in [21], see also [22], Section 4.1.4. Here we show that our version of McDiarmid's bounded differences inequality for Markov chains in fact also implies concentration for hidden Markov chains.

Corollary 2.16 (McDiarmid's inequality for hidden Markov chains)**.** *Let $\tilde{\tau}(\epsilon)$ denote the mixing time of the underlying chain $\tilde{X}_1, \dots, \tilde{X}_N$, then Corollary 2.10 also applies to hidden Markov chains, with $\tau(\epsilon)$ replaced by $\tilde{\tau}(\epsilon)$ in* (2.7)*.*

Proof. It suffices to notice that $(X_1, \tilde{X}_1), (X_2, \tilde{X}_2), \dots$ is a Markov chain, whose mixing time is upper bounded by the mixing time of the underlying chain, $\tilde{\tau}(\epsilon)$. Since the function f satisfies (2.4) as a function of $X_1, \dots, X_N$, and it does not depends on $\tilde{X}_1, \dots, \tilde{X}_N$, it also satisfies this condition as a function of $(X_1, \tilde{X}_1)$, $(X_2, \tilde{X}_2)$, …, $(X_N, \tilde{X}_N)$. Therefore the result follows from Corollary 2.10. □

Example 2.17 (Convergence of empirical distribution in total variational distance)**.** Let $X_1, \dots, X_n$ be a uniformly ergodic Markov chain with countable state space Ω, unique stationary distribution π, and mixing time t_{mix}. In this example, we are going to study how fast is the empirical distribution, defined as $\pi_{em}(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i = x]$ for $x \in \Omega$, converges to the stationary distribution π in total variational distance. The following proposition shows a concentration bound for this distance, $d(X_1, \dots, X_n) := d_{\mathrm{TV}}(\pi_{em}(x), \pi)$.

Proposition 2.18. *For any $t \geq 0$,*

$$\mathbb{P}(|d(X_1, \dots, X_n) - \mathbb{E}(d)| \geq t) \leq 2\exp\left(-\frac{t^2 \cdot n}{4.5 t_{\mathrm{mix}}}\right).$$

Proof. The result is an immediate consequence of Corollary 2.10, by noticing that the function d satisfies (2.4) with $c_i = 1/n$ for $1 \leq i \leq n$. □

This proposition shows that the distance $d_{\mathrm{TV}}(\pi_{em}(x), \pi)$ is highly concentrated around its mean. In Example 3.15 of Section 3, we are going to bound the expectation $\mathbb{E}(d)$ in terms of spectral properties of the chain. When taken together, our results generalise the well-known Dvoretzky-Kiefer-Wolfowitz inequality (see [11], [37]) to the total variational distance case, for Markov chains.

Note that a similar bound was obtained in [20]. The main advantage of Proposition 2.18 is that the constants in the exponent of our inequality are proportional to the mixing time of the chain. This is sharper than the inequality in Theorem 2 of [20], where the constants are proportional to t_{mix}^2.

3 Spectral methods

In this section, we prove concentration inequalities for sums of the form $f_1(X_1) + \ldots + f_n(X_n)$, with $X_1, \ldots, X_n$ being a time homogeneous Markov chain. The proofs are based on spectral methods, due to [28].

Firstly, in Section 3.1, we introduce the spectral gap for reversible chains, and explain how to get bounds on the spectral gap from the mixing time and vice-versa. We then define a new quantity called the "pseudo spectral gap", for non-reversible chains. We show that its relation to the mixing time is very similar to that of the spectral gap in the reversible case.

After this, our results are presented in Section 3.2, where we state variance bounds and Bernstein-type inequalities for stationary Markov chains. For reversible chains, the constants depend on the spectral gap of the chain, while for non-reversible chains, the pseudo spectral gap takes the role of the spectral gap in the inequalities.

In Section 3.3, we state propositions that allow us to extend these results to non-stationary chains, and to unbounded functions.

Finally, Section 3.4 gives some applications of these bounds, including hypothesis testing, and estimating the total variational distance of the empirical measure from the stationary distribution.

In order to avoid unnecessary repetitions in the statement of our results, we will make the following assumption.

Assumption 3.1. *Everywhere in this section, we assume that $X = (X_1, \ldots, X_n)$ is a time homogenous, ϕ-irreducible, aperiodic Markov chain. We assume that its state space is a Polish space Ω, and that it has a Markov kernel $P(x, \mathrm{d}y)$ with unique stationary distribution π.*

3.1 Preliminaries

We call a Markov chain $X_1, X_2, \ldots$ on state space Ω with transition kernel $P(x, dy)$ *reversible* if there exists a probability measure π on Ω satisfying the detailed balance conditions,

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx) \text{ for every } x, y \in \Omega. \tag{3.1}$$

In the discrete case, we simply require $\pi(x)P(x, y) = \pi(y)P(y, x)$. It is important to note that reversibility of a probability measures implies that it is a stationary distribution of the chain.

Let $L^2(\pi)$ be the Hilbert space of complex valued measurable functions on Ω that are square integrable with respect to π. We endow $L^2(\pi)$ with the inner product $\langle f, g\rangle_\pi = \int f g^* \mathrm{d}\pi$, and norm $\|f\|_{2,\pi} := \langle f, f\rangle_\pi^{1/2} = (\mathbb{E}_\pi\left(f^2\right))^{1/2}$. P can be then viewed as a linear operator on $L^2(\pi)$, denoted by $\boldsymbol{P}$, defined as $(\boldsymbol{P}f)(x) := \mathbb{E}_{P(x,\cdot)}(f)$, and reversibility is equivalent to the self-adjointness of $\boldsymbol{P}$. The operator $\boldsymbol{P}$ acts on measures to the left, creating a measure $\mu\boldsymbol{P}$, that is, for every measurable subset A of Ω, $\mu\boldsymbol{P}(A) := \int_{x\in\Omega} P(x, A)\mu(\mathrm{d}x)$. For a Markov chain with stationary distribution π, we define the

spectrum of the chain as

$$S_2 := \Big\{ \lambda \in \mathbb{C} \setminus 0 : (\lambda \mathbf{I} - \boldsymbol{P})^{-1} \text{ does not exist as a bounded linear operator on } L^2(\pi) \Big\}.$$

For reversible chains, S_2 lies on the real line. We define the *spectral gap* for reversible chains as

$$\gamma := 1 - \sup\{\lambda : \lambda \in S_2, \lambda \neq 1\} \quad \text{if eigenvalue 1 has multiplicity 1,}$$
$$\gamma := 0 \quad \text{otherwise.}$$

For both reversible, and non-reversible chains, we define the *absolute spectral gap* as

$$\gamma^* := 1 - \sup\{|\lambda| : \lambda \in S_2, \lambda \neq 1\} \quad \text{if eigenvalue 1 has multiplicity 1,}$$
$$\gamma^* := 0 \quad \text{otherwise.}$$

In the reversible case, obviously, $\gamma \geq \gamma^*$. For a Markov chain with transition kernel $P(x, dy)$, and stationary distribution π, we defined the time reversal of P as the Markov kernel

$$P^*(x, dy) := \frac{P(y, dx)}{\pi(dx)} \cdot \pi(dy). \tag{3.2}$$

Then the linear operator $\boldsymbol{P}^*$ is the adjoint of the linear operator $\boldsymbol{P}$, on $L^2(\pi)$. We define a new quantity, called the *pseudo spectral gap* of $\boldsymbol{P}$, as

$$\gamma_{\mathrm{ps}} := \max_{k \geq 1} \left\{ \gamma((\boldsymbol{P}^*)^k \boldsymbol{P}^k)/k \right\}, \tag{3.3}$$

where $\gamma((\boldsymbol{P}^*)^k \boldsymbol{P}^k)$ denotes the spectral gap of the self-adjoint operator $(\boldsymbol{P}^*)^k \boldsymbol{P}^k$.

Remark 3.2. The pseudo spectral gap is a generalization of spectral gap of the multiplicative reversiblization ($\gamma(\boldsymbol{P}^* \boldsymbol{P})$), see [13]. We apply it to hypothesis testing for coin tossing (Example 3.19). Another application is given in [41], where we estimate the pseudo spectral gap of the Glauber dynamics with systemic scan in the case of the Curie-Weiss model. In these examples, the spectral gap of the multiplicative reversiblization is 0, but the pseudo spectral gap is positive.

If a distribution q on Ω is absolutely continuous with respect to π, we denote

$$N_q := \mathbb{E}_\pi \left(\left(\frac{\mathrm{d}\, q}{\mathrm{d}\, \pi} \right)^2 \right) = \int_{x \in \Omega} \frac{\mathrm{d}\, q}{\mathrm{d}\, \pi}(x) q(\mathrm{d}x). \tag{3.4}$$

If we q is not absolutely continuous with respect to π, then we define $N_q := \infty$. If q is localized on x, that is, $q(x) = 1$, then $N_q = 1/\pi(x)$.

The relations between the mixing and spectral properties for reversible, and non-reversible chains are given by the following two propositions (the proofs are included in Section 5.2).

Proposition 3.3 (Relation between mixing time and spectral gap)**.** *Suppose that our chain is reversible. For uniformly ergodic chains, for* $0 \leq \epsilon < 1$*,*

$$\gamma^* \geq \frac{1}{1 + \tau(\epsilon)/\log(1/\epsilon)}, \text{ in particular, } \gamma^* \geq \frac{1}{1 + t_{\mathrm{mix}}/\log(2)}. \tag{3.5}$$

For arbitrary initial distribution q*, we have*

$$d_{\mathrm{TV}}\left(q\boldsymbol{P}^n, \pi\right) \leq \frac{1}{2}(1 - \gamma^*)^n \cdot \sqrt{N_q - 1}, \tag{3.6}$$

implying that for reversible chains on finite state spaces, for $0 \leq \epsilon \leq 1$,

$$t_{\text{mix}}(\epsilon) \leq \frac{2\log(1/(2\epsilon)) + \log(1/\pi_{\min})}{2\gamma^*}, \text{ in particular,} \tag{3.7}$$

$$t_{\text{mix}} \leq \frac{2\log(2) + \log(1/\pi_{\min})}{2\gamma^*}, \tag{3.8}$$

with $\pi_{\min} = \min_{x\in\Omega} \pi(x)$.

Proposition 3.4 (Relation between mixing time and pseudo spectral gap)**.** *For uniformly ergodic chains, for* $0 \leq \epsilon < 1$,

$$\gamma_{\text{ps}} \geq \frac{1-\epsilon}{\tau(\epsilon)}, \text{ in particular, } \gamma_{\text{ps}} \geq \frac{1}{2t_{\text{mix}}}. \tag{3.9}$$

For arbitrary initial distribution q*, we have*

$$d_{\text{TV}}\left(q\boldsymbol{P}^n, \pi\right) \leq \frac{1}{2}(1-\gamma_{\text{ps}})^{(n-1/\gamma_{\text{ps}})/2} \cdot \sqrt{N_q - 1}, \tag{3.10}$$

implying that for chains with finite state spaces, for $0 \leq \epsilon \leq 1$,

$$t_{\text{mix}}(\epsilon) \leq \frac{1 + 2\log(1/(2\epsilon)) + \log(1/\pi_{\min})}{\gamma_{\text{ps}}}, \text{ in particular,} \tag{3.11}$$

$$t_{\text{mix}} \leq \frac{1 + 2\log(2) + \log(1/\pi_{\min})}{\gamma_{\text{ps}}}. \tag{3.12}$$

3.2 Results

In this section, we are going to state variance bounds and Bernstein-type concentration inequalities, for reversible and non-reversible chains (the proofs are included in Section 5.2). We state these inequalities for stationary chains (that is, $X_1 \sim \pi$), and use the notation $\mathbb{P}_\pi$ and $\mathbb{E}_\pi$ to emphasise this fact. In Proposition 3.10 of the next section, we will generalise these bounds to the non-stationary case.

Theorem 3.1 (Variance bound for reversible chains)**.** *Let* $X_1, \ldots, X_n$ *be a stationary, reversible Markov chain with spectral gap* γ*, and absolute spectral gap* γ^**. Let* f *be a measurable function in* $L^2(\pi)$*. Define* $V_f := \text{Var}_\pi(f)$*, and define the asymptotic variance* σ^2_{as} *as*

$$\sigma^2_{\text{as}} := \lim_{N\to\infty} N^{-1}\text{Var}_\pi\left(f(X_1) + \ldots + f(X_N)\right). \tag{3.13}$$

Then

$$\text{Var}_\pi\left[f(X_1) + \ldots + f(X_n)\right] \leq \frac{2nV_f}{\gamma}, \tag{3.14}$$

$$\left|\text{Var}_\pi\left[f(X_1) + \ldots + f(X_n)\right] - n\sigma^2\right| \leq 4V_f/\gamma^2. \tag{3.15}$$

More generally, let $f_1, \ldots, f_n$ *be functions in* $L^2(\pi)$*, then*

$$\text{Var}_\pi\left[f_1(X_1) + \ldots + f_n(X_n)\right] \leq \frac{2}{\gamma^*}\sum_{i=1}^{n}\text{Var}_\pi\left[f_i(X_i)\right]. \tag{3.16}$$

Remark 3.5. For empirical sums, the bound depends on the spectral gap, while for more general sums, on the absolute spectral gap. This difference is not just an artifact of the proof. If we consider a two state ($\Omega = \{0,1\}$) periodical Markov chain with transition matrix $\boldsymbol{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, then $\pi = (1/2, 1/2)$ is the stationary distribution, the

chain is reversible, and $-1, 1$ are the eigenvalues of $\boldsymbol{P}$. Now $\gamma = 2$, and $\gamma^* = 0$. When considering a function f defined as $f(0) = 1, f(1) = -1$, then $\sum_{i=1}^n f(X_i)$ is indeed highly concentrated, as predicted by (3.14). However, if we define functions $f_j(x) := (-1)^j \cdot f(x)$, then for stationary chains, $\sum_{i=1}^n f_i(X_i)$ will take values n and $-n$ with probability $1/2$, thus the variance is n^2. So indeed, we cannot replace γ^* by γ in (3.16).

Theorem 3.2 (Variance bound for non-reversible chains)**.** *Let $X_1, \ldots, X_n$ be a stationary Markov chain with pseudo spectral gap γ_{ps}. Let f be a measurable function in $L^2(\pi)$. Let V_f and σ_{as}^2 be as in Theorem 3.1. Then*

$$\mathrm{Var}_\pi\left[f(X_1) + \ldots + f(X_n)\right] \leq \frac{4nV_f}{\gamma_{\mathrm{ps}}}, \text{ and} \tag{3.17}$$

$$\left|\mathrm{Var}_\pi\left[f(X_1) + \ldots + f(X_n)\right] - n\sigma_{\mathrm{as}}^2\right| \leq 16V_f/\gamma_{\mathrm{ps}}^2. \tag{3.18}$$

More generally, let $f_1, \ldots, f_n$ be functions in $L^2(\pi)$, then

$$\mathrm{Var}_\pi\left[f_1(X_1) + \ldots + f_n(X_n)\right] \leq \frac{4}{\gamma_{\mathrm{ps}}} \sum_{i=1}^n \mathrm{Var}_\pi\left[f_i(X_i)\right]. \tag{3.19}$$

Theorem 3.3 (Bernstein inequality for reversible chains)**.** *Let $X_1, \ldots, X_n$ be a stationary reversible Markov chain with spectral gap γ, and absolute spectral gap γ^*. Let $f \in L^2(\pi)$, with $|f(x) - \mathbb{E}_\pi(f)| \leq C$ for every $x \in \Omega$. Let V_f and σ_{as}^2 be as in Theorem 3.1. Let $S := \sum_{i=1}^n f(X_i)$, then*

$$\mathbb{P}_\pi(|S - \mathbb{E}_\pi(S)| \geq t) \leq 2\exp\left(-\frac{t^2}{2n(\sigma_{\mathrm{as}}^2 + 0.8V_f) + 10tC/\gamma}\right), \tag{3.20}$$

and we also have

$$\mathbb{P}_\pi(|S - \mathbb{E}_\pi(S)| \geq t) \leq 2\exp\left(-\frac{t^2 \cdot \gamma}{4nV_f + 10tC}\right). \tag{3.21}$$

More generally, let $f_1, \ldots, f_n$ be $L^2(\pi)$ functions satisfying that $|f_i(x) - \mathbb{E}_\pi(f_i)| \leq C$ for every $x \in \Omega$. Let $S' := \sum_{i=1}^n f_i(X_i)$, and $V_{S'} := \sum_{i=1}^n \mathrm{Var}_\pi(f_i)$, then

$$\mathbb{P}_\pi(|S' - \mathbb{E}_\pi(S')| \geq t) \leq 2\exp\left(-\frac{t^2 \cdot (2\gamma^* - (\gamma^*)^2)}{8V_{S'} + 20tC}\right), \tag{3.22}$$

Remark 3.6. The inequality (3.20) is an improvement over the earlier result of [28], because it uses the asymptotic variance σ_{as}^2. In fact, typically $\sigma_{\mathrm{as}}^2 \gg V_f$, so the bound roughly equals $2\exp\left(-\frac{t^2}{2n\sigma_{\mathrm{as}}^2}\right)$ for small values of t, which is the best possible given the asymptotic normality of the sum. Note that a result very similar to (3.20) has been obtained for continuous time Markov processes by [29].

Theorem 3.4 (Bernstein inequality for non-reversible chains)**.**
Let $X_1, \ldots, X_n$ be a stationary Markov chain with pseudo spectral gap γ_{ps}. Let $f \in L^2(\pi)$, with $|f(x) - \mathbb{E}_\pi(f)| \leq C$ for every $x \in \Omega$. Let V_f be as in Theorem 3.1. Let $S := \sum_{i=1}^n f(X_i)$, then

$$\mathbb{P}_\pi(|S - \mathbb{E}_\pi(S)| \geq t) \leq 2\exp\left(-\frac{t^2 \cdot \gamma_{\mathrm{ps}}}{8(n + 1/\gamma_{\mathrm{ps}})V_f + 20tC}\right). \tag{3.23}$$

More generally, let $f_1, \ldots, f_n$ be $L^2(\pi)$ functions satisfying that $|f_i(x) - \mathbb{E}_\pi(f_i)| \leq C$ for every $x \in \Omega$. Let $S' := \sum_{i=1}^n f_i(X_i)$, and $V_{S'} := \sum_{i=1}^n \mathrm{Var}_\pi(f_i)$. Suppose that k_{ps} is a the smallest positive integer such that

$$\gamma_{\mathrm{ps}} = \gamma((\boldsymbol{P}^*)^{k_{\mathrm{ps}}}\boldsymbol{P}^{k_{\mathrm{ps}}})/k_{\mathrm{ps}}.$$

For $1 \le i \le k_{\mathrm{ps}}$, *let* $V_i := \sum_{j=0}^{\lfloor (n-i)/k_{\mathrm{ps}} \rfloor} \mathrm{Var}_\pi(f_{i+jk_{\mathrm{ps}}})$, *and let*

$$M := \left(\sum_{1 \le i \le k_{\mathrm{ps}}} V_i^{1/2} \right) \Big/ \min_{1 \le i \le k_{\mathrm{ps}}} V_i^{1/2}.$$

Then

$$\mathbb{P}_\pi(|S' - \mathbb{E}_\pi(S')| \ge t) \le 2 \exp\left(- \frac{t^2 \cdot \gamma_{\mathrm{ps}}}{8V_{S'} + 20tC \cdot M/k_{\mathrm{ps}}} \right). \tag{3.24}$$

Remark 3.7. The bound (3.24) is of similar form as (3.23) (nV_f is replaced by $V_{S'}$), the main difference is that instead of $20tC$, now we have $20tC \cdot M/k_{\mathrm{ps}}$ in the denominator. We are not sure whether the M/k_{ps} term is necessary, or it can be replaced by 1. Note that the bound (3.24) also applies if we replace V_i by $V_i' \ge V_i$ for each $1 \le i \le n$. In such a way, M/k_{ps} can be decreased, at the cost of increasing $V_{S'}$.

Remark 3.8. Theorems 3.3 and 3.4 can be applied to bound the error of MCMC simulations, see [17] for more details and examples. The generalisation to sums of the form $f_1(X_1) + \ldots f_n(X_n)$ can be used for "time discounted" sums, see Example 3.17.

Remark 3.9. The results of this paper generalise to continuous time Markov processes in a very straightforward way. To save space, we have not included such results in this paper, the interested reader can consult [40].

3.3 Extension to non-stationary chains, and unbounded functions

In the previous section, we have stated variance bounds and Bernstein-type inequalities for sums of the form $f_1(X_1) + \ldots + f_n(X_n)$, with $X_1, \ldots, X_n$ being a stationary time homogeneous Markov chain. Our first two propositions in this section generalise these bounds to the non-stationary case, when $X_1 \sim q$ for some distribution q (in this case, we will use the notations $\mathbb{P}_q$, and $\mathbb{E}_q$). Our third proposition extends the Bernstein-type inequalities to unbounded functions by a truncation argument. The proofs are included in Section 5.2.

Proposition 3.10 (Bounds for non-stationary chains)**.** *Let* $X_1, \ldots, X_n$ *be a time homogenous Markov chain with state space* Ω, *and stationary distribution* π. *Suppose that* $g(X_1, \ldots, X_n)$ *is real valued measurable function. Then*

$$\mathbb{P}_q(g(X_1, \ldots, X_n) \ge t) \le N_q^{1/2} \cdot \left[\mathbb{P}_\pi(g(X_1, \ldots, X_n) \ge t) \right]^{1/2}, \tag{3.25}$$

for any distribution q *on* Ω (N_q *was defined in* (3.4)*). Now suppose that we "burn" the first* t_0 *observations, and we are interested in bounds on a function* h *of* $X_{t_0+1}, \ldots, X_n$. *Firstly,*

$$\mathbb{P}_q(h(X_{t_0+1}, \ldots, X_n) \ge t) \le N_{q\boldsymbol{P}^{t_0}}^{1/2} \cdot \left[\mathbb{P}_\pi(h(X_1, \ldots, X_n) \ge t) \right]^{1/2}, \tag{3.26}$$

moreover,

$$\mathbb{P}_q(h(X_{t_0+1}, \ldots, X_n) \ge t) \le \mathbb{P}_\pi(h(X_{t_0+1}, \ldots, X_n) \ge t) + d_{\mathrm{TV}}\left(q\boldsymbol{P}^{t_0}, \pi\right). \tag{3.27}$$

Proposition 3.11 (Further bounds for non-stationary chains)**.** *In Proposition 3.10,* $N_{q\boldsymbol{P}^{t_0}}$ *can be further bounded. For reversible chains, we have*

$$N_{q\boldsymbol{P}^{t_0}} \le 1 + (N_q - 1) \cdot (1 - \gamma^*)^{2t_0}, \tag{3.28}$$

while for non-reversible chains,

$$N_{q\boldsymbol{P}^{t_0}} \le 1 + (N_q - 1) \cdot (1 - \gamma_{\mathrm{ps}})^{2(t_0 - 1/\gamma_{\mathrm{ps}})}. \tag{3.29}$$

Similarly, $d_{\mathrm{TV}}\left(q\boldsymbol{P}^n, \pi\right)$ *can be further bounded too. For reversible chains, we have, by* (3.6)*,*

$$d_{\mathrm{TV}}\left(q\boldsymbol{P}^n, \pi\right) \leq \frac{1}{2}(1-\gamma^*)^n \cdot \sqrt{N_q - 1}.$$

For non-reversible chains, by (3.10)*,*

$$d_{\mathrm{TV}}\left(q\boldsymbol{P}^n, \pi\right) \leq \frac{1}{2}(1-\gamma_{\mathrm{ps}})^{(n-1/\gamma_{\mathrm{ps}})/2} \cdot \sqrt{N_q - 1}.$$

Finally, for uniformly ergodic Markov chains,

$$d_{\mathrm{TV}}\left(q\boldsymbol{P}^n, \pi\right) \leq \inf_{0\leq\epsilon<1} \epsilon^{\lfloor n/\tau(\epsilon)\rfloor} \leq 2^{-\lfloor n/t_{\mathrm{mix}}\rfloor}. \tag{3.30}$$

The Bernstein-type inequalities assume boundedness of the summands. In order to generalise such bounds to unbounded summands, we can use truncation. For $a, b \in \mathbb{R}$, $a < b$, define

$$\mathcal{T}_{[a,b]}(x) = x \cdot \mathbb{1}[x \in [a,b]] + a \cdot \mathbb{1}[x < a] + b \cdot \mathbb{1}[x > b],$$

then we have the following proposition.

Proposition 3.12 (Truncation for unbounded summands)**.**
Let $X_1, X_2, \ldots, X_n$ *be a stationary Markov chain. Let* $f : \Omega \to \mathbb{R}$ *be a measurable function. Then for any* $a < b$*,*

$$\begin{aligned}
&\mathbb{P}_\pi\left(\sum_{i=1}^n f(X_i) \geq t\right) \\
&\leq \mathbb{P}_\pi\left(\sum_{i=1}^n \mathcal{T}_{[a,b]}(f(X_i)) \geq t\right) + \mathbb{P}_\pi\left(\min_{1\leq i\leq n} f(X_i) < a\right) + \mathbb{P}_\pi\left(\max_{1\leq i\leq n} f(X_i) > b\right) \\
&\leq \mathbb{P}_\pi\left(\sum_{i=1}^n \mathcal{T}_{[a,b]}(f(X_i)) \geq t\right) + \sum_{1\leq i\leq n} \mathbb{P}_\pi(f(X_i) \leq a) + \sum_{1\leq i\leq n} \mathbb{P}_\pi(f(X_i) \geq b).
\end{aligned}$$

Remark 3.13. A similar bound can be given for sums of the form $\sum_{i=1}^n f_i(X_i)$. One might think that such truncation arguments are rather crude, but in the Appendix of [40], we include a counterexample showing that it is not possible to obtain concentration inequalities for sums of unbounded functions of Markov chains that are of the same form as inequalities for sums of unbounded functions of independent random variables.

Remark 3.14. Note that there are similar truncation arguments in the literature for ergodic averages of unbounded functions of Markov chains, see [1], [2], and [38]. These rely on regeneration-type arguments, and thus apply to a larger class of Markov chains. However, our bounds are simpler, and the constants depend explicitly on the spectral properties of the Markov chain, whereas the constants in the previous bounds are less explicit.

3.4 Applications

In this section, we state four applications of our results, to the convergence of the empirical distribution in total variational distance, "time discounted" sums, bounding the Type-I and Type-II errors in hypothesis testing, and finally to coin tossing.

Example 3.15 (Convergence of empirical distribution in total variational distance revisited)**.** Let $X_1, \ldots, X_n$ be a uniformly ergodic Markov chain with countable state space Λ, unique stationary distribution π. We denote its empirical distribution by $\pi_{em}(x) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}[X_i = x]$. In Example 2.17, we have shown that the total variational distance of the empirical distribution and the stationery distribution, $d_{\mathrm{TV}}(\pi_{em}, \pi)$, is highly concentrated around its expected value. The following proposition bounds the expected value of this quantity.

Proposition 3.16. *For stationary, reversible chains,*

$$\mathbb{E}_\pi(d_{\mathrm{TV}}(\pi_{em},\pi)) \le \sum_{x\in\Lambda} \min\left(\sqrt{\frac{2\pi(x)}{n\gamma}}, \pi(x)\right). \tag{3.31}$$

For stationary, non-reversible chains, (3.31) *holds with* γ *replaced by* $\gamma_{\mathrm{ps}}/2$.

Proof. It is well known that the total variational distance equals

$$d_{\mathrm{TV}}(\pi_{em},\pi) = \sum_{x\in\Lambda} (\pi(x) - \pi_{em}(x))_+.$$

Using (3.14), we have

$$\mathbb{E}_\pi\left((\pi(x)-\pi_{em}(x))_+^2\right) \le \mathrm{Var}_\pi(\pi(x)-\pi_{em}(x)) \le \frac{2\pi(x)(1-\pi(x))}{n\gamma}.$$

By Jensen's inequality, we obtain that

$$\mathbb{E}_\pi[(\pi(x)-\pi_{em}(x))_+] \le \min\left(\sqrt{\frac{2\pi(x)}{n\gamma}}, \pi(x)\right),$$

and the statement follows by summing up. The proof of the non-reversible case is similar, using (3.17) to bound the variance. □

It is easy to see that for any stationary distribution π, our bound (3.31) tends to 0 as the sample size n tends to infinity. In the particular case of when π is an uniform distribution on a state space consisting of N elements, we obtain that

$$\mathbb{E}_\pi(d_{\mathrm{TV}}(\pi_{em},\pi)) \le \sqrt{\frac{2N}{n\gamma}},$$

thus $n \gg N/\gamma$ samples are necessary.

Example 3.17 (A vineyard model)**.** Suppose that we have a vineyard, which in each year, depending on the weather, produces some wine. We are going to model the weather with a two state Markov chain, where 0 corresponds to bad weather (freeze destroys the grapes), and 1 corresponds to good weather (during the whole year). For simplicity, assume that in bad weather, we produce no wine, while in good weather, we produce 1$ worth of wine. Let $X_1, X_2, \dots$ be a Markov chain of the weather, with state space $\Omega = \{0,1\}$, stationary distribution π, and absolute spectral gap γ^* (it is easy to prove that any irreducible two state Markov chain is reversible). We suppose that it is stationary, that is, $X_1 \sim \pi$.

Assuming that the rate of interest is r, the present discounted value of the wine produced is

$$W := \sum_{i=1}^{\infty} X_i (1+r)^{-i}. \tag{3.32}$$

It is easy to see that $\mathbb{E}(W) = \mathbb{E}_\pi(X_1)/r$. We can apply Bernstein's inequality for reversible Markov chains (Theorem 3.3) with $f_i(X_i) = X_i(1+r)^{-i}$ and $C = 1$, and use a limiting argument, to obtain that

$$\begin{aligned}\mathbb{P}(|W - \mathbb{E}_\pi(X_1)/r| \ge t) &\le 2\exp\left(-\frac{t^2\cdot(\gamma^* - (\gamma^*)^2/2)}{4\mathrm{Var}_\pi(X_1)\sum_{i=1}^{\infty}(1+r)^{-2i} + 10t}\right) \\ &= 2\exp\left(-\frac{t^2\cdot(\gamma^* - (\gamma^*)^2)}{4\mathrm{Var}_\pi(X_1)(1+r)^2/(r^2+2r) + 10t}\right).\end{aligned}$$

If the price of the vineyard on the market is p, satisfying $p < \mathbb{E}_\pi(X_1)/r$, then we can use the above formula with $t = \mathbb{E}_\pi(X_1)/r - p$ to upper bound the probability that the vineyard is not going to earn back its price.

If we would model the weather with a less trivial Markov chain that has more than two states, then it could be non-reversible. In that case, we could get a similar result using Bernstein's inequality for non-reversible Markov chains (Theorem 3.4).

Example 3.18 (Hypothesis testing)**.** The following example was inspired by [18]. Suppose that we have a sample $X = (X_1, X_2, \dots, X_n)$ from a stationary, finite state Markov chain, with state space Ω. Our two hypotheses are the following.

$$H_0 := \{\text{transition matrix is } P_0, \text{ with stationary dist. } \pi_0, \text{ and } X_1 \sim \pi_0\},$$
$$H_1 := \{\text{transition matrix is } P_1, \text{ with stationary dist. } \pi_1, \text{ and } X_1 \sim \pi_1\}.$$

Then the log-likelihood function of X given the two hypotheses are

$$l_0(X) := \log \pi_0(X_1) + \sum_{i=1}^{n-1} \log P_0(X_i, X_{i+1}),$$
$$l_1(X) := \log \pi_1(X_1) + \sum_{i=1}^{n-1} \log P_1(X_i, X_{i+1}).$$

Let

$$T(X) := l_0(X) - l_1(X) = \log\left(\frac{\pi_0(X_1)}{\pi_1(X_1)}\right) + \sum_{i=1}^{n-1} \log\left(\frac{P_0(X_i, X_{i+1})}{P_1(X_i, X_{i+1})}\right).$$

The most powerful test between these two hypotheses is the Neyman-Pearson likelihood ratio test, described as follows. For some $\xi \in \mathbb{R}$,

$$T(X)/(n-1) > \xi \Rightarrow \text{ Stand by } H_0, \qquad T(X)/(n-1) \le \xi \Rightarrow \text{ Reject } H_0.$$

Now we are going to bound the Type-I and Type-II errors of this test using our Bernstein-type inequality for non-reversible Markov chains.

Let $Y_i := (X_i, X_{i+1})$ for $i \ge 1$. Then $(Y_i)_{i\ge1}$ is a Markov chain. Denote its transition matrix by $\boldsymbol{Q}_0$, and $\boldsymbol{Q}_1$, respectively, under hypotheses H_0 and H_1 (these can be easily computed from $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$). Denote

$$\hat{T}(Y) := \sum_{i=1}^{n-1} \log\left(\frac{P_0(Y_i)}{P_1(Y_i)}\right) = \sum_{i=1}^{n-1} \log\left(\frac{P_0(X_i, X_{i+1})}{P_1(X_i, X_{i+1})}\right), \tag{3.33}$$

then

$$\frac{T(X)}{n-1} = \frac{\log(\pi_0(X_1)/\pi_1(X_1))}{n-1} + \frac{\hat{T}(Y)}{n-1}. \tag{3.34}$$

Let

$$\delta_0 := \max_{x,y\in\Omega} \log P_0(x,y) - \min_{x,y\in\Omega} \log P_0(x,y),$$

and similarly,

$$\delta_1 := \max_{x,y\in\Omega} \log P_1(x,y) - \min_{x,y\in\Omega} \log P_1(x,y),$$

and let $\delta := \delta_0 + \delta_1$. Suppose that $\delta < \infty$. Then $\left|\frac{\log(\pi_0(X_1)/\pi_1(X_1))}{n-1}\right| \le \frac{\delta}{n-1}$, implying that $|T(X)/(n-1) - \hat{T}(Y)/(n-1)| \le \delta/(n-1)$. Moreover, we also have $|\log P_0(Y_i) - \log P_1(Y_i)| \le \delta$.

It is easy to verify that the matrices $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_1$, except in some trivial cases, always correspond to non-reversible chains (even when P_0 and P_1 are reversible). Let

$$J_0 := \mathbb{E}_0\left(\log\frac{P_0(X_1,X_2)}{P_1(X_1,X_2)}\right), \text{ and } J_1 := \mathbb{E}_1\left(\log\frac{P_0(X_1,X_2)}{P_1(X_1,X_2)}\right).$$

Note that J_0 can be written as the relative entropy of two distributions, and thus it is positive, and J_1 is negative. By the stationary assumption, $\mathbb{E}_0(\hat{T}(Y)) = (n-1)J_0$ and $\mathbb{E}_1(\hat{T}(Y)) = (n-1)J_1$.

By applying Theorem 3.4 on $\hat{T}(Y)$, we have the following bounds on the Type-I and Type-II errors. Assuming that $J_0 - \delta/(n-1) \geq \xi \geq J_1 + \delta/(n-1)$,

$$\mathbb{P}_0\left(\frac{T(X)}{n-1} \leq \xi\right) \leq \exp\left(-\frac{(J_0-\delta/(n-1)-\xi)^2(n-1)\gamma_{\mathrm{ps}}(\boldsymbol{Q}_0)}{8V_0 + 20\delta\cdot(J_0-\delta/(n-1)-\xi)}\right), \tag{3.35}$$

$$\mathbb{P}_1\left(\frac{T(X)}{n-1} \geq \xi\right) \leq \exp\left(-\frac{(\xi-J_1-\delta/(n-1))^2(n-1)\gamma_{\mathrm{ps}}(\boldsymbol{Q}_1)}{8V_1 + 20\delta\cdot(\xi-J_1-\delta/(n-1))}\right). \tag{3.36}$$

Here $V_0 = \mathrm{Var}_0\left(\log\left(\frac{P_0(X_1,X_2)}{P_1(X_1,X_2)}\right)\right)$, $V_1 = \mathrm{Var}_1\left(\log\left(\frac{P_0(X_1,X_2)}{P_1(X_1,X_2)}\right)\right)$, and $\gamma_{\mathrm{ps}}(\boldsymbol{Q}_0)$ and $\gamma_{\mathrm{ps}}(\boldsymbol{Q}_1)$ are the pseudo spectral gaps of $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_1$.

Example 3.19 (Coin tossing)**.** Let $X_1,\ldots,X_n$ be the realisation of n coin tosses (1 corresponds to heads, and 0 corresponding to tails). It is natural to model them as i.i.d. Bernoulli random variables, with mean $1/2$. However, since the well-known paper of [7], we know that in practice, the coin is more likely to land on the same side again than on the opposite side. This opens up the possibility that coin tossing can be better modelled by a two state Markov chain with a non-uniform transition matrix. To verify this phenomenon, we have performed coin tosses with a Singapore 50 cent coin (made in 2011). We have placed the coin in the middle of our palm, and thrown it up about 40-50cm high repeatedly. We have included our data of 10000 coin tosses in the Appendix of [40]. Using Example 3.18, we can make a test between the following hypotheses.

H_0 - i.i.d. Bernoulli trials, i.e. transition matrix $\boldsymbol{P}_0 := \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$, and

H_1 - stationary Markov chain with transition matrix $\boldsymbol{P}_1 = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}$.

For these transition matrices, we have stationary distributions $\pi_0(0) = \pi_0(1) = 1/2$ and $\pi_1(0) = 1-\pi_1(1) = 1/2$. A simple computation gives that for these transition probabilities, using the notation of Example 3.18, we have $\delta_0 = 0$, $\delta_1 = \log(0.6) - \log(0.4) = 0.4055$, $J_0 = 2.0411\cdot 10^{-2}$, $J_1 = -2.0136\cdot 10^{-2}$, and $\delta = \delta_0 + \delta_1 = 0.4055$. The matrices $\boldsymbol{Q}_0$ and $\boldsymbol{Q_1}$ are

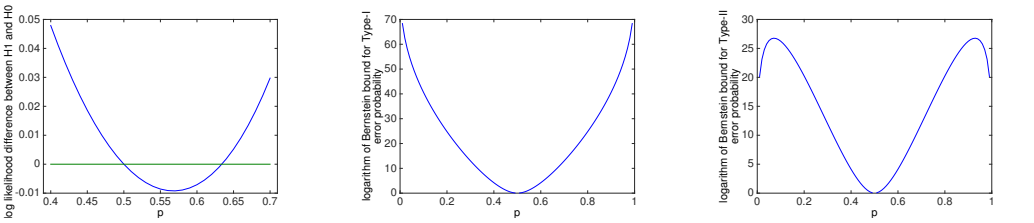
$$\boldsymbol{Q}_0 = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \end{pmatrix}, \text{ and } \boldsymbol{Q}_1 = \begin{pmatrix} 0.6 & 0.4 & 0 & 0 \\ 0 & 0 & 0.4 & 0.6 \\ 0.6 & 0.4 & 0 & 0 \\ 0 & 0 & 0.4 & 0.6 \end{pmatrix}.$$

We can compute $\boldsymbol{Q}_0^*$ and $\boldsymbol{Q}_1^*$ using (3.2),

$$\boldsymbol{Q}_0^* = \begin{pmatrix} 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0 & 0.5 \end{pmatrix}, \text{ and } \boldsymbol{Q}_1^* = \begin{pmatrix} 0.6 & 0 & 0.4 & 0 \\ 0.6 & 0 & 0.4 & 0 \\ 0 & 0.4 & 0 & 0.6 \\ 0 & 0.4 & 0 & 0.6 \end{pmatrix}.$$

As we can see, Q_0 and Q_1 are non-reversible. The spectral gap of their multiplicative reversiblization is $\gamma(\boldsymbol{Q}_0^*\boldsymbol{Q}_0) = \gamma(\boldsymbol{Q}_1^*\boldsymbol{Q}_1) = 0$. However, $\gamma((\boldsymbol{Q}_0^*)^2\boldsymbol{Q}_0^2) = 1$ and

Figure 1: Hypothesis testing for different values of the parameter p



(a) Log-likelihood difference (b) Logarithm of Type-I error bound (c) Logarithm of Type-II error bound

$\gamma((\boldsymbol{Q}_1^*)^2\boldsymbol{Q}_1^2) = 0.96$, thus $\gamma_{\mathrm{ps}}(\boldsymbol{Q}_0) = 0.5$, $\gamma_{\mathrm{ps}}(\boldsymbol{Q}_1) = 0.48$. The stationary distributions for Q_0 is $[0.25, 0.25, 0.25, 0.25]$, and for Q_1 is $[0.3, 0.2, 0.2, 0.3]$ (these probabilities correspond to the states $00, 01, 10$, and 11, respectively). A simple calculation gives $V_0 = 4.110 \cdot 10^{-2}$, $V_1 = 3.946 \cdot 10^{-2}$. By substituting these to (3.35) and (3.36), and choosing $\xi = 0$, we obtain the following error bounds.

$$\text{Type-I error.} \quad \mathbb{P}_0(T(X)/(n-1) \le \xi) \le \exp(-4.120) = 0.0150, \tag{3.37}$$

$$\text{Type-II error.} \quad \mathbb{P}_1(T(X)/(n-1) \ge \xi) \le \exp(-4.133) = 0.0160. \tag{3.38}$$

The actual value of $T(X)/(n-1)$ on our data is $\tilde{T}/(n-1) = -7.080 \cdot 10^{-3}$. Since $\tilde{T}/(n-1) < \xi$, we reject H0 (Bernoulli i.i.d. trials).

The choice of the transition matrix P_1 was somewhat arbitrary in the above argument. Indeed, we can consider a more general transition matrix of the form $\boldsymbol{P}_1 = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}$. We have repeated the above computations with this transition matrix, and found that for the interval $p \in (0.5, 0.635)$, H0 is rejected, while outside of this interval, we stand by H0. Three plots in Figure 1 show the log-likelihood differences, and the absolute value of the logarithm of the Bernstein bound on the Type-I and Type-II errors, respectively, for different values of p (in the first plot, we have restricted the range of p to $[0.4, 0.7]$ for better visibility). As we can see, the further away p is from 0.5, the smaller our error bounds become, which is reasonable since it becomes easier to distinguish between H0 and H1. Finally, from the first plot we can see that maximal likelihood estimate of p is $\hat{p} \approx 0.57$.

4 Comparison with the previous results in the literature

The literature of concentration inequalities for Markov chains is quite large, with many different approaches for both sums, and more general functions.

The first result in the case of general functions satisfying a form of the bounded differences condition (2.4) is Proposition 1 of [32], a McDiarmid-type inequality with constants proportional on $1/(1-a)^2$ (with a being the total variational distance contraction coefficient of the Markov chain in on steps, see (2.2)). The proof is based on the transportation cost inequality method. [33, 34, 36] extends this result, and proves Talagrand's convex distance inequality for Markov chains, with constants $1/(1-a)^2$ times worse than in the independent case. [45] extends Talagrand's convex distance inequality to more general dependency structures, and introduces the coupling matrix to quantify the strength of dependence between random variables. Finally, [35] further develops the results of [45], and introduces the coupling structure that we call Marton coupling in this paper. There are further extensions of this method to more general distances, and mixing conditions, see [42], [8], and [46]. Alternative, simpler approaches

to show McDiarmid-type inequalities for dependent random variables were developed in [4] (using an elementary martingale-type argument) and [23] (using martingales and linear algebraic inequalities). For time homogeneous Markov chains, their results are similar to Proposition 1 of [32].

In this paper, we have improved upon the previous results by showing a McDiarmid-type bounded differences inequality for Markov chains, with constants proportional to the mixing time of the chain, which can be much sharper than the previous bounds.

In the case of sums of functions of elements of Markov chains, there are two dominant approaches in the literature.

The first one is spectral methods, which use the spectral properties of the chain. The first concentration result of this type is [14], which shows a Hoeffding-type inequality for reversible chains. The method was further developed in [28], where Bernstein-type inequalities are obtained. A sharp version of Hoeffding's inequality for reversible chains was proven in [25].

The second popular approach in the literature is by regeneration-type minorisation conditions, see [15] and [10] for Hoeffding-type inequalities, and [2] for Bernstein-type inequalities. Such regeneration-type assumptions can be used to obtain bounds for a larger class of Markov chains than spectral methods would allow, including chains that are not geometrically ergodic. However, the bounds are more complicated, and the constants are less explicit.

In this paper, we have sharpened the bounds of [28]. In the case of reversible chains, we have proven a Bernstein-type inequality that involves the asymptotic variance, making our result essentially sharp. For non-reversible chains, we have proven Bernstein-type inequalities using the pseudo spectral gap, improving upon the earlier bounds of [28].

5 Proofs

5.1 Proofs by Marton couplings

Proof of Proposition 2.4. The main idea is that we divide the index set into mixing time sized parts. We define the following partition of X. Let $n = \left\lceil \frac{N}{\tau(\epsilon)} \right\rceil$, and

$$\begin{aligned}\hat{X} &:= (\hat{X}_1, \dots, \hat{X}_n) \\ &:= \left(\left(X_1, \dots, X_{\tau(\epsilon)}\right), \left(X_{\tau(\epsilon)+1}, \dots, X_{2\tau(\epsilon)}\right), \dots, \left(X_{(n-1)\tau(\epsilon)}, \dots, X_N\right) \right).\end{aligned}$$

Such a construction has the important property that $\hat{X}_1, \dots, \hat{X}_n$ is now a Markov chain, with ϵ-mixing time $\hat{\tau}(\epsilon) = 2$ (the proof of this is left to the reader as an exercise). Now we are going to define a Marton coupling for $\hat{X}$, that is, for $1 \leq i \leq n$, we need to define the couplings $\left(\hat{X}^{(\hat{x}_1,\dots,\hat{x}_i,\hat{x}_i')}, \hat{X'}^{(\hat{x}_1,\dots,\hat{x}_i,\hat{x}_i')}\right)$. These couplings are simply defined according to Proposition 2.6. Now using the Markov property, it is easy to show that for any $1 \leq i < j \leq n$, the total variational distance of $\mathcal{L}(\hat{X}_j, \dots \hat{X}_n | \hat{X}_1 = \hat{x}_1, \dots, \hat{X}_i = \hat{x}_i)$ and $\mathcal{L}(\hat{X}_j, \dots \hat{X}_n | \hat{X}_1 = \hat{x}_1, \dots, \hat{X}_{i-1} = \hat{x}_{i-1}, \hat{X}_i = \hat{x}_i')$ equals to the total variational distance of $\mathcal{L}(X_j \mid \hat{X}_1 = \hat{x}_1, \dots, \hat{X}_i = \hat{x}_i)$ and $\mathcal{L}(X_j | \hat{X}_1 = \hat{x}_1, \dots, \hat{X}_{i-1} = \hat{x}_{i-1}, \hat{X}_i = \hat{x}_i')$, and this can be bounded by ϵ^{j-i-1}, so the statement of the proposition follows. □

We will use the following Lemma in the proof of Theorem 2.1 (due to [6]).

Lemma 5.1. *Suppose $\mathcal{F}$ is a sigma-field and Z_1, Z_2, V are random variables such that*

1. $Z_1 \leq V \leq Z_2$
2. $\mathbb{E}(V|\mathcal{F}) = 0$
3. *Z_1 and Z_2 are $\mathcal{F}$-measurable.*

Then for all $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}(e^{\lambda V}|\mathcal{F}) \le e^{\lambda^2 (Z_2 - Z_1)^2/8}.$$

Proof of Theorem 2.1. We prove this result based on the martingale approach of [4] (a similar proof is possible using the method of [22]). Let $\hat{f}(\hat{X}) := f(X)$, then it satisfies that for every $\hat{x}, \hat{y} \in \hat{\Lambda}$,

$$\hat{f}(\hat{x}) - \hat{f}(\hat{y}) \le \sum_{i=1}^{n} \mathbb{1}[\hat{x}_i \neq \hat{y}_i] \cdot C_i(c).$$

Because of this property, we are going to first show that

$$\log \mathbb{E}\left(e^{\lambda(f(X) - \mathbb{E}f(X))}\right) \le \frac{\lambda^2 \cdot \|\Gamma \cdot c\|^2}{8} \tag{5.1}$$

under the assumption that there is a Marton coupling for X with mixing matrix Γ. By applying this inequality to $\hat{X}$, (2.5) follows.

Now we will show (5.1). Let us define $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$ for $i \le N$, and write $f(X) - \mathbb{E}f(X) = \sum_{i=1}^{N} V_i(X)$, with

$$\begin{aligned}
&V_i(X) := \mathbb{E}(f(X)|\mathcal{F}_i) - \mathbb{E}(f(X)|\mathcal{F}_{i-1}) \\
&= \int_{z_{i+1},\dots,z_N} \mathbb{P}(X_{i+1} \in \mathrm{d}z_{i+1}, \dots, X_N \in \mathrm{d}z_n | X_1, \dots, X_i) \\
&\quad \cdot f(X_1, \dots, X_i, z_{i+1}, \dots, z_N) \\
&- \int_{z_i,\dots,z_N} \mathbb{P}(X_i \in \mathrm{d}z_i, \dots, X_N \in \mathrm{d}z_n | X_1, \dots, X_{i-1}) \\
&\quad \cdot f(X_1, \dots, X_{i-1}, z_i, \dots, z_N) \\
&= \int_{z_{i+1},\dots,z_N} \mathbb{P}(X_{i+1} \in \mathrm{d}z_{i+1}, \dots, X_N \in \mathrm{d}z_n | X_1, \dots, X_i) \\
&\quad \cdot f(X_1, \dots, X_i, z_{i+1}, \dots, z_N) \\
&- \int_{z_i} \mathbb{P}(X_i \in \mathrm{d}z_i | X_1, \dots, X_{i-1}) \cdot \\
&\cdot \int_{z_{i+1},\dots,z_N} \mathbb{P}(X_{i+1} \in \mathrm{d}z_i, \dots, X_N \in \mathrm{d}z_n | X_1, \dots, X_{i-1}, X_i = z_i) \cdot \\
&\cdot f(X_1, \dots, X_{i-1}, z_i, \dots, z_N) \\
&\le \sup_{a \in \Lambda_i} \int_{z_{i+1},\dots,z_N} \mathbb{P}(X_{i+1} \in \mathrm{d}z_{i+1}, \dots, X_N \in \mathrm{d}z_n | X_1, \dots, X_{i-1}, X_i = a) \cdot \\
&\cdot f(X_1, \dots, X_{i-1}, a, z_{i+1}, \dots, z_N) \\
&- \inf_{b \in \Lambda_i} \int_{z_{i+1},\dots,z_N} \mathbb{P}(X_{i+1} \in \mathrm{d}z_{i+1}, \dots, X_N \in \mathrm{d}z_n | X_1, \dots, X_{i-1}, X_i = b) \cdot \\
&\cdot f(X_1, \dots, X_{i-1}, b, z_{i+1}, \dots, z_N) \\
&=: M_i(X) - m_i(X),
\end{aligned}$$

here $M_i(X)$ is the supremum, and $m_i(X)$ is the infimum, and we assume that these values are taken at a and b, respectively (one can take the limit in the following arguments if they do not exist).

After this point, [4] defines a coupling between the distributions

$$\begin{aligned}
&\mathcal{L}(X_{i+1}, \dots, X_N | X_1, \dots, X_{i-1}, X_i = a), \\
&\mathcal{L}(X_{i+1}, \dots, X_N | X_1, \dots, X_{i-1}, X_i = b)
\end{aligned}$$

as a maximal coupling of the two distributions. Although this minimises the probability that the two sequences differ in at least one coordinate, it is not always the best choice. We use a coupling between these two distributions that is induced by the Marton coupling for X, that is

$$(X^{(X_1,\dots,X_{i-1},a,b)}, X'^{(X_1,\dots,X_{i-1},a,b)}).$$

From the definition of the Marton coupling, we can see that

$$\begin{aligned}
&M_i(Y) - m_i(Y) = \mathbb{E}\left(f(X^{(X_1,\dots,X_{i-1},a,b)}) - f(X'^{(X_1,\dots,X_{i-1},a,b)}) \middle| X_1,\dots,X_{i-1}\right) \\
&\leq \mathbb{E}\left(\sum_{j=i}^{N} \mathbb{1}\left[X_j^{(X_1,\dots,X_{i-1},a,b)} \neq X'^{(X_1,\dots,X_{i-1},a,b)}_j \right] \cdot c_j \middle| X_1,\dots,X_{i-1} \right) \\
&\leq \sum_{j=i}^{N} \Gamma_{i,j} c_j.
\end{aligned}$$

Now using Lemma 5.1 with $V = V_i$, $Z_1 = m_i(X) - \mathbb{E}(f(X)|\mathcal{F}_{i-1})$, $Z_2 = M_i(X) - \mathbb{E}(f(X)|\mathcal{F}_{i-1})$, and $\mathcal{F} = \mathcal{F}_{i-1}$, we obtain that

$$\mathbb{E}\left(e^{\lambda V_i(X)} \middle| \mathcal{F}_{i-1} \right) \leq \exp\left(\frac{\lambda^2}{8} \left(\sum_{j=i}^{n} \Gamma_{i,j} c_j \right)^2 \right).$$

By taking the product of these, we obtain (5.1), and as a consequence, (2.5). The tail bounds follow by Markov's inequality. □

Proof of Corollary 2.10. We use the Marton coupling of Proposition 2.4. By the simple fact that $\|\Gamma\| \leq \sqrt{\|\Gamma\|_1 \|\Gamma\|_\infty}$, we have $\|\Gamma\| \leq 2/(1-\epsilon)$, so applying Theorem 2.1 and taking infimum in ϵ proves the result. □

5.2 Proofs by spectral methods

Proof of Proposition 3.3. The proof of the first part is similar to the proof of Proposition 30 of [39]. Let $L^\infty(\pi)$ be the set of π-almost surely bounded functions, equipped with the $\|\cdot\|_\infty$ norm ($\|f\|_\infty := \operatorname{ess\,sup}_{x\in\Omega} |f(x)|$). Then $L^\infty(\pi)$ is a Banach space. Since our chain is reversible, $\boldsymbol{P}$ is a self-adjoint, bounded linear operator on $L^2(\pi)$. Define the operator $\boldsymbol{\pi}$ on $L^2(\pi)$ as $\boldsymbol{\pi}(f)(x) := \mathbb{E}_\pi(f)$. This is a self-adjoint, bounded operator. Let $\boldsymbol{M} := \boldsymbol{P} - \boldsymbol{\pi}$, then we can express the absolute spectral gap γ^* of $\boldsymbol{P}$ as

$$\begin{aligned}
&\gamma^* = 1 - \sup\{|\lambda| : \lambda \in S_2(\boldsymbol{M})\}, \text{ with } S_2(\boldsymbol{M}) := \\
&\{\lambda \in \mathbb{C} \setminus 0 : (\lambda \mathbf{I} - \boldsymbol{M})^{-1} \text{ does not exists as a bounded lin. op. on } L^2(\pi)\}.
\end{aligned}$$

Thus $1-\gamma^*$ equals to the spectral radius of $\boldsymbol{M}$ on $L^2(\pi)$. It is well-known that the Banach space $L^\infty(\pi)$ is a dense subspace of the Hilbert space $L^2(\pi)$. Denote the restriction of $\boldsymbol{M}$ to $L^\infty(\pi)$ by $\boldsymbol{M}_\infty$. Then this is a bounded linear operator on a Banach space, so by Gelfand's formula, its spectral radius (with respect to the $\|\|_\infty$ norm) is given by $\lim_{k\to\infty} \|\boldsymbol{M}_\infty^k\|_\infty^{1/k}$. For some $0 \leq \epsilon < 1$, it is easy to see that $\|\boldsymbol{M}_\infty^{\tau(\epsilon)}\|_\infty \leq 2\epsilon$, and for $l \geq 1$, $\tau(\epsilon^l) \leq l\tau(\epsilon)$, thus $\|\boldsymbol{M}_\infty^{l\tau(\epsilon)}\|_\infty \leq 2\epsilon^l$. Therefore, we can show that

$$\lim_{k\to\infty} \|\boldsymbol{M}_\infty^k\|_\infty^{1/k} \leq \epsilon^{1/\tau(\epsilon)}. \tag{5.2}$$

For self-adjoint, bounded linear operators on Hilbert spaces, it is sufficient to control their spectral radius on a dense subspace, and therefore $\boldsymbol{M}$ has the same spectral radius

as $\boldsymbol{M}_\infty$. This implies that

$$\gamma^* \geq 1 - \epsilon^{1/\tau(\epsilon)} = 1 - \exp(-\log(1/\epsilon)/\tau(\epsilon)) \geq \frac{1}{1+\tau(\epsilon)/\log(1/\epsilon)}.$$

Now we turn to the proof of (3.6). For Markov chains on finite state spaces, (3.6) is a reformulation of Theorem 2.7 of [13] (using the fact that for reversible chains, the multiplicative reversiblization can be written as P^2). The same proof works for general state spaces as well. □

Proof of Proposition 3.4. In the non-reversible case, it is sufficient to bound

$$\gamma((\boldsymbol{P}^*)^{\tau(\epsilon)}\boldsymbol{P}^{\tau(\epsilon)}) = \gamma^*((\boldsymbol{P}^*)^{\tau(\epsilon)}\boldsymbol{P}^{\tau(\epsilon)}),$$

for some $0 \leq \epsilon < 1$. This is done similarly as in the reversible case. Firstly, note that $\gamma^*((\boldsymbol{P}^*)^{\tau(\epsilon)}\boldsymbol{P}^{\tau(\epsilon)})$ can be expressed as the spectral radius of the matrix $\boldsymbol{Q}_2 := (\boldsymbol{P}^*)^{\tau(\epsilon)}\boldsymbol{P}^{\tau(\epsilon)} - \boldsymbol{\pi}$. Denote the restriction of $\boldsymbol{Q}_2$ to $L^\infty(\pi)$ by $\boldsymbol{Q}_\infty$. Then by Gelfand's formula, $\boldsymbol{Q}_\infty$ has spectral radius $\lim_{k\to\infty}\|\boldsymbol{Q}_\infty^k\|_\infty^{1/k}$, which can be upper bounded by ϵ. Again, it is sufficient to control the spectral radius on a dense subspace, thus $\boldsymbol{Q}_2$ has the same spectral radius as $\boldsymbol{Q}_\infty$, and therefore $\gamma((\boldsymbol{P}^*)^{\tau(\epsilon)}\boldsymbol{P}^{\tau(\epsilon)}) \geq 1-\epsilon$. The result now follows from the definition of γ_{ps}.

Finally, we turn to the proof of (3.10). Note that for any $k \geq 1$,

$$d_{\mathrm{TV}}\left(q\boldsymbol{P}^n(\cdot), \pi\right) \leq d_{\mathrm{TV}}\left(q(\boldsymbol{P}^k)^{\lfloor n/k \rfloor}(\cdot), \pi\right).$$

Now using Theorem 2.7 of [13] with $\boldsymbol{M} = (\boldsymbol{P}^*)^k\boldsymbol{P}^k$, we obtain

$$d_{\mathrm{TV}}\left(q\boldsymbol{P}^n(\cdot), \pi\right) \leq \frac{1}{2}(1-\gamma((\boldsymbol{P}^*)^k\boldsymbol{P}^k))^{\lfloor n/k \rfloor/2} \cdot \sqrt{N_q - 1}.$$

Finally, we choose the k such that $\gamma((\boldsymbol{P}^*)^k\boldsymbol{P}^k) = k\gamma_{\mathrm{ps}}$, then

$$\begin{aligned} d_{\mathrm{TV}}\left(q\boldsymbol{P}^n(\cdot), \pi\right) &\leq \frac{1}{2}(1-k\gamma_{\mathrm{ps}})^{\lfloor n/k \rfloor/2} \cdot \sqrt{N_q - 1} \\ &\leq \frac{1}{2}(1-\gamma_{\mathrm{ps}})^{(n-k)/2} \cdot \sqrt{N_q - 1} \leq \frac{1}{2}(1-\gamma_{\mathrm{ps}})^{(n-1/\gamma_{\mathrm{ps}})/2} \cdot \sqrt{N_q - 1}. \end{aligned}$$ □

Proof of Theorem 3.1. Without loss of generality, we assume that $\mathbb{E}_\pi(f) = 0$, and $\mathbb{E}_\pi(f_i) = 0$, for $1 \leq i \leq n$. For stationary chains,

$$\mathbb{E}_\pi(f(X_i)f(X_j)) = \mathbb{E}_\pi(f\boldsymbol{P}^{j-i}(f)) = \mathbb{E}_\pi(f(\boldsymbol{P}-\boldsymbol{\pi})^{j-i}(f)),$$

for $1 \leq i \leq j \leq n$. By summing up in j from 1 to n, we obtain

$$\mathbb{E}_\pi\left(f(X_i)\sum_{j=1}^n f(X_j)\right) = \left\langle f, \left(\sum_{j=1}^n (\boldsymbol{P}-\boldsymbol{\pi})^{|j-i|}\right) f\right\rangle_\pi, \tag{5.3}$$

where

$$\begin{aligned} &\sum_{j=1}^n (\boldsymbol{P}-\boldsymbol{\pi})^{|j-i|} = \boldsymbol{I} + \sum_{k=1}^{i-1}(\boldsymbol{P}-\boldsymbol{\pi})^k + \sum_{k=1}^{n-i}(\boldsymbol{P}-\boldsymbol{\pi})^k = (\boldsymbol{I} - (\boldsymbol{P}-\boldsymbol{\pi})^i) \\ &\cdot (\boldsymbol{I} - (\boldsymbol{P}-\boldsymbol{\pi}))^{-1} + (\boldsymbol{I} - (\boldsymbol{P}-\boldsymbol{\pi})^{n-i+1}) \cdot (\boldsymbol{I} - (\boldsymbol{P}-\boldsymbol{\pi}))^{-1} - \boldsymbol{I}. \end{aligned}$$

Since $\boldsymbol{P}$ is reversible, the eigenvalues of $\boldsymbol{P}-\boldsymbol{\pi}$ lie in the interval $[-1, 1-\gamma]$. It is easy to show that for any $k \geq 1$ integer, the function $x \to (1-x^k)/(1-x)$ is non-negative on the

interval $[-1, 1-\gamma]$, and its maximum is less than or equal to $\max(1/\gamma, 1)$. This implies that for $x \in [-1, 1-\gamma]$, for $1 \le i \le n$,

$$-1 \le (1-x^i)/(1-x) + (1-x^{n-i+1})/(1-x) - 1 \le 2\max(1/\gamma, 1) - 1.$$

Now using the fact that $0 < \gamma \le 2$, we have $|(1-x^i)/(1-x)+(1-x^{n-i+1})/(1-x)-1| \le 2/\gamma$, and thus

$$\left\| \sum_{j=1}^{n} (\boldsymbol{P}-\boldsymbol{\pi})^{|j-i|} \right\|_{2,\pi} \le \frac{2}{\gamma}, \text{ thus } \mathbb{E}\left(f(X_i) \sum_{j=1}^{n} f(X_j) \right) \le \frac{2}{\gamma} \mathbb{E}_\pi\left(f^2\right).$$

Summing up in i leads to (3.14).

Now we turn to the proof of (3.15). Summing up (5.3) in i leads to

$$\mathbb{E}\left(\left(\sum_{i=1}^{n} f(X_i) \right)^2 \right) = \Big\langle f, \big[(2n\boldsymbol{I} - 2\left(\boldsymbol{I} - (\boldsymbol{P}-\boldsymbol{\pi})^{n-1}\right)(\boldsymbol{I} - (\boldsymbol{P}-\boldsymbol{\pi}))^{-1} \big) \cdot (\boldsymbol{I} - (\boldsymbol{P}-\boldsymbol{\pi}))^{-1} - n\boldsymbol{I} \big] f \Big\rangle_\pi, \tag{5.4}$$

so by the definition of σ_{as}^2, we can see that

$$\begin{aligned} &\sigma_{\text{as}}^2 = \left\langle f, \left[2(\boldsymbol{I} - (\boldsymbol{P}-\boldsymbol{\pi}))^{-1} - \boldsymbol{I} \right] f \right\rangle_\pi, \text{ and} \\ &\left| \mathrm{Var}_\pi \left(\sum_{i=1}^{n} f(X_i) \right) - n\sigma_{\text{as}}^2 \right| \\ &= \left| \left\langle f, \left[2(\boldsymbol{I} - (\boldsymbol{P}-\boldsymbol{\pi})^{n-1}) \cdot (\boldsymbol{I} - (\boldsymbol{P}-\boldsymbol{\pi}))^{-2} \right] f \right\rangle_\pi \right| \le 4V_f/\gamma^2. \end{aligned}$$

Now we turn to the proof of (3.16). For stationary chains, for $1 \le i, j \le n$,

$$\begin{aligned} \mathbb{E}_\pi(f_i(X_i) f_j(X_j)) &= \mathbb{E}_\pi(f_i \boldsymbol{P}^{j-i}(f_j)) = \mathbb{E}_\pi(f_i(\boldsymbol{P}-\boldsymbol{\pi})^{j-i}(f_j)) \\ &\le \|f_i\|_{2,\pi} \|f_j\|_{2,\pi} \|\boldsymbol{P}-\boldsymbol{\pi}\|_{2,\pi}^{j-i} \le \frac{1}{2}\mathbb{E}_\pi(f_i^2 + f_j^2)(1-\gamma^*)^{i-j}, \end{aligned}$$

and thus for any $1 \le i, j \le n$, $\mathbb{E}(f_i(X_i) f_j(X_j)) \le \frac{1}{2}\mathbb{E}_\pi(f_i^2 + f_j^2)(1-\gamma^*)^{|i-j|}$. Summing up in i and j proves (3.16). □

Proof of Theorem 3.2. Without loss of generality, we assume that $\mathbb{E}_\pi(f) = 0$, and $\mathbb{E}_\pi(f_i) = 0$ for $1 \le i \le n$. Now for $1 \le i, j \le n$,

$$\mathbb{E}_\pi(f(X_i) f(X_j)) = \mathbb{E}_\pi(f \boldsymbol{P}^{j-i}(f)) = \mathbb{E}_\pi(f(\boldsymbol{P}-\boldsymbol{\pi})^{j-i}(f)) \le V_f \left\| (\boldsymbol{P}-\boldsymbol{\pi})^{j-i} \right\|_{2,\pi},$$

and for any integer $k \ge 1$, we have

$$\left\| (\boldsymbol{P}-\boldsymbol{\pi})^{|j-i|} \right\| \le \left\| (\boldsymbol{P}-\boldsymbol{\pi})^k \right\|_{2,\pi}^{\lceil \frac{|j-i|}{k} \rceil} = \left\| (\boldsymbol{P}^*-\boldsymbol{\pi})^k (\boldsymbol{P}-\boldsymbol{\pi})^k \right\|_{2,\pi}^{\frac{1}{2}\lceil \frac{|j-i|}{k} \rceil}.$$

Let k_{ps} be the smallest positive integer such that $k_{\text{ps}}\gamma_{\text{ps}} = \gamma\left((\boldsymbol{P}^*)^{k_{\text{ps}}} \boldsymbol{P}^{k_{\text{ps}}}\right) = 1 - \left\| (\boldsymbol{P}^*-\boldsymbol{\pi})^k (\boldsymbol{P}-\boldsymbol{\pi})^k \right\|_{2,\pi}$, then $\mathbb{E}(f(X_i) f(X_j)) \le V_f(1 - k\gamma_{\text{ps}})^{\frac{1}{2}\lceil \frac{j-i}{k_{\text{ps}}} \rceil}$. By summing up in i and j, and noticing that

$$\sum_{l=0}^{\infty} (1-k_{\text{ps}}\gamma_{\text{ps}})^{\frac{1}{2}\lceil \frac{l}{k_{\text{ps}}} \rceil} \le 2\sum_{l=0}^{\infty} (1-k_{\text{ps}}\gamma_{\text{ps}})^{\lceil \frac{l}{k_{\text{ps}}} \rceil} = \frac{2k_{\text{ps}}}{k_{\text{ps}}\gamma_{\text{ps}}} = \frac{2}{\gamma_{\text{ps}}},$$

we can deduce (3.17). By the definition of σ_{as}^2, it follows that

$$\sigma_{\text{as}}^2 = \left\langle f, \left[2(\boldsymbol{I} - (\boldsymbol{P}-\boldsymbol{\pi}))^{-1} - \boldsymbol{I} \right] f \right\rangle_\pi,$$

and by comparing this with (5.4), we have

$$\left|\mathrm{Var}_\pi\left(\sum_{i=1}^n f(X_i)\right) - n\sigma_{\mathrm{as}}^2\right| = \left|\left\langle f, \left[2(\boldsymbol{I} - (\boldsymbol{P}-\boldsymbol{\pi})^{n-1})\cdot(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi}))^{-2}\right] f\right\rangle_\pi\right|.$$

In the above expression, $\|(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi})^{n-1})\|_{2,\pi} \le 2$, and for any $k \ge 1$,

$$\|(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi}))^{-1}\|_{2,\pi} \le \sum_{i=0}^\infty \|(\boldsymbol{P}-\boldsymbol{\pi})^i\|_{2,\pi} \le k\sum_{i=0}^\infty \|(\boldsymbol{P}-\boldsymbol{\pi})^k\|_{2,\pi}^i$$
$$= \frac{k}{1-\sqrt{1-\gamma((\boldsymbol{P}^*)^k\boldsymbol{P}^k)}} \le \frac{2k}{\gamma((\boldsymbol{P}^*)^k\boldsymbol{P}^k)}.$$

Optimizing in k gives $\|(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi}))^{-1}\|_{2,\pi} \le 2/\gamma_{\mathrm{ps}}$, and (3.18) follows. Finally, the proof of (3.19) is similar, and is left to the reader as exercise. □

Before starting the proof of the concentration bounds, we state a few lemmas that will be useful for the proofs.

Lemma 5.2. *Let $X_1,\dots,X_n$ be a time homogeneous, stationary Markov chain, with state space Ω, and stationary distribution π. Suppose that $f:\Omega\to\mathbb{R}$ is a bounded function in $L^2(\pi)$, and let $S := f(X_1)+\ldots+f(X_n)$. Then for any θ,*

$$\mathbb{E}_\pi(\exp(\theta S)) = \left\langle \mathbf{1}, (e^{\theta \boldsymbol{D}_f}\boldsymbol{P})^n \mathbf{1}\right\rangle_\pi \le \|e^{\theta\boldsymbol{D}_f/2}\boldsymbol{P}e^{\theta\boldsymbol{D}_f/2}\|_{2,\pi}^{n-1}\|e^{\theta f/2}\|_{2,\pi}^2, \tag{5.5}$$

here $\mathbf{1}$ is the constant 1 function on Ω, and $\boldsymbol{D}_f$ is the bounded linear operator on $L^2(\pi)$ corresponding to $\boldsymbol{D}_f(g)(x) = f(x)g(x)$ for every $x\in\Omega$, $g\in L^2(\pi)$.

More generally, if $f_1,\dots,f_n$ are bounded functions in $L^2(\pi)$, and $S' := f_1(X_1)+\ldots+f_n(X_n)$, then for any θ,

$$\begin{aligned}\mathbb{E}_\pi(\exp(\theta S')) &= \left\langle \mathbf{1}, \left(e^{\theta\boldsymbol{D}_{f_1}}\boldsymbol{P}\right)\cdot\ldots\cdot\left(e^{\theta\boldsymbol{D}_{f_n}}\boldsymbol{P}\right)\mathbf{1}\right\rangle_\pi \\ &= \left\langle \mathbf{1}, \left(\boldsymbol{P}e^{\theta\boldsymbol{D}_{f_1}}\right)\cdot\ldots\cdot\left(\boldsymbol{P}e^{\theta\boldsymbol{D}_{f_n}}\right)\mathbf{1}\right\rangle_\pi \\ &\le \|\boldsymbol{P}e^{\theta\boldsymbol{D}_{f_1}}\|_{2,\pi}\cdot\ldots\cdot\|\boldsymbol{P}e^{\theta\boldsymbol{D}_{f_n}}\|_{2,\pi}.\end{aligned} \tag{5.6}$$

Proof. This result is well known, it follows by a straightforward application of the Markov property. □

Lemma 5.3. *Suppose that $f\in L^2(\pi)$, $-1\le f\le 1$, $\mathbb{E}_\pi(f)=0$, then for reversible $\boldsymbol{P}$, for $0<\theta<\gamma/10$, we have*

$$\|e^{\theta\boldsymbol{D}_f}\boldsymbol{P}e^{\theta\boldsymbol{D}_f}\|_{2,\pi} \le 1 + \frac{4V_f}{\gamma}\cdot\theta^2\cdot\left(1-\frac{10\theta}{\gamma}\right)^{-1}, \text{ and} \tag{5.7}$$

$$\|e^{\theta\boldsymbol{D}_f}\boldsymbol{P}e^{\theta\boldsymbol{D}_f}\|_{2,\pi} \le 1 + 2(\sigma_{\mathrm{as}}^2 + 0.8V_f)\cdot\theta^2\cdot\left(1-\frac{10\theta}{\gamma}\right)^{-1}, \tag{5.8}$$

where $V_f := \mathbb{E}_\pi(f^2)$ and $\sigma_{\mathrm{as}}^2 := \lim_{N\to\infty}\frac{1}{N}\mathrm{Var}_\pi(f(X_1)+\ldots+f(X_N))$.

Proof. (5.7) is proven in [27] (pages 47 and 97), see also [28]. We prove (5.8) using a refinement of the same argument. Let us assume, without loss of generality, that our Markov chain has a finite state space (the general state space case can be proven analogously, see page 97 of [27]). We start by noting that the positive definite matrix $e^{\theta\boldsymbol{D}_f}\boldsymbol{P}e^{\theta\boldsymbol{D}_f}$ is similar to the matrix $\boldsymbol{P}(2\theta) := \boldsymbol{P}e^{2\theta\boldsymbol{D}_f}$. Using the Ferron-Probenius

theorem, it follows that $\boldsymbol{P}(2\theta)$ has real eigenvalues, and $\|e^{\theta \boldsymbol{D}_f}\boldsymbol{P}e^{\theta \boldsymbol{D}_f}\|_{2,\pi} = \lambda_{\max}(\boldsymbol{P}(2\theta))$ (the maximal eigenvalue).

Define the operator $\boldsymbol{\pi}$ on $L^2(\pi)$ as $\boldsymbol{\pi}(f)(x) = \mathbb{E}_\pi(f)$ for any $x \in \Omega$. Denote

$$\boldsymbol{Z} := \sum_{n=0}^{\infty}(\boldsymbol{P}^n - \boldsymbol{\pi}) = \sum_{n=0}^{\infty}(\boldsymbol{P} - \boldsymbol{\pi})^n = (\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{\pi})^{-1},$$

$\boldsymbol{Z}^{(0)} := -\boldsymbol{\pi}$, and $\boldsymbol{Z}^{(k)} := \boldsymbol{Z}^k$ for $k \geq 1$. Then we have $\|\boldsymbol{Z}\|_\pi = 1/\gamma$. By page 46 of [27], using the theory of linear perturbations, for $0 \leq r \leq \gamma/3$, we have

$$\lambda_{\max}(\boldsymbol{P}(r)) = 1 + \sum_{n=1}^{\infty}\beta^{(n)}r^n, \text{ with}$$

$$\beta^{(n)} = \sum_{p=1}^{n}\frac{-1}{p}\sum_{\substack{\nu_1+\ldots+\nu_p=n\\ k_1+\ldots+k_p=p-1\\ v_i\geq 1, k_j\geq 0}}\frac{1}{\nu_1!\cdots\nu_p!}\mathrm{tr}\left[\boldsymbol{P}\boldsymbol{D}_f^{\nu_1}\boldsymbol{Z}^{(k_1)}\cdots\boldsymbol{P}\boldsymbol{D}_f^{\nu_p}\boldsymbol{Z}^{(k_p)}\right].$$

Now for every integer valued vector $(k_1, \ldots, k_p)$ satisfying $k_1 + \ldots + k_p = p - 1$, $k_i \geq 0$, at least one of the indices must be 0. Suppose that the lowest such index is i, then we define $(k'_1, \ldots, k'_p) := (k_{i+1}, \ldots, k_p, k_1, \ldots, k_i)$, (a "rotation" of the original vector). We define $(\nu'_1, \ldots, \nu'_p)$ analogously. Using the fact that such rotation of matrices does not change the trace, and that $\boldsymbol{Z}^{(k'_p)} = \boldsymbol{Z}^{(0)} = -\boldsymbol{\pi}$, we can write

$$\beta^{(n)} = \sum_{p=1}^{n}\frac{1}{p}\sum_{\substack{\nu_1+\ldots+\nu_p=n\\ k_1+\ldots+k_p=p-1\\ v_i\geq 1, k_j\geq 0}}\frac{1}{\nu_1!\cdots\nu_p!}\left\langle f^{\nu'_1}, \boldsymbol{Z}^{(k'_1)}\boldsymbol{P}\boldsymbol{D}_f^{\nu_2}\cdots\boldsymbol{Z}^{(k'_{p-1})}\boldsymbol{P}f^{\nu'_p}\right\rangle_\pi. \tag{5.9}$$

After a simple calculation, we obtain $\beta^{(1)} = 0$, and $\beta^{(2)} = \langle f, \boldsymbol{Z}f\rangle_\pi - (1/2)\langle f, f\rangle_\pi$. By page 48-49 of [27], $\langle f, \boldsymbol{Z}f\rangle_\pi = \sigma_{\mathrm{as}}^2 + (1/2)\langle f, f\rangle_\pi$, thus $\beta^{(2)} = \sigma_{\mathrm{as}}^2$. For $n = 3$, after some calculations, using the fact that $\boldsymbol{Z}$ and $\boldsymbol{P}$ commute, we have

$$\begin{aligned}\beta^{(3)} &= \langle f, \boldsymbol{Z}\boldsymbol{P}\boldsymbol{D}_f\boldsymbol{Z}\boldsymbol{P}f\rangle_\pi + \left\langle f, \boldsymbol{Z}\boldsymbol{P}f^2\right\rangle_\pi + \frac{1}{6}\mathbb{E}_\pi(f^3)\\ &= \left\langle \boldsymbol{Z}^{1/2}f, \boldsymbol{Z}^{1/2}\boldsymbol{P}\boldsymbol{D}_f\boldsymbol{P}\boldsymbol{Z}^{1/2}(\boldsymbol{Z}^{1/2}f)\right\rangle_\pi + \left\langle f, \boldsymbol{Z}\boldsymbol{P}f^2\right\rangle_\pi + \frac{1}{6}\langle f, \boldsymbol{D}_f f\rangle_\pi,\end{aligned}$$

and we have $\left\langle f, \boldsymbol{Z}\boldsymbol{P}f^2\right\rangle_\pi \leq \frac{V_f}{\gamma}$, $\frac{1}{6}\langle f, \boldsymbol{D}_f f\rangle_\pi \leq \frac{1}{6}V_f$,

$$\begin{aligned}&\left\langle \boldsymbol{Z}^{1/2}f, \boldsymbol{Z}^{1/2}\boldsymbol{P}\boldsymbol{D}_f\boldsymbol{P}\boldsymbol{Z}^{1/2}(\boldsymbol{Z}^{1/2}f)\right\rangle_\pi \qquad \leq \|\boldsymbol{Z}^{1/2}f\|_{2,\pi}^2 \cdot \|\boldsymbol{Z}^{1/2}\boldsymbol{P}\boldsymbol{D}_f\boldsymbol{P}\boldsymbol{Z}^{1/2}\|_{2,\pi}\\ &\leq \frac{1}{\gamma}\langle f, \boldsymbol{Z}f\rangle_\pi = \frac{1}{\gamma}\left(\sigma_{\mathrm{as}}^2 + V_f/2\right),\end{aligned}$$

thus $|\beta^{(3)}| \leq \sigma_{\mathrm{as}}^2/\gamma + (3/2)V_f/\gamma + (1/6)V_f$. Suppose now that $n \geq 4$. First, if $p = n$, then $\nu_1 = \ldots = \nu_p = 1$, thus each such term in (5.9) looks like

$$\begin{aligned}&\left\langle f, \boldsymbol{Z}^{(k'_1)}\boldsymbol{P}\boldsymbol{D}_f\cdots\boldsymbol{Z}^{(k'_{n-1})}\boldsymbol{P}D_f\boldsymbol{Z}^{(k'_{n-1})}\boldsymbol{P}f\right\rangle_\pi\\ &= \left\langle f, \boldsymbol{Z}^{(k'_1)}\boldsymbol{P}\boldsymbol{D}_f\cdots\boldsymbol{Z}^{(k'_{n-1})}\boldsymbol{P}D_f\boldsymbol{P}\boldsymbol{Z}^{(k'_{n-1})}f\right\rangle_\pi.\end{aligned}$$

If k'_1 or k'_{n-1} are 0, then such terms equal zero (since $\boldsymbol{\pi}(f) = 0$). If they are at least one, then we can bound the absolute value of this by

$$\begin{aligned}&\left|\left\langle Z^{1/2}f, \boldsymbol{Z}^{k'_1-1/2}\boldsymbol{P}\boldsymbol{D}_f\cdots\boldsymbol{Z}^{(k'_{n-1})}\boldsymbol{P}D_f\boldsymbol{P}\boldsymbol{Z}^{k'_{n-1}-1/2}(\boldsymbol{Z}^{1/2}f)\right\rangle_\pi\right|\\ &\leq \frac{\langle f, \boldsymbol{Z}f\rangle_\pi}{2\gamma^{n-2}} \leq \frac{\sigma_{\mathrm{as}}^2 + V_f}{2\gamma^{n-2}}.\end{aligned}$$

It is easy to see that there are $\binom{2(n-1)}{n-1}$ such terms. For $1 \le p < n$, we have

$$\left\| \left\langle f^{\nu_1'}, \boldsymbol{Z}^{(k_1')} \boldsymbol{P} \boldsymbol{D}_f^{\nu_2} \cdots \boldsymbol{Z}^{(k_{p-1}')} \boldsymbol{P} f^{\nu_p'} \right\rangle_\pi \right\| \le \frac{V_f}{\gamma^{p-1}},$$

and there are $\binom{n-1}{p-1}\binom{2(p-1)}{p-1}$ such terms. By summing up, and using the fact that $\nu_1! \cdots \nu_p! \ge 2^{n-p}$, and $2/\gamma \ge 1$, we obtain

$$\begin{aligned} |\beta^{(n)}| &\le \frac{1}{n}\binom{2(n-1)}{n-1}\frac{\sigma_{\mathrm{as}}^2 + V_f}{2\gamma^{n-2}} + \sum_{p=1}^{n-1}\frac{1}{p}\binom{n-1}{p-1}\binom{2(p-1)}{p-1}\frac{1}{2^{n-p}}\cdot\frac{V_f}{\gamma^{p-1}} \\ &\le \frac{1}{n}\binom{2(n-1)}{n-1}\frac{\sigma_{\mathrm{as}}^2 + V_f}{2\gamma^{n-2}} + \frac{V_f}{2^{n-1}}\sum_{p=1}^{n-1}\frac{1}{p}\binom{n-1}{p-1}\binom{2(p-1)}{p-1}\left(\frac{2}{\gamma}\right)^{n-2}. \end{aligned}$$

Now by (1.11) on page 20 of [27], we have $\binom{2(n-1)}{n-1} \le \frac{4^{(n-1)}}{\sqrt{(n-1)\pi}}$. Define $D(n) := \sum_{p=1}^{n}\frac{1}{p}\binom{n-1}{p-1}\binom{2(p-1)}{p-1}$, then by page 47 of [27], for $n \ge 3$, $D(n) \le 5^{n-2}$. Thus for $n \ge 4$, we have

$$\begin{aligned} |\beta^{(n)}| &\le \frac{4^{n-1}}{n\sqrt{(n-1)\pi}}\frac{\sigma_{\mathrm{as}}^2 + V_f}{2\gamma^{n-2}} + \frac{5^{n-3}}{2\gamma^{n-2}}V_f \qquad (5.10) \\ &\le \frac{5^{n-2}}{\gamma^{n-2}}\left(\frac{\sigma_{\mathrm{as}}^2 + V_f}{2}\cdot\frac{1}{4} + \frac{V_f}{10}\right) \le \frac{5^{n-2}}{\gamma^{n-2}}\left(\frac{\sigma_{\mathrm{as}}^2 + 0.8V_f}{2}\right). \end{aligned}$$

By comparing this with our previous bounds on $\beta^{(2)}$ and $\beta^{(3)}$, we can see that (5.10) holds for every $n \ge 2$. By summing up, we obtain

$$\lambda_{\max}(\boldsymbol{P}(r)) = 1 + \sum_{n=1}^{\infty}\beta^{(n)}r^n \le 1 + \frac{\sigma_{\mathrm{as}}^2 + 0.8V_f}{2}\cdot\frac{r^2}{1 - 5r/\gamma},$$

and substituting $r = 2\theta$ gives (5.8). □

Proof of Theorem 3.3. We can assume, without loss of generality, that $C = 1$. First, we will prove the bounds for S, then for S'.

By (5.5), we have

$$\mathbb{E}_\pi(\exp(\theta S)) \le \|e^{\theta \boldsymbol{D}_f/2}\boldsymbol{P}e^{\theta \boldsymbol{D}_f/2}\|_2^{n-1}\cdot\mathbb{E}_\pi\left(e^{\theta f}\right). \qquad (5.11)$$

By (5.7), and (5.8), we have that for $0 \le \theta \le \gamma/5$,

$$\|e^{\theta \boldsymbol{D}_f/2}\boldsymbol{P}e^{\theta \boldsymbol{D}_f/2}\|_{2,\pi} \le \exp\left(\frac{V_f}{\gamma}\cdot\theta^2\cdot\left(1 - \frac{5\theta}{\gamma}\right)^{-1}\right), \text{ and} \qquad (5.12)$$

$$\|e^{\theta \boldsymbol{D}_f/2}\boldsymbol{P}e^{\theta \boldsymbol{D}_f/2}\|_{2,\pi} \le \exp\left(\frac{\sigma_{\mathrm{as}}^2 + 0.8V_f}{2}\cdot\theta^2\cdot\left(1 - \frac{5\theta}{\gamma}\right)^{-1}\right). \qquad (5.13)$$

Now using the fact that $-1 \le f(x) \le 1$, $\mathbb{E}_\pi(f) = 0$, it is easy to show that for any $\theta \ge 0$,

$$\mathbb{E}_\pi\left(e^{\theta f}\right) \le \exp\left(V_f(e^\theta - \theta - 1)\right),$$

and it is also easy to show that this can be indeed further bounded by the right hand sides of (5.12) and (5.13). Therefore, we obtain that for $0 \le \theta \le \gamma/5$,

$$\mathbb{E}_\pi(\exp(\theta S)) \le \exp\left(\frac{nV_f}{\gamma}\cdot\theta^2\cdot\left(1 - \frac{5\theta}{\gamma}\right)^{-1}\right), \text{ and}$$

$$\mathbb{E}_\pi(\exp(\theta S)) \le \exp\left(\frac{n\sigma_{\rm as}^2 + 0.8V_f}{2}\cdot\theta^2\cdot\left(1-\frac{5\theta}{\gamma}\right)^{-1}\right).$$

Now the bounds (3.21) and (3.20) follow by Markov's inequality, for the optimal choice

$$\theta = \frac{t\gamma}{V_f(1+5t/V_f+\sqrt{1+5t/V_f})}, \text{ and } \theta = \frac{t}{5t/\gamma + K(1+\sqrt{1+5t/(\gamma K)})},$$

with $K = 0.5\sigma_{\rm as}^2 + 0.4V_f$.

Now we are going to prove (3.22). Firstly, by (5.6), we have

$$\mathbb{E}_\pi(\exp(\theta S')) \le \|\boldsymbol{P}e^{\theta \boldsymbol{D}_{f_1}}\|_{2,\pi}\cdot\ldots\cdot\|\boldsymbol{P}e^{\theta\boldsymbol{D}_{f_n}}\|_{2,\pi}. \tag{5.14}$$

Now for $0\le\theta\le\gamma(\boldsymbol{P}^2)/10$, each of these terms can be further bounded by (5.7) as

$$\|\boldsymbol{P}e^{\theta\boldsymbol{D}_{f_i}}\|_{2,\pi} = \|e^{\theta\boldsymbol{D}_{f_i}}\boldsymbol{P}^2e^{\theta\boldsymbol{D}_{f_i}}\|_{2,\pi}^{1/2} \le \exp\left(\frac{2\mathbb{E}_\pi(f_i^2)}{\gamma(\boldsymbol{P}^2)}\cdot\theta^2\cdot\left(1-\frac{10\theta}{\gamma(\boldsymbol{P}^2)}\right)^{-1}\right).$$

By taking the product for $1\le i\le n$, we obtain that for $0\le\theta\le\gamma(\boldsymbol{P}^2)/10$,

$$\mathbb{E}_\pi(\exp(\theta S')) \le \exp\left(\frac{2V_{S'}}{\gamma(\boldsymbol{P}^2)}\cdot\theta^2\cdot\left(1-\frac{10\theta}{\gamma(\boldsymbol{P}^2)}\right)^{-1}\right), \tag{5.15}$$

and (3.22) follows by Markov's inequality. □

Proof of Theorem 3.4. We will treat the general case concerning S' first. The proof is based on a trick of [19]. First, we divide the sequence $f_1(X_1),\ldots,f_n(X_n)$ into $k_{\rm ps}$ parts,

$$\left(f_1(X_1), f_{k_{\rm ps}+1}(X_{k_{\rm ps}+1}),\ldots,\right),\ldots,\left(\left(f_{k_{\rm ps}}(X_{k_{\rm ps}}), f_{2k_{\rm ps}}(X_{2k_{\rm ps}}),\ldots,\right)\right).$$

Denote the sums of each part by $S_1',\ldots,S_{k_{\rm ps}}'$, then $S' = \sum_{i=1}^{k_{\rm ps}} S_i'$. By Yensen's inequality, for any weights $0\le p_1,\ldots,p_{k_{\rm ps}}\le 1$ with $\sum_{i=1}^{k_{\rm ps}} p_i = 1$,

$$\mathbb{E}_\pi\exp(\theta S') \le \sum_{i=1}^{k_{\rm ps}} p_i\mathbb{E}_\pi\exp((\theta/p_i)\cdot S_i'). \tag{5.16}$$

Now we proceed the estimate the terms $\mathbb{E}\exp(\theta S_i')$.

Notice that $X_i, X_{i+k_{\rm ps}},\ldots,X_{i+k_{\rm ps}\lfloor(n-i)/k_{\rm ps}\rfloor}$ is a Markov chain with transition kernel $\boldsymbol{P}^{k_{\rm ps}}$. Using (5.6) on this chain, we have

$$\mathbb{E}_\pi(\exp(\theta S_i')) \le \|\boldsymbol{P}^{k_{\rm ps}}e^{\theta\boldsymbol{D}_{f_i}}\|_{2,\pi}\cdot\ldots\cdot\|\boldsymbol{P}^{k_{\rm ps}}e^{\theta\boldsymbol{D}_{f_{i+k_{\rm ps}\lfloor(n-i)/k_{\rm ps}\rfloor}}}\|_{2,\pi}.$$

Now

$$\left\|\boldsymbol{P}^{k_{\rm ps}}e^{\theta\boldsymbol{D}_{f_j}}\right\|_{2,\pi} = \left\|e^{\theta\boldsymbol{D}_{f_j}}\left(\boldsymbol{P}^*\right)^{k_{\rm ps}}\boldsymbol{P}^{k_{\rm ps}}e^{\theta\boldsymbol{D}_{f_j}}\right\|_{2,\pi}^{1/2}.$$

By (5.7), and using the assumption $\mathbb{E}_\pi(f_j)=0$,

$$\begin{aligned}&\left\|\boldsymbol{P}^k e^{\theta\boldsymbol{D}_{f_j}}\right\|_{2,\pi}\\ &\le \|e^{\theta\boldsymbol{D}_f}\boldsymbol{P}e^{\theta\boldsymbol{D}_f}\|_{2,\pi} \le \exp\left(\frac{2\mathrm{Var}_\pi(f_j)}{\gamma((\boldsymbol{P}^*)^{k_{\rm ps}}\boldsymbol{P}^{k_{\rm ps}})}\cdot\theta^2\cdot\left(1-\frac{10\theta}{\gamma((\boldsymbol{P}^*)^{k_{\rm ps}}\boldsymbol{P}^{k_{\rm ps}})}\right)^{-1}\right).\end{aligned}$$

By taking the product of these, we have

$$\begin{aligned}&\mathbb{E}_\pi(\exp(\theta S_i'))\\&\leq \exp\left(\frac{2\sum_{j=0}^{\lfloor (n-i)/k_{\mathrm{ps}}\rfloor}\mathrm{Var}_\pi(f_{i+jk_{\mathrm{ps}}})}{\gamma((\boldsymbol{P}^*)^{k_{\mathrm{ps}}}\boldsymbol{P}^{k_{\mathrm{ps}}})}\cdot\theta^2\cdot\left(1-\frac{10\theta}{\gamma((\boldsymbol{P}^*)^{k_{\mathrm{ps}}}\boldsymbol{P}^{k_{\mathrm{ps}}})}\right)^{-1}\right).\end{aligned}$$

These bounds hold for every $1\leq i\leq k_{\mathrm{ps}}$. Setting p_i in (5.16) as

$$p_i := V_i^{1/2}/\left(\sum_{i=1}^k V_i^{1/2}\right),$$

and using the inequality $(\sum_{i=1}^{k_{\mathrm{ps}}} V_i^{1/2})^2\leq k_{\mathrm{ps}}\sum_{i=1}^n V_i$, we obtain

$$\begin{aligned}\mathbb{E}_\pi(\exp(\theta S'))&\leq \exp\left(\frac{2k_{\mathrm{ps}}\sum_{j=1}^n\mathrm{Var}_\pi(f_j)}{\gamma((\boldsymbol{P}^*)^{k_{\mathrm{ps}}}\boldsymbol{P}^{k_{\mathrm{ps}}})}\cdot\theta^2\cdot\left(1-\frac{10\theta\cdot M}{\gamma((\boldsymbol{P}^*)^{k_{\mathrm{ps}}}\boldsymbol{P}^{k_{\mathrm{ps}}})}\right)^{-1}\right)\\&\leq \exp\left(\frac{2\sum_{j=1}^n\mathrm{Var}_\pi(f_j)}{\gamma_{\mathrm{ps}}}\cdot\theta^2\cdot\left(1-\frac{10\theta\cdot M}{k_{\mathrm{ps}}\gamma_{\mathrm{ps}}}\right)^{-1}\right),\end{aligned}$$

and (3.24) follows by Markov's inequality. In the case of (3.23), we have

$$\begin{aligned}&\mathbb{E}_\pi(\exp(\theta S_i'))\\&\leq \exp\left(\frac{2\lceil n/k_{\mathrm{ps}}\rceil}{\gamma((\boldsymbol{P}^*)^{k_{\mathrm{ps}}}\boldsymbol{P}^{k_{\mathrm{ps}}})}\cdot\theta^2\cdot\left(1-\frac{10\theta}{\gamma((\boldsymbol{P}^*)^{k_{\mathrm{ps}}}\boldsymbol{P}^{k_{\mathrm{ps}}})}\right)^{-1}\right),\end{aligned}$$

which implies that

$$\mathbb{E}_\pi(\exp(\theta S))\leq \exp\left(\frac{2k_{\mathrm{ps}}\lceil n/k_{\mathrm{ps}}\rceil\mathrm{Var}_\pi(f)}{\gamma_{\mathrm{ps}}}\cdot\theta^2\cdot\left(1-\frac{10\theta}{\gamma_{\mathrm{ps}}}\right)^{-1}\right).$$

Now (3.23) follows by Markov's inequality and $k_{\mathrm{ps}}\lceil n/k_{\mathrm{ps}}\rceil\leq n+1/\gamma_{\mathrm{ps}}$. □

Proof of Proposition 3.10. Inequalities (3.25) and (3.26) follow by writing

$$\begin{aligned}\mathbb{P}_q\left(g(X_1,\dots,X_n)\geq t\right)&=\mathbb{E}_q\left(\mathbb{1}[g(X_1,\dots,X_n)\geq t]\right)\\&=\mathbb{E}_\pi\left(\frac{\mathrm{d}q}{\mathrm{d}\pi}\cdot\mathbb{1}[g(X_1,\dots,X_n)\geq t]\right),\end{aligned}$$

and then applying Cauchy-Schwartz inequality. Inequality (3.27) follows by noticing that by the Markov property, the two distributions

$$\mathcal{L}(X_{t_0+1},\dots,X_n|X_1\sim q)\text{ and }\mathcal{L}(X_{t_0+1},\dots,X_n|X_1\sim\pi)$$

have total variational distance equal to the total variational distance of

$$\mathcal{L}(X_{t_0+1}|X_1\sim q)\text{ and }\mathcal{L}(X_{t_0+1}|X_1\sim\pi).$$ □

Proof of Proposition 3.11. Inequalities (3.28) and (3.29) follow from (2.11) on page 68 of [13], similarly to the proof of Proposition 3.4 (by noticing that the χ^2 distance can be written as N_q-1). Finally, (3.30) follows from the definition of $\tau(\epsilon)$ and t_{mix}. □

Proof of Proposition 3.12. This follows by a straightforward coupling argument. The details are left to the reader. □

Acknowledgments. The author thanks his thesis supervisors, Louis Chen and Adrian Röllin, for their useful advices. He thanks Emmanuel Rio, Laurent Saloff-Coste, Olivier Wintenberger, Zhipeng Liao, and Daniel Rudolf for their useful comments. He thanks Doma Szász and Mogyi Tóth for infecting him with their enthusiasm of probability. Finally, many thanks to Roland Paulin for the enlightening discussions.

References

[1] Radosław Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13:no. 34, 1000–1034, 2008. MR-2424985

[2] Radosław Adamczak and Witold Bednorz. Exponential concentration inequalities for additive functionals of markov chains. arXiv:1201.3569.

[3] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. MR-3185193

[4] J.-R. Chazottes, P. Collet, C. Külske, and F. Redig. Concentration inequalities for random fields via coupling. *Probab. Theory Related Fields*, 137(1-2):201–225, 2007. MR-2278456

[5] Jean-René Chazottes and Frank Redig. Concentration inequalities for Markov processes via coupling. *Electron. J. Probab.*, 14:no. 40, 1162–1180, 2009. MR-2511280

[6] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001. MR-1843146

[7] Persi Diaconis, Susan Holmes, and Richard Montgomery. Dynamical bias in the coin toss. *SIAM Rev.*, 49(2):211–235, 2007. MR-2327054

[8] H. Djellout, A. Guillin, and L. Wu. Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *Ann. Probab.*, 32(3B):2702–2732, 2004. MR-2078555

[9] Wolfang Doeblin. Exposé de la théorie des chaınes simples constantes de markova un nombre fini d'états. *Mathématique de l'Union Interbalkanique*, 2(77-105):78–80, 1938.

[10] Randal Douc, Eric Moulines, Jimmy Olsson, and Ramon van Handel. Consistency of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.*, 39(1):474–513, 2011. MR-2797854

[11] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27:642–669, 1956. MR-0083864

[12] Doris Fiebig. Mixing properties of a class of Bernoulli-processes. *Trans. Amer. Math. Soc.*, 338(1):479–493, 1993. MR-1102220

[13] James Allen Fill. Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *Ann. Appl. Probab.*, 1(1):62–87, 1991. MR-1097464

[14] David Gillman. A Chernoff bound for random walks on expander graphs. *SIAM J. Comput.*, 27(4):1203–1220, 1998. MR-1621958

[15] Peter W. Glynn and Dirk Ormoneit. Hoeffding's inequality for uniformly ergodic Markov chains. *Statist. Probab. Lett.*, 56(2):143–146, 2002. MR-1881167

[16] Sheldon Goldstein. Maximal coupling. *Z. Wahrsch. Verw. Gebiete*, 46(2):193–204, 1978/79. MR-0516740

[17] B. Gyori and D. Paulin. Non-asymptotic confidence intervals for MCMC in practice. *arXiv preprint*, 2014.

[18] ShuLan Hu. Transportation inequalities for hidden Markov chains and applications. *Sci. China Math.*, 54(5):1027–1042, 2011. MR-2800925

[19] Svante Janson. Large deviations for sums of partly dependent random variables. *Random Structures Algorithms*, 24(3):234–248, 2004. MR-2068873

[20] A. Kontorovich and R. Weiss. Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for Markov chains and related processes. arXiv:1207.4678. MR-3301291

[21] L. Kontorovich. Measure concentration of hidden Markov processes. arXiv:math/0608064, 2006.

[22] L. Kontorovich. *Measure Concentration of Strongly Mixing Processes with Applications*. 2007. Ph.D. dissertation, Carnegie Mellon University, Available at `http://www.cs.bgu.ac.il/~karyeh/thesis.pdf`. MR-2710649

[23] Leonid Kontorovich and Kavita Ramanan. Concentration inequalities for dependent random variables via the martingale method. *Ann. Probab.*, 36(6):2126–2158, 2008. MR-2478678

[24] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001. MR-1849347

[25] C. A. León and F. Perron. Optimal Hoeffding bounds for discrete reversible Markov chains. *Ann. Appl. Probab.*, 14(2):958–970, 2004. MR-2052909

[26] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, Providence, RI, 2009. With a chapter by James G. Propp and David B. Wilson. MR-2466937

[27] P. Lezaud. *Etude quantitative des chaînes de Markov par perturbation de leur noyau*. 1998. Thèse doctorat mathématiques appliquées de l'Université Paul Sabatier de Toulouse, Available at `http://pom.tls.cena.fr/papers/thesis/these_lezaud.pdf`.

[28] Pascal Lezaud. Chernoff-type bound for finite Markov chains. *Ann. Appl. Probab.*, 8(3):849–867, 1998. MR-1627795

[29] Pascal Lezaud. Chernoff and Berry-Esséen inequalities for Markov processes. *ESAIM Probab. Statist.*, 5:183–201, 2001. MR-1875670

[30] Torgny Lindvall. *Lectures on the coupling method*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1992. A Wiley-Interscience Publication. MR-1180522

[31] Eyal Lubetzky and Allan Sly. Cutoff for the Ising model on the lattice. *Invent. Math.*, 191(3):719–755, 2013. MR-3020173

[32] K. Marton. Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration. *Ann. Probab.*, 24(2):857–866, 1996. MR-1404531

[33] K. Marton. A measure concentration inequality for contracting Markov chains. *Geom. Funct. Anal.*, 6(3):556–571, 1996. MR-1392329

[34] K. Marton. Erratum to: "A measure concentration inequality for contracting Markov chains" [Geom. Funct. Anal. **6** (1996), no. 3, 556–571; MR1392329 (97g:60082)]. *Geom. Funct. Anal.*, 7(3):609–613, 1997. MR-1392329

[35] K. Marton. Measure concentration and strong mixing. *Studia Sci. Math. Hungar.*, 40(1-2):95–113, 2003. MR-2002993

[36] Katalin Marton. Measure concentration for a class of random processes. *Probab. Theory Related Fields*, 110(3):427–439, 1998. MR-1616492

[37] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 1990. MR-1062069

[38] Florence Merlevède, Magda Peligrad, and Emmanuel Rio. A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probab. Theory Related Fields*, 151(3-4):435–474, 2011. MR-2851689

[39] Yann Ollivier. Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.*, 256(3):810–864, 2009. MR-2484937

[40] D. Paulin. *Concentration inequalities for dependent random variables*. 2014. Thesis (Ph.D.), National University of Singapore. Available at `http://www.scholarbank.nus.edu.sg/handle/10635/118229`.

[41] D. Paulin. Mixing and concentration by Ricci curvature. *arXiv preprint*, 2014.

[42] Emmanuel Rio. Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(10):905–908, 2000. MR-1771956

[43] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probab. Surv.*, 1:20–71, 2004. MR-2095565

[44] Jeffrey S. Rosenthal. Faithful couplings of Markov chains: now equals forever. *Adv. in Appl. Math.*, 18(3):372–381, 1997. MR-1436487

[45] Paul-Marie Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.*, 28(1):416–461, 2000. MR-1756011

[46] O. Wintenberger. Weak transport inequalities and applications to exponential inequalities and oracle inequalities. *ArXiv e-prints,* 2012.